Numéro d'identification :

THÈSE

pour l'obtention du Grade de
DOCTEUR DE L'UNIVERSITÉ DE POITIERS

(Faculté des Sciences Fondamentales et Appliquées)
(Diplôme National - Arrêté du 7 août 2006)

École Doctorale : Sciences et Ingénierie pour l'Information
Secteur de Recherche : Traitement du signal et des images, informatique et applications

Présentée par : **Stéphane SOCHACKI**

DE LA COMBINAISON/COMPÉTITION DE CLASSIFIEURS VERS LA SÉLECTION DYNAMIQUE D'OPÉRATEURS DE TRAITEMENT D'IMAGE

Un problème de métrique, de qualité, de précision et de combinaison intelligente

Directeur de Thèse : **Christine FERNANDEZ-MALOIGNE**
Co-Directeur de Thèse : **Noël RICHARD**

Soutenue le : **05 Décembre 2011** devant la Commission d'examen composée de :

M. Olivier COLOT (Professeur), Université Lille I, LAGIS	Examinateur
M. Didier COQUIN (Maître de Conférences - HDR), Polytech Annecy-Chambéry, LISTIC	Rapporteur
M. Laurent WENDLING (Professeur), Université Paris V, LIPADE	Rapporteur
M. Jean-Yves RAMEL (Professeur), Polytech'Tours, LI	Examinateur
Mme. Christine FERNANDEZ-MALOIGNE (Professeur), Université de Poitiers, XLIM-SIC	Directrice de Thèse
M. Noël RICHARD (Maître de Conférences), Université de Poitiers, XLIM-SIC	Co-directeur de Thèse
Mme. Anne-Sophie CAPELLE-LAIZÉ (Maître de Conférences), Université de Poitiers, XLIM-SIC	Co-encadrante

Résumé

L'évolution des technologies a permis le développement du multimédia et des grandes bases de données hétérogènes (textes, images, sons, vidéo, ...). Cependant cette évolution a amplifié le besoin de solutions de traitement automatique. Bien que de nombreuses années de recherche aient abouti au développement d'opérateurs de plus en plus performants, leur interopérabilité et parfois même leurs réglages restent à la charge des experts. Cette limite nuit au développement de systèmes pluri-média spécialisés et grand public. Ce travail de recherche se partage en trois parties principales.

Gérer de façon automatique une chaîne de traitement d'images allant du prétraitement à la décision, impose de construire un système rétro-bouclé. Chaque étage nécessite la conception d'informations liées à la qualité ou la précision de l'action effectuée mais aussi d'informations indiquant la pertinence de cette action selon l'objectif visé. Cette première partie s'appuie sur la Théorie de l'Evidence. Une phase intermédiaire de validation sera mise en oeuvre autour des classes d'opérateurs de calcul d'attributs et de classification.

Les variabilités intrinsèques aux documents patrimoniaux et par extension aux bases multimédia introduisent la difficulté du choix d'une séquence unique d'opérateurs. Afin de décider dynamiquement de la meilleure chaîne pour le document d'entrée permettant d'atteindre l'objectif visé, le système doit pouvoir mettre en concurrence/compétition les différents opérateurs dans une même classe. Cette deuxième partie exploite les critères de précision/pertinence établis précédemment. Une validation intermédiaire sur la gestion des attributs de caractérisation des primitives graphiques et des classifieurs hiérarchiques sera utilisée.

La classification se basant sur un ensemble de données, les attributs, il faut également que la chaîne de traitement manipule des opérateurs comme des ana-

lyses de texture ou de forme. Mais pour les intégrer, nous avons besoin d'en connaître la précision par l'utilité et la conservation de l'information. Ce sera la troisième grande étape, au travers de l'étude d'attributs de descripteurs de formes.

Le déroulement de ce travail de thèse permettra de développer ces différents points dans le cadre d'une chaîne de traitement complète. Le cadre de travail permettra d'intégrer ces développements dans une application dédiée à l'analyse et la valorisation de documents patrimoniaux. Des éléments intermédiaires de validation seront développés pour quantifier la qualité et la stabilité des solutions proposées. Ce travail de thèse s'inscrit dans la suite de celui de D. Arrivault entre le laboratoire XLIM-SIC et la société RC-Soft, il rentre également dans le cadre des travaux dont le résultat est exploité au sein du programme des Environnements Documentaires Interactifs auxquels ont participé les deux entités.

Mots clés

image, reconnaissance de forme, exactitude et attributs numériques, combinaison/compétition d'attributs, chaîne de traitement d'image, classification

Title

From classifiers combination/competition to image processing operators dynamic selection

Abstract

Technology evolutions bring multimedia and huge heterogeneous databases (texts, images, sounds, video etc.) development. Nevertheless, this evolution amplified the need of automatic processing solutions. Whereas many years of research resulted in more and more powerful operators development, their interoperability and sometimes their tuning remain supported by the expert. This limit harms the specialized multi-media and consumer systems development. This research work is divided in three main parts.

The automatic processing chain management, from preprocessing to decision, leads a reverse looped system to be build. Each stage needs the conception of information relative to the quality or the accuracy of the done action, and information indicating the relevance of this action according to the desired goal as well. This first part is based on the Theory of Evidence. An intermediate validation phase will be achieved on classes of attributes computing and classification operators.

Patrimonial documents, and by extension multimedia databases, intrinsic variability introduce the choice trouble of a unique operators sequence. In order to dynamically decide the best chain for the input document, allowing to reach the desired goal, the system would have to put in competition the different operators from a same class. This second part exploits accuracy/relevance previously proposed. An intermediate validation about graphical primitives description attributes and hierarchical classifiers management will be used.

As classification is based on a set of data, the attributes, the processing chain will have to manipulate operators like texture or shape analysis. But for their integration, we need to know the accuracy by usefulness and information preservation. This will be the third main step, via a shape descriptors attributes study.

The progress of this thesis work will help to develop this different points within a complete processing chain. All those development will be integrated in an application dedicated to the analysis and valorization of patrimonial documents. Intermediate validation elements will be developed to evaluate the quality and the stability of all proposed solutions. This thesis work comes next to D. Arrivault's between the XLIM-SIC laboratory and the RC-Soft company and comes within works which results are exploited in the Interactive Documentary Environments to which the two entities participated.

Keywords

image, pattern recognition, accuracy and numeric attributes, attributes combination/competition, image processing chain, classification

Remerciements

Le travail que vous allez découvrir a été réalisé au Laboratoire XLIM-SIC de l'Université de Poitiers. Il n'aurait jamais abouti sans l'aide et le soutien de nombreuses personnes.

En premier lieu je tiens à remercier Messieurs Didier Coquin et Laurent Wendling pour avoir accepté de rapporter sur ce travail de Thèse. Un grand merci à Messieurs Olivier Colot et Jean-Yves Ramel pour en être examinateurs. Merci également à Anne-Sophie Capelle-Laizé pour son apport en tant que membre de jury évidement mais également pour l'aide apporté tout au long de ce travail.

Merci à Monsieur Noël Richard, mon encadrant et co-Directeur de Thèse, pour le temps, la disponibilité, les conseils et les idées qu'il m'a accordés durant ces années de collaboration.

Merci également à Madame Christine Fernandez-Maloigne, ma Directrice de Thèse et également Directrice du laboratoire XLIM-SIC, pour m'avoir accueilli au sein du laboratoire et mis à disposition les moyens nécessaires à l'aboutissement de ce travail de Thèse.

Un grand merci, pour tout, à Monsieur Philippe Bouyer, mon ami. Sans la magie de sa rencontre au moment où l'idée de ce projet germait dans nos deux têtes, sans notre grain de folie, sans son soutien et ses conseils, ce projet n'aurait jamais vu le jour.

Mes remerciements vont également aux membres du Jury.

Je n'oublie pas tous ceux qui ont toujours été là pour moi, à me supporter et m'accompagner, autant dans mes moments de joie que mes moments de doutes.

Merci à mes proches pour m'accepter comme je suis.

Merci à ceux grâce à qui j'en suis là aujourd'hui, notamment Monsieur Pierre Loonis qui su me transmettre le goût pour la recherche alors que j'étais son élève à l'université de La Rochelle.

Avant de finir, un grand merci inconditionnel à mes parents pour m'avoir tenu, soutenu et encouragé à simplement être moi et faire ce que je voulais, et pour les sacrifices qu'ils ont fait pour cela. Merci aussi à Ness, d'être là tout simplement.

Ce travail a été financé à la fois par la société RCSOFT à Angoulême et par l'ASSEDIC de la région Poitou-Charentes.

A Sarah...

Table des matières

"*On commence par dire : cela est impossible pour se dispenser de le tenter, et cela devient impossible, en effet, parce qu'on ne le tente pas.*"
Charles Fourier

Table des notations/symboles

Symbole	Signification
x	Individu connu
$\{x\}$	Ensemble des individus connus
x^*	Individu inconnu
$\{x^*\}$	Ensemble des individus inconnus
$\mathcal{L}$	Ensemble d'apprentissage ($\mathcal{L} = \{x\}$)
N_x	Nombre d'individus dans l'espace d'apprentissage : $N_x = \|L\|$
$\mathcal{T}$	Ensemble de test ($\mathcal{T} = \{x^*\}$)
N_{x^*}	Nombre d'individus dans l'espace de test : $N_{x^*} = \|T\|$
$\mathcal{E}$	Ensemble de tous les individus x et x^*
ω_q	Classe q
Ω	Ensemble de toutes les classes possibles $\Omega = \{\omega_1, ..., \omega_{N_\omega}\}$
N_ω	Nombre de classes : $N_\omega = \mathsf{card}(\Omega)$
$P(\omega_q)$	Probabilité d'apparition de la classe ω_q
$A_j(x)$	Attributs de l'individu x selon l'opérateur j
$\mathcal{A}$	Espace des attributs
N_a	Nombre d'opérateurs de calcul d'attributs
α_j	Précision mesurée en sortie de l'opérateur de calcul d'attributs j
$C_i(\omega_q\|x)$	Distance de l'individu x à la classe ω_q selon le classifieur i
$C_i(\omega_q\|A_j(x))$	Distance de l'individu x à la classe ω_q selon ses attributs A_j et le classifieur i
$C_i(x)$	Classe affectée à l'individu x selon le classifieur i
$\mathcal{C}$	Ensemble des classifieurs
N_c	Nombre de classifieurs
γ_i	Précision mesurée en sortie du classifieur i
$L(\omega_q\|x)$	Fonction de vraisemblance
$m(B)$	Masse de croyance
m_{ijq}	Masse de croyance pour la classe q, selon les attributs j et le classifieur i
$B(x^*)$	Voisinage de x^* de taille k
$LA_{j,k}(x^*)$	Précision locale estimée du classifieur j dans le voisinage $B(x^*)$

Chapitre 1

Introduction

1.1 Contexte

1.1.1 Document

Durant toute l'histoire de l'humanité, l'écriture, ou son équivalent pictural, s'est toujours faite par dépôt d'un pigment (comme de l'encre par exemple) ou altération d'une surface matérielle. La mise en forme, la modification ou l'effacement d'une production étaient donc très limités, et la recopie était souvent la seule solution pour y opérer une modification. L'arrivée des systèmes informatiques permet aujourd'hui la création de documents modifiables à volonté, et mis en page avant impression.

Cependant, le support utilisé est d'une certaine façon virtuel, et a peu de liens avec le papier, le parchemin, le papyrus ou la tablette d'argile. Se pose le problème de la transmission de l'écrit sur le support informatique. Dans le sens fichier texte vers un support papier, le problème est résolu par les périphériques d'impression. Dans l'autre sens, le problème est plus complexe. Les outils utilisés comme périphériques d'acquisition (scanners, caméras ou appareils photo numériques) traitent un document comme étant une image dans sa globalité, et ont des difficultés à faire la différence entre un caractère alphabétique et un dessin. Ils n'interprètent pas.

La recherche pour créer des outils pouvant reconnaître des caractères à partir d'une image et les convertir en caractères compréhensibles par l'informatique a débuté dès les années cinquante. Différentes voies ont été explorées (matricielles, topologiques et neuronales), mais ce n'est que récemment que des outils permettant un taux de reconnaissance acceptable sont apparus. Il y a encore peu, ces outils (OCR [1]) étaient limités à la reconnaissance de caractères imprimés dans un alphabet donné. Pour reconnaître un document en hébreu, il faut disposer d'un logiciel traitant spécifiquement cette langue. Cette limitation est paradoxale, car à l'inverse plusieurs polices d'alphabets différents sont disponibles sur tout ordinateur courant. Aujourd'hui, le contexte a évolué car ces logiciels proposent de traiter plusieurs langues, mais nous en sommes encore aux débuts d'une technologie qui demande à évoluer.

D'autre part, la reconnaissance d'écriture manuscrite n'offre toujours pas des critères de rapidité et de robustesse industrielles, alors que la plupart des documents passant chaque jour entre nos mains sont manuscrits. Cet enjeu industriel correspond actuellement au traitement électronique du flux de documents papier entrant dans une entreprise.

La numérisation d'un maximum de documents avec une bonne résolution apparaît aujourd'hui comme une évidence économique pour nombre d'entreprises ou services publics, car une fois le document numérisé tout devient possible : la sauvegarde du document original, le rapprochement visuel de documents épars, la consultation d'un même document par plusieurs personnes ou le traitement automatique de formulaires.

La sauvegarde devra concerner avant tout les documents les plus fragiles. La

1. Optical Recognition Character ou reconnaissance optique de caractère

consultation d'un même document par plusieurs personnes peut se faire via les serveurs des institutions scientifiques, des universités ou des systèmes aujourd'hui grand public (Google etc.).

De plus, contrairement au lecteur de microfilm dont les ajustements, même sur les meilleures machines, demeurent assez rudimentaires, la numérisation permet de retravailler l'image, c'est-à-dire d'éclaircir les zones ombragées, d'ajuster le contraste et la luminosité, non seulement pour l'ensemble de la page, mais aussi pour n'importe quelle section de cette dernière, d'augmenter le nombre de pixel par pouce (DPI) et surtout de modifier à sa guise la taille de l'image en fonction de ses besoins (impression en format A3, A4 ou encore page Web).

L'obstacle principal, à l'heure actuelle, est encore l'espace de mémoire nécessaire. Ces gros fichiers posent principalement le problème de la distribution en ligne.

La création de fonds d'archives numériques et leur déploiement via des technologies d'accès rapides (IP, ATM) dans des contextes normalisés (MPEG-4/7, CORBA) ont motivé de nombreux projets européens, qu'il s'agisse :

- d'archives historiques précieuses : ESPRIT-4 VENIVA (VENetIan Virtual Archive), INCO SINAMMA (Serveur d'Images Numerisées d'Archives de la Mediterranée et du Monde Arabe) ;
- d'archives muséographiques : RACE-2 RAMA (Remote Access to Museum Archives), IMPACT-1 NARCISSE (Network of Art Research Computer Image SystemS in Europe).

Une orientation forte des récents programmes de recherche concerne la définition de services (accès par le contenu et protection des contenus) au sein de systèmes dédiés à la gestion électronique de fonds d'archives. Dans ce contexte, l'accent est mis sur les archives audiovisuelles tant au niveau des programmes na-

tionaux (projets RNRT AGIR : développement d'un système d'indexation audiovisuelle et de recherche par le contenu ; programmes PRIAMM : Programme pour l'Innovation dans l'Audiovisuel, et le Multimédia du ministère de l'industrie dans le cadre du projet VISTéO : Système temps-réel intégrant vidéos et mondes virtuels et OPPIDUM : sécurité signature électronique, cryptographie, carte à puce etc.) qu'au niveau européen (projet ESPRIT-4 DIVAN ; programme Information Society Technologies, IST, du 5ème PCRD ; programme Information and Communication Technologies, ICT, du 7ème PCRD). Quelques initiatives existent cependant en ce qui concerne la conservation et la valorisation du patrimoine archivé, telles le projet ESPRIT-4 MENHIR (Multimedia European network of high quality image registration) dans le domaine muséologique.

La manipulation de tous ces documents va obligatoirement amener à la nécessité de manipuler et gérer en temps réel des bases de données de taille gigantesque. Ceci impose forcément des contraintes de temps de réponse ; des applications en ligne, destinées au grand public, se doivent de remplir leur rôle en quelques secondes. Or, comment donner une réponse fiable, précise, en un temps très faible ? D'ailleurs, un internaute qui interroge le système par curiosité, a t'il les mêmes attentes en termes de précision (performance en terme de taux de bonne reconnaissance, qualité de la décision) et de temps de réponse qu'un égyptologue qui travaille sur le document de sa vie ? Il y a de fortes chances que ce dernier accepte d'attendre quelques jours, si le système lui promet une quasi certitude quant au résultat.

C'est dans ce cadre que Philippe Bouyer a initié en 2002 le projet EUREKA COROC (Cognitive Optical Recognition for Old Characters), porté par la société RCSOFT installée à Angoulême. L'idée de ce projet était de monter un système dont la base serait capable d'apprendre à reconnaître n'importe quel type de caractère, des hiéroglyphes égyptiens aux écritures cursives des actes notariés antérieurs au XXème siècle.

Le projet COROC, tel qu'il a été écrit au départ, devait répondre aux contraintes suivantes :

- Adaptabilité : choix *a priori* d'une qualité de réponse pour un coût (temps) donné. L'utilisateur peut choisir, avant le début des traitements, la précision de la reconnaissance (i.e. 90% des caractères, 75% de tous les mots de la page, 100% des mots clé du texte etc.)

- Dynamique : le système est capable de détecter les cas nouveaux et de les mémoriser. Une première série d'enchaînements du système est effectuée hors ligne, *a priori*, à partir d'une analyse statistique d'un certain nombre de cas connus servant de base d'apprentissage. En phase d'utilisation (dite en ligne), nous devons nous attendre à voir apparaître des cas inconnus, des cas particuliers pour lesquels la connaissance du système est insuffisante. A ce moment là, le système devra être capable de détecter ces cas, de les marquer comme "inédits", de les mémoriser pour analyse future et de proposer une réponse immédiate même non optimale

- Autonomie : le système propose en ligne des solutions pré-établies, mais travaille hors-ligne à découvrir les solutions inédites. À la suite du point précédent, une fois un cas inédit détecté et mémorisé, le système se doit encore de découvrir comment résoudre ce cas même s'il a déjà proposé à l'utilisateur une solution immédiate, car cette solution ne sera pas forcément idéale. Hors ligne, le système peut également reprendre des cas déjà résolus et tenter de découvrir de nouvelles solutions encore plus optimales, soit en temps, soit en résultat

- Ouverture : apprendre de nouveaux caractères, de nouvelles langues. Les premiers types de caractères appris par le système sont le grec ancien et les

hiéroglyphes grecs. Une solution de reconnaissance de mots dans les actes notariés français des XVIéme et XVIIème siècles a été ajoutée par la suite. Face au grand nombre d'applications possibles liées au document (reconnaissance de caractères, de mots, de structure, de contenu etc.), le système se doit d'être le plus générique possible

Nous parlons d'analyse de document, de reconnaissance de caractères, et ce genre de système nécessite une version numérique du document (scanner, photographie...) afin de le traiter. Or, un document numérisé ne devient pour le système rien d'autre qu'une image dont il faut analyser le contenu. Nous pouvons donc dire qu'une application d'analyse de documents manuscrits numérisés est une application de traitement d'image au même titre que, par exemple, un outil d'aide à la décision médicale par analyse de radiographies numériques. Nous en arrivons donc aux problèmes liés à la vision artificielle.

1.1.2 Image

La vision industrielle est l'application de la vision assistée par ordinateur aux domaines industriels de production et de recherche. Les productions de masse à haute cadence, le souci constant d'amélioration de la Qualité et la recherche de gain économique poussent de plus en plus les industriels à automatiser les moyens de production. La vision industrielle est une réponse à ces préoccupations pour les opérations de contrôles de la production. En effet les machines de vision industrielle permettent un contrôle de la production à haute cadence et assurent une bonne répétabilité du contrôle (à la différence d'un opérateur, une machine n'est jamais fatiguée et ses critères de décision ne varient pas). Une autre utilisation est la gestion des flux d'objets. Par exemple la lecture optique d'un code à barres ou d'une adresse postale sur un colis pour l'orienter dans un centre de tri. Ou encore le tri de pommes par couleurs différentes avant emballage. Cela peut enfin être un

moyen de guidage pour un système mobile autonome (comme un robot) lorsque ses mouvements ne peuvent pas être déterminés par avance, comme par exemple la préhension d'objets sur un tapis roulant. Une caméra est alors embarquée sur la tête du robot et permet le positionnement de celui-ci au point désiré.

Cette discipline s'appuie essentiellement sur la reconnaissance de formes ou de caractères. Or, une image est un objet qui contient une grande quantité d'informations (formes, couleurs etc.) selon le niveau d'analyse : l'image par elle même, chacun de ses objets ou chacun de ses pixels. Dans notre optique d'instrumentation d'une partie de la vision humaine, nous sommes amenés à nous poser la question de l'automatisation des traitements de l'image. Il existe aujourd'hui un grand nombre d'opérations informatiques applicables à une image. Le problème est que celles-ci sont dépendantes du but recherché ; rehausser un contraste ne mettra pas en jeu les mêmes actions que l'extraction des contours des objets. Il en va de même pour l'extraction des formes contenues dans l'image ou bien encore, tout simplement, le débruitage de celle-ci. Le choix des opérateurs, ainsi que leur séquencement va donc dépendre des données disponibles ainsi que du but recherché.

Ces questions ne se posent pas uniquement dans le cadre de la reconnaissance de formes et de caractères : toute application de vision artificielle pose ces interrogations. Il existe en effet bon nombre de méthodes, d'approches, d'algorithmes etc. qui permettent aux experts de ce domaine de développer des applications spécifiques à un problème ou un champ d'application. L'inconvénient est que de part la complexité de l'objet image, et la spécificité des applications, il faut, la plupart du temps, pour chaque nouveau problème redévelopper une application entière.

1.1.3 Chaîne de traitement

Les travaux présentés dans ce document sont basés sur l'hypothèse selon laquelle toute application de traitement d'image peut être représentée comme une chaîne d'opérateurs unitaires de traitement d'image. Parmi ces opérateurs nous retrouvons toutes les étapes de pré-traitement (correction de l'illumination de l'image, de son orientation, de son échelle etc.), de débruitage (image couleur, niveaux de gris ou noir et blanc), de concentration des données dites utiles (seuillage, binarisation etc.), de segmentation, de calculs d'attributs sur les éléments issus de la segmentation le plus souvent suivies d'une étape de classification et de décision. Nos précédents travaux sur l'identification des taches de traitement d'image, présentés en 2003 en fin de Master[78], nous ont amenés à une représentation de ces enchaînements illustrée par la figure 1.1). Le débruitage est représenté en dehors des opérations de pré-traitement de part sa nature complexe. En effet, il est tout à fait possible d'enchaîner les opérations (débruitage d'une image couleur, conversion de l'image en niveaux de gris, débruitage d'une image en niveaux de gris, binarisation, débruitage d'une image binaire), ou de corriger différentes origines de bruit (bruit lié au document lui-même, bruit lié au système de numérisation...).

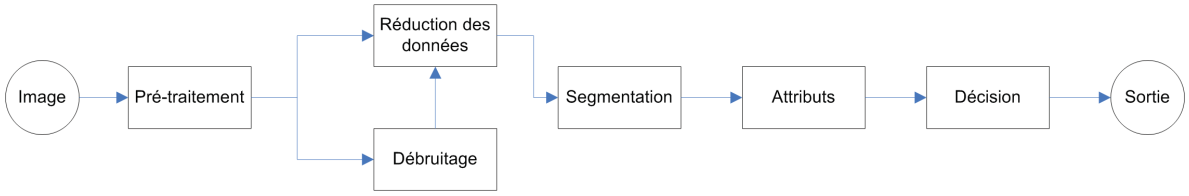

FIGURE 1.1 – Une application de traitement d'images vue sous forme de chaîne de traitements

Ainsi que nous l'avons énoncé, il faut dans la majorité des cas, développer intégralement une nouvelle application de traitement d'image pour chaque nouveau problème qui se pose, ce qui implique de refaire les enchaînements, avec de nouveaux opérateurs. Dans ce cas, certes les traitements sont différents, mais ils représentent fonctionnellement le même rôle ; un calcul d'attributs via les moments de Zernike se fera dans exactement le même but que via les moments de Fourier. Il

en va de même pour les moyens techniques de réaliser un débruitage (qu'importe la méthode, le but est d'enlever le bruit), une classification (qu'importe la méthode, le but est d'attribuer une classe aux attributs) etc. Si le schéma fonctionnel reste le même, pouvons-nous imaginer fournir au système des ensembles d'opérateurs, sortes de boîtes à outils par étape de la chaîne de traitements ? Ainsi, un système d'analyse des sorties des opérateurs et de bouclage entre les étapes apporterait une chaîne de traitements dynamique et adaptative, résolvant ainsi nombre de contraintes.

Face à la complexité du problème, et pour pouvoir respecter les délais imposés par un travail de Thèse, nous avons été obligés de contraindre notre problème à une partie précise de la chaîne de traitement de l'image. Notre hypothèse principale est que toute chaîne de traitement d'images devrait fonctionner de façon bouclée. La question principale portant dès lors sur le critère de sortie de la boucle de traitement. Selon un tel formalisme, le critère de bouclage doit porter sur les informations issues du traitement, soit la décision d'étiquetage typiquement. Face à cette question, il nous apparaît évident que pour justifier le bouclage, la qualité de la décision doit être estimée pour servir de critère. En s'arrêtant à cette question, les solutions sont nombreuses pour produire à la fois décision et estimation de la qualité de celle-ci, mais notre raisonnement induit que la qualité de la décision est liée à la qualité des données, la qualité des valeurs qui caractérisent ces données et la qualité des systèmes de décision ou d'étiquetage intermédiaires. Dès lors, rien n'interdit la mise en œuvre de différents niveaux de bouclage pour obtenir un système tel que décrit par Fayyad[27][26] dans ses travaux donnant lieu aux bases du "Data Mining" et du "Knowledge Discovery in Databases" (KDD), illustré par la figure 1.2.

Le modèle de Fayyad, adapté à notre problème, peut être représenté par la figure 1.3. Or, si l'approche totale de Fayyad est intellectuellement séduisante, elle paraît concrètement comme très difficile à mettre en œuvre du fait de l'as-

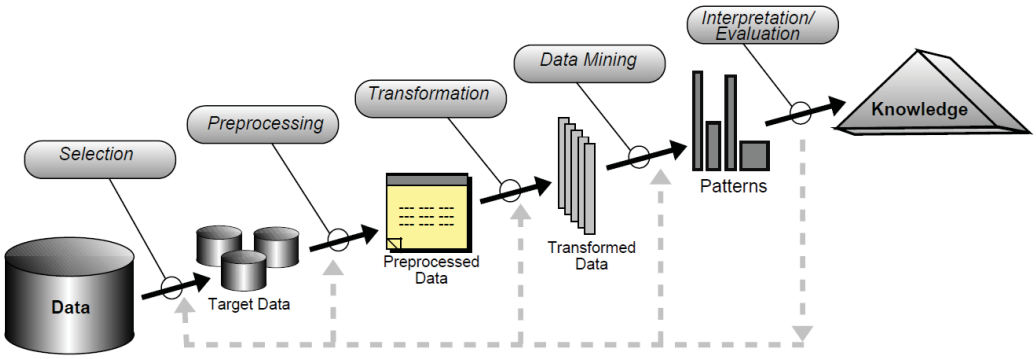

FIGURE 1.2 – Les étapes de l'extraction de la connaissance selon Fayyad[27]

pect fortement non linéaire des différents traitements. À l'opposé d'un système de traitement complètement automatique dans sa gestion interne, Régis Clouard[11] s'appuie sur la forte dépendance à l'expert pour la production d'une chaîne de traitement d'images.

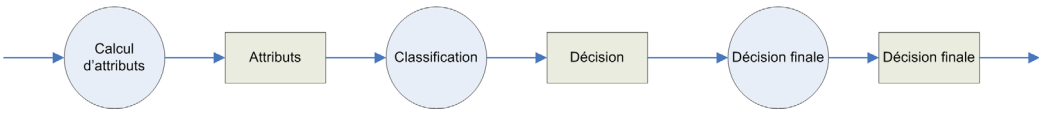

FIGURE 1.3 – Chaîne de traitement linéaire simple

Selon lui, cette discipline a atteint une certaine apogée, de part les méthodes développées mais manque actuellement de modélisations génériques et d'outils de construction automatique d'applications, d'où le recours quasi systématique aux experts en traitement d'images. Dans son modèle, la complexité de l'analyse réside dans le fait que les applications de traitement d'images sont toujours dédiées, et d'autre part que le choix des opérations est souvent si ce n'est toujours soumis aux choix de l'expert en traitement d'images. Sous ces deux contraintes, la production du système de traitement d'images réside dans la collaboration fine entre l'expert de l'application et l'expert en traitement d'images, mais induit deux problèmes :

– un système dédié peut se révéler très performant dans le cadre d'un do-

maine ou d'une application bien précise, mais il est souvent très difficile, voire impossible, de reprendre le même système et de l'adapter à une autre application (aspect objectif et dépendance au cadre applicatif)

– la suite d'opérations choisie par un expert peut être fondamentalement différente de celle qu'aurait choisie un autre expert mais les résultats peuvent être similaires.

Ces problèmes sont typiques des chaînes de décision linéaires. Bien évidement, les objectifs du projet COROC sont de pallier à ces problèmes en autorisant des bouclages intermédiaires (**aspect "feedback" classique de l'automatique**), une remise en cause d'un choix d'opérateurs dans la boucle (**aspect non linéaire**), ou encore d'essayer une autre combinaison d'opérateurs si le temps restant imposé par l'application le permet. Du fait de la nécessité de boucler sur n'importe quelle étape de la chaîne, à n'importe quel moment, notre approche sera donc forcément dynamique et non linéaire.

Même si d'un point de vue conceptuel cette chaîne non linéaire et ses bouclages semble apporter des réponses à notre cahier des charges, sa réalisation semble très complexe ; comment faire en sorte que le système "décide" de revenir sur un traitement pour en essayer un autre ? La littérature apporte des bases à partir de méthodes de type statistique, théorie de la décision, logique floue ou encore théorie de l'évidence. Nous verrons plus loin que cette dernière aura l'avantage de prendre en compte une incertitude établie au préalable, c'est à dire un formalisme intégrant le manque d'information ou son incomplétude.

Dés lors le cadre formel de notre travail est établi et implique la sélection dynamique des opérateurs de calcul. Nous développerons, pour le montrer, le cas de la sélection dynamique des opérateurs de calcul d'attributs caractéristiques. Comme indiqué précédemment, les règles de sélection ou de bouclage doivent être établies

à partir de critères qualité estimés au moment du calcul de l'attribut et au moment de la prise de décision (figure 1.4).

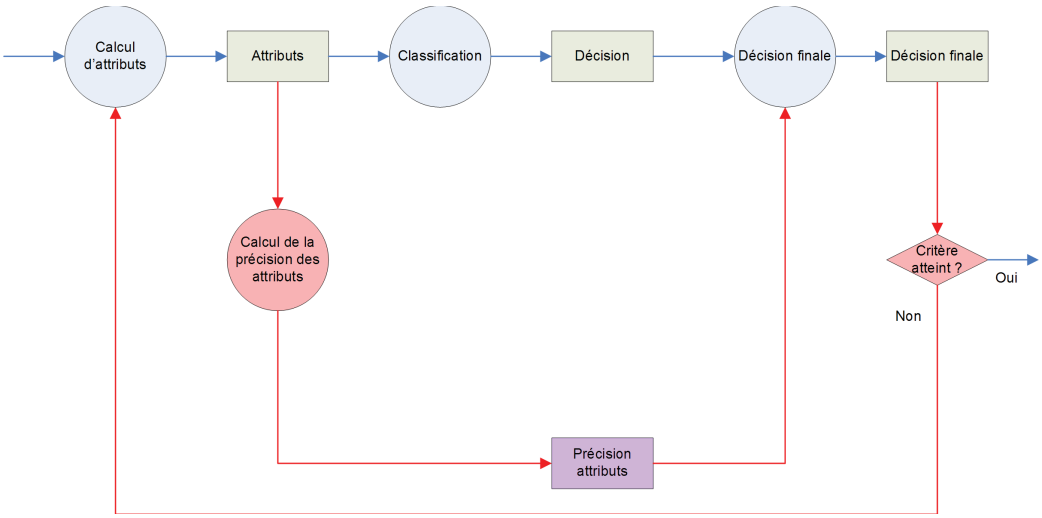

FIGURE 1.4 – Chaîne de traitement intégrant la précision des attributs

L'approche la plus classique dans la conception d'une application de traitement d'image, et de choisir une fois pour toutes, *a priori*, une seule méthode de classification. Or, et nous le verrons dans ce document, il est plus intéressant en terme de performance (nombre moyen de bonnes classification) de combiner les méthodes de classification, permettant ainsi de compenser les erreurs de chacune des méthodes. Il existe plusieurs façons d'aborder la combinaison de classifieurs, dont une dynamique que nous développerons par la suite, à partir du calcul de la précision locale de chaque classifieur. La précision locale d'un classifieur, telle que définie par Woods[87] et Giacinto[30], est une mesure sur la décision locale de cet opérateur. Il est possible d'intégrer cette mesure en cours de traitement. De cette manière, nous obtenons une chaîne de traitements schématisée par la figure 1.5.

Que ce soit pour le calcul des attributs, ou la classification, la décision de garder, ou non, le résultat d'un traitement, se prend à partir d'une mesure de la qualité

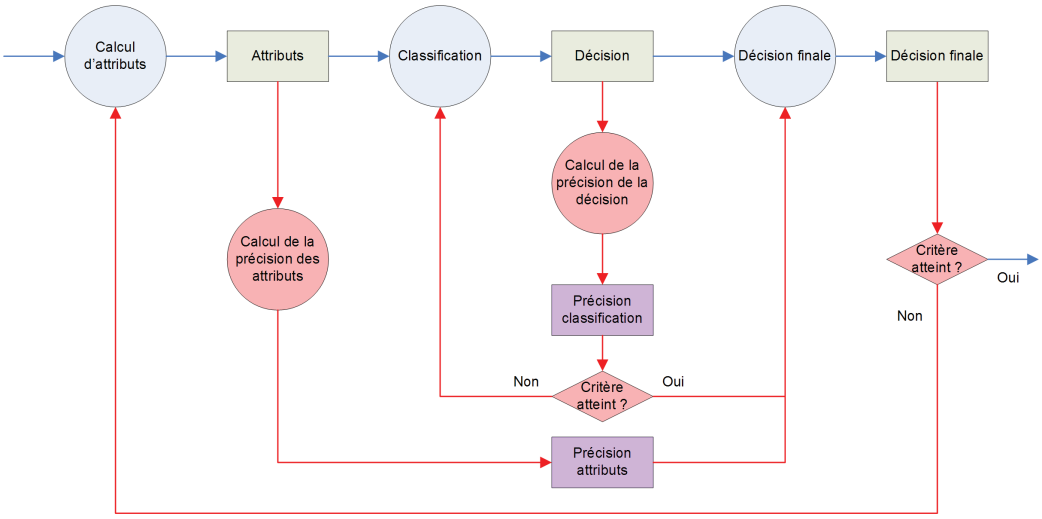

FIGURE 1.5 – Chaîne de traitement intégrant la précision des attributs et de la classification

des données issues de ce traitement. L'impact de ces données sur la décision finale, est essentiellement lié à sa qualité, à sa précision. C'est pour cela que dans le cadre de ce travail, nous avons décidé de parler de la **précision des données en sortie des opérateurs**.

L'une des caractéristiques visées du projet COROC était son adaptabilité. Le choix laissé à l'utilisateur devant se limiter aux réglages d'un paramètre liant le temps de calcul à la qualité de la décision, étant admis que l'accroissement de cette dernière induit un coût combinatoire qui peut être très important.

1.2 Introduction au travail réalisé et au document

L'objectif de ce projet est d'arriver à un prototype d'application de reconnaissance de caractères par sélection dynamique d'opérateurs simples de traitement de l'image, plus précisément, une application composée uniquement d'opérateurs de calcul d'attributs et de classification. Pour savoir si l'objectif est atteint, il faut

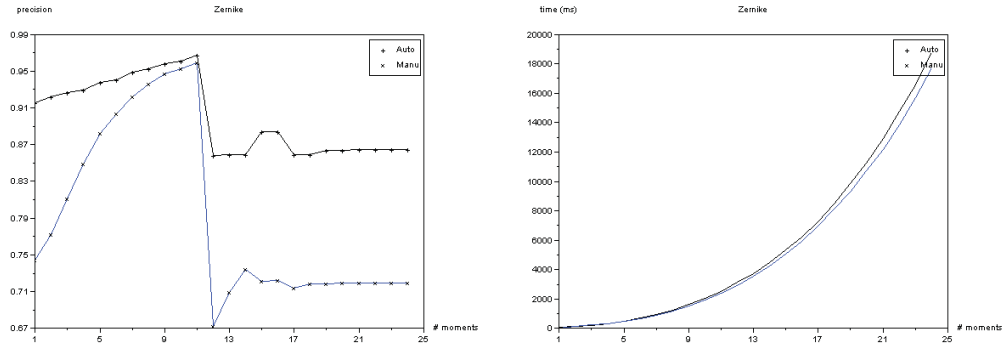

FIGURE 1.6 – Évolution de la précision et du temps de calcul en fonction de l'ordre pour les moments de Zernike

mesurer le résultat obtenu et analyser sa distance au résultat escompté. Nous nous baserons sur les éléments suivants :

– Avons-nous produit une application de reconnaissance de caractères ?

– Cette application se construit-elle dynamiquement par sélection dynamique des opérateurs disponibles parmi une bibliothèque d'opérateurs simples ?

– Les performances en terme de taux de bonne reconnaissance sont elles meilleures que celles obtenues par le meilleur enchaînement statique des opérateurs utilisés ?

– Ce travail apporte t'il une base de réflexion théorique au problème ?

Avant de pouvoir répondre à ces question, il nous faut savoir sur quoi se baser pour comparer les approches, c'est pourquoi nous commencerons par présenter les données utilisées pour nos tests. Nous avons besoin de plusieurs sources de données pour pouvoir tester les mesures de précision sur les sorties des classifieurs, d'autre types de données pour analyser les mesures de précision sur les

opérateurs de calcul d'attributs et enfin de données s'inscrivant dans le cadre du projet COROC, son champ d'exploitation étant bien spécifique. Toutes les bases de test seront présentées avec leur origine, les raisons de leur sélection ainsi que des informations statistiques sur les individus les composant.

Le cœur du problème étant la décision, qu'elle soit ponctuelle sur le choix d'un opérateur, ou finale. Nous enchaînerons donc sur les outils permettant la réalisation de cette partie, en présentant ce qui existe actuellement dans la littérature, notre propre proposition et les résultats obtenus sur de la reconnaissance de caractères. C'est au cours de ce chapitre que nous verrons à partir de quoi réaliser une chaîne de traitements avec sélection dynamique des opérateurs.

Le schéma de décision étant spécifié en fonction de nos contraintes, nous pourrons formaliser ce qui est attendu comme informations pour exercer ce schéma. Les deux chapitres suivants traiteront de la mesure de la précision des opérateurs de calcul d'attributs, plus précisément les descripteurs de formes utilisés en reconnaissance de caractères, et des classifieurs. Ce dernier point se fera en présentant le principe des systèmes à classifieurs multiples, et en analysant ce que propose la littérature en matière de construction de tels systèmes, de façon statique ou dynamique, grâce à des mesures de qualité ou de précision des classifieurs.

À ce moment enfin, ayant à la fois présenté les données de tests, notre proposition quant au système de sélection dynamique d'opérateur et les critères sur lesquels se base cette sélection, nous pourrons conclure sur les résultats obtenus en théorie et pratiquement sur les données du projet COROC, la réalisation de nos objectifs, et les pistes de recherche induites par notre travail.

"*Le scepticisme est la plus facile des philosophies.*"
Robert Kemp

Chapitre 2

Les données de validation du propos

Sommaire

2.1 Pourquoi ?

La société RCSOFT s'est intéressée aux applications liées à l'analyse numérique de documents manuscrits, et notamment à la reconnaissance de caractères anciens (grec, hiéroglyphes égyptiens).

Le travail présenté dans ce document se concentre essentiellement sur le calcul des attributs (représenter l'image du caractère en cours d'analyse par une série de données numériques) et la classification de ces attributs (les attributs analysés ressemblent-ils plus à des attributs du caractère α ou plutôt à ceux du caractère γ ?).

Il existe dans la littérature, de nombreux attributs utilisés en reconnaissance de caractères, ainsi que plusieurs outils de classification. Notre propos n'est pas de fixer a priori quels attributs et quels classifieurs vont être utilisés. Au contraire, l'idée est de fournir au système une boîte à outils de méthodes pour chacune de ces étapes. Le choix des méthodes (opérateurs) les plus adaptées aux contraintes posées par l'utilisation (le critère coût/précision) étant opéré de manière totalement dynamique. Pour effectuer cette sélection, plusieurs paramètres sont disponibles : la connaissance sur les traitements disponibles, l'attente de l'utilisateur en terme de qualité du résultat en fonction du coût en temps de calcul etc.

Les données manipulées sont enfin très particulières. Il s'agit de symboles ou de caractères manuscrits. Nous savons tous par expérience qu'il est très rare, voire improbable, de rencontrer deux écritures strictement identiques. Nous pouvons également constater au quotidien, qu'un même scripteur peut très difficilement reproduire deux fois de suite exactement le même dessin, particulièrement dans le cadre d'une écriture rapide, non appliquée.

Au final, notre système produira des résultats, qu'ils soient intermédiaires ou

non. Nous allons donc mesurer ces résultats afin de nous situer par rapport aux contraintes fixées, c'est à dire les questions de la sélection dynamique d'opérateurs et du bouclage. Nous allons également nous baser sur des méthodes existantes dans la littérature, dont les résultats sont connus. C'est pourquoi la mesure de ce que nous allons obtenir doit apporter deux types d'information. Dans un premier temps, nous devons savoir si les méthodes choisies nous permettent d'atteindre nos objectifs (temps de calcul, taux de réussite etc.). Dans un second temps, nous devons également comparer ces résultats par rapport à ceux obtenus par les méthodes déjà existantes, pour savoir si nos choix apportent un gain ou non.

Pour cela, nous devons disposer d'un ensemble de données, que nous appelons bases de tests, afin de tester et mesurer l'ensemble de nos traitements. Ces bases devront évidement représenter la complexité et la variabilité des données que le système rencontrera en utilisation réelle. De plus, nous devrons disposer à la fois d'images pour travailler sur le calcul des attributs, mais également de données statistiques pour la classification et la prise de décision. Pour la réalisation, nous devons donc répondre aux exigences suivantes :

- reprendre des bases de données existantes pour confronter nos résultats à ceux obtenus par la communauté du traitement d'images,
- avoir des bases de caractères d'origines différentes (artificielles ou naturelles) afin de tester l'impact de l'origine du scripteur sur les résultats,
- posséder des bases de données de caractères en corrélation avec le contexte de ce travail,
- disposer de bases de données de symboles pour les tests validant l'origine du caractères,
- fournir des bases de données de mesures quelconques pour valider les méthodes statistiques.

En ce qui concerne la classification et la prise de décision, nous pouvons tester ces parties seules, en s'affranchissant du contexte lié à la reconnaissance de carac-

tères. En effet, ces opérations étant basées sur de l'analyse statistique de données, il nous suffit de concevoir une base de données "directes". Ces données seront en partie générées artificiellement afin de maîtriser un ensemble de paramètres (comme la répartition des données), pour valider le propos. Une autre partie sera composée de données naturelles, afin de tester le comportement des opérations dans un cadre réel.

Ensuite, nous travaillerons sur une base d'images à partir desquelles seront calculées les attributs, pour valider notre propos sur les opérations de calcul d'attributs et sur la décision finale en fin de chaîne. Là encore, ces données seront en parties générées artificiellement, et en partie d'origine naturelle, afin d'étudier l'influence du scripteur et la variabilité des données sur les résultats. Par rapport au contexte, nous disposerons d'une base de caractères grecs anciens. Pour ce qui concerne la validation scientifique du propos, nous intégrons aussi d'autres bases de données liées au problème de l'analyse de document numérisé, mais déjà utilisées par la communauté, et sur lesquels nous possédons des résultats auxquels nous pouvons nous comparer.

2.2 Bases à données directes

L'ensemble des traitements de classification ne nécessitent pas obligatoirement de données images, proches de celles rencontrées en utilisation réelle du système. En effet, ces opérateurs se basant sur des éléments statistiques, des jeux de données brutes, directes, suffisent amplement. L'intérêt de ces données est d'être indépendants des artefacts liés à leur extraction.

2.2.1 Données artificielles

Afin de comprendre les résultats obtenus, le comportement des opérateurs en fonction de certains paramètres, et étant donné le caractère statistique des méthodes de classification, il apparaît important de travailler sur des jeux de données dont les caractéristiques sont maîtrisées. Parmi les notions que nous souhaitons voir contrôler, se situe la répartition des individus dans l'espace des attributs (les individus sont-ils répartis équitablement dans tout l'espace ?), l'existence de regroupements permettant la définition de classes, la forme de ces classes, etc.

Dans le cadre de ces données artificielles, issues de générateurs aléatoires, nous utiliserons quatre types de générateurs nommés Banana, Complex, Difficult et Simple, avec pour chacun 250 individus par classe, dans un espace à deux dimensions (voir la figure 2.1), selon les règles suivantes :

– Jeu de données Banana : Génération d'un jeu de données à deux classes et deux dimensions selon une distribution en forme de "banane". Les données sont uniformément distribuées le long des régions en forme de "banane" et superposées selon une distribution normale avec un écart-type de 1 dans toutes les directions, les deux classes étant equiprobables.

– Jeu de données Complex : Génération d'un jeu de données à deux classes concentriques dans un espace à deux dimensions. Chaque classe possède une distribution Gaussienne sphérique. Les deux classes sont centrées autour du même point, mais présentent une covariance différente. La covariance de la première classe correspond à une matrice identité ce qui induit un nuage sphérique très concentré autour du centre. Par opposition, la seconde classe est distribuée uniformément dans un cercle de plus grand diamètre (rapport 4). Les deux classes sont équiprobables.

– Jeu de données Difficult : Génération d'un jeu de données à deux classes

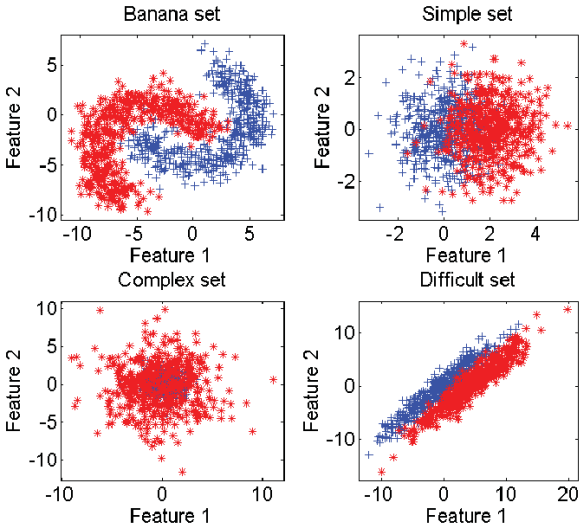

FIGURE 2.1 – Nuages d'individus pour les 4 générateurs artificiels utilisés

équiprobables dans un espace à deux dimensions. Les variances sont fortement différentes, la séparation étant ainsi, pour les échantillons de faible taille, "difficile". La différence entre les moyennes pour la première dimension est de 3, de même pour la différence entre les moyennes de la seconde dimension. Les deux matrices de covariance sont égales, avec une variance de 1 pour la première dimension, et 40 pour la seconde. Les 2 dimensions subissent une rotation de 45 degrés afin de construire une forte corrélation.

– Jeu de données Simple : Génération d'un jeu de données à deux classes équiprobables dans un espace à deux dimensions. Les nuages obtenus sont de forme sphérique et de même rayon (matrice de covariance de type identitaire). Comme les centres de ces nuages sont à une distance unitaire, les deux nuages s'intersectent, induisant des difficultés d'étiquetage.

L'ensemble des données générées, pour chaque type de problème, possède les caractéristiques statistiques globales présentées en table 2.1.

Données	Distance moyenne	Écart-type des distances
Banana set	$66,7$	4.10^3
Complex set	$71,13$	$4,4.10^3$
Difficult set	$68,14$	$3,9.10^3$
Simple set	$65,7$	$3,8.10^3$

TABLE 2.1 – Distance moyenne et écart-type des distances entre individus pour les jeux de données artificielles

Ces différentes configurations spatiales (différents types de classes), ainsi que la maîtrise de leurs paramètres (moyenne, variance, forme) servent à valider notre propos sur un cadre théorique, ainsi qu'à analyser et expliquer le comportement de nos outils. Cependant, ces données artificielles possèdent un coté "parfait" (notamment leur régularité dans la répartition des individus) qui se retrouve rarement dans le cadre de données "naturelles". Typiquement, l'écriture manuscrite correspond au cas où la variabilité dans la représentation est très forte mais surtout où l'equiprobabilité n'est jamais respectée, ni l'uniformité des distributions. C'est pour cela que nous devons également nous construire des bases de données naturelles, pour analyser le comportement des outils de classification sur des cas réels.

2.2.2 Données naturelles

En plus de l'analyse du comportement des opérateurs de classification sur des cas réels, nous devrons comparer nos résultats avec ceux déjà existants dans la littérature. Pour ce faire, nous devons constituer des bases de données issues de cas réels, mais également utilisées par la communauté. Ainsi, nous avons choisi cinq types de problèmes issus de la base communautaire *UCI Repository of machine learning databases* [1]. Chacun de ces problèmes comporte deux classes, et leurs

1. http ://www.ics.uci.edu/∼mlearn/MLRepository.html

caractéristiques sont décrites en table 2.2.

Les cinq problèmes utilisés, et leurs caractéristiques, sont les suivants :

– Pima Indians Diabetes : Le diagnostique, variable binaire étudiée, indique si le patient montre des signes de diabète selon les critères du World Health Organization. La population habite près de Phoenix en Arizona aux USA. Leur algorithme ADAP produit une valeur de prédiction réelle comprise entre 0 et 1, transformée en décision binaire en utilisant un seuil de 0.448. Plusieurs contraintes ont été placées sur la sélection des individus à partir d'une base de données plus importante. En particulier, tous les patients ici sont de sexe féminin, au moins âgés de 21 ans et descendants des indiens Pima. La base utilisée contient 768 individus et présente 8 attributs :

  1. Nombre de grossesses ;
  2. Concentration en glucose à 2 heures (test oral de tolérance au glucose) ;
  3. Pression diastolique (mm Hg) ;
  4. Epaisseur de pli du triceps (mm) ;
  5. Taux d'insuline à 2 heures (μU/ml) ;
  6. Indice de masse corporelle (poids en kg/(taille en m)^2) ;
  7. Fonction de pedigree diabétique ;
  8. Age (années).

  La distribution des classes est la suivante (la classe 1 étant interprétée comme "testée positive au diabète") :

Classe	Nombre d'individus
0	500
1	268

  Les attributs sont répartis selon les statistiques suivantes :

Attribut	Moyenne	Écart-type
1	3.8	3.4
2	120.9	32.0
3	69.1	19.4
4	20.5	16.0
5	79.8	115.2
6	32.0	7.9
7	0.5	0.3
8	33.2	11.8

– Wisconsin breast cancer : Ce jeu de données illustre le problème du diagnostique du cancer du sein dans une population sélectionnée géographiquement. Il est composé de 201 individus pour une classe et 85 pour l'autre. Chacun d'eux est décrit par 9 attributs linéaires correspondants à :

1. tranches d'âge : 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99 ;
2. ménopause (-40ans, +40ans, pré-ménopause) : lt40, ge40, premeno ;
3. tranches de taille de la tumeur : 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 ;
4. inv-nodes : 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39 ;
5. node-caps : yes, no ;
6. degré de malignité : 1, 2, 3 ;
7. sein : left, right ;
8. position sur le sein : left-up, left-low, right-up, right-low, central ;
9. rayons : yes, no.

– Sonar (Mines et Rochers) : le but est de différencier les signaux sonar renvoyés par des cylindres métalliques de ceux renvoyés par des rochers cylindriques bruts. La classe des mines est une collection de 111 signaux obtenus

par signaux sonar envoyés sur des cylindres métalliques à différents angles et sous différentes conditions. La classe des rochers est composée de 97 signaux renvoyés par des rochers dans des conditions similaires. Chaque individu est décrit par un vecteur de 60 éléments réels compris entre 0 et 1.

– Ionosphère : Ces données radar on été collectées à Goose Bay au Labrador par un groupe de 16 antennes haute fréquence ciblant les électrons libres de l'ionosphère. Une classe est donc le groupe de retours montrant l'existence d'une forme de structure dans l'ionosphère, l'autre regroupant les signaux qui sont passés à travers l'ionosphère. Les signaux reçus on été traités par autocorrélation. Pour les 351 mesures, il y a 17 impulsions par individu, dont les valeurs complexes (2 attributs par impulsion) forment le vecteur d'attributs de chaque échantillon. Les données sont donc décrites par 34 attributs réels.

Nom	Nb individus classe 1	Nb individus classe 2	Dimension
Wisconsin breast cancer	458	241	10
Ionosphere	225	126	34
Pima Indians diabetes	500	268	8
Sonar	97	111	60
German	700	300	24

TABLE 2.2 – Caractéristiques des données naturelles

L'ensemble de ces données naturelles possède les caractéristiques statistiques présentées en table 2.3.

A présent nous possédons à la fois des bases de données artificielles, ainsi que des bases de données naturelles. Pour ces dernières, nous utilisons des données issues de bases utilisées par la communauté pour la confrontation des résultats. Nous avons donc toutes les données nécessaires pour valider notre propos quant

Données	Distance moyenne	Écart-type des distances
Wisconsin breast cancer	$7,62.10^{11}$	$7,53.10^{25}$
Pima Indians diabetes	$3,03.10^{4}$	4.10^{9}
German	$2,17.10^{3}$	$1,21.10^{7}$
Ionosphere	$18,99$	$164,50$
Sonar	$3,99$	$3,71$

TABLE 2.3 – Distance moyenne et écart-type des distances entre individus issus de données naturelles

aux opérateurs de classification, en utilisant des données directes. Voyons à présent les données utilisées pour les autres parties de notre travail.

2.3 Données au travers des attributs

Les deux bases de données précédentes sont indépendantes de toute application de traitement d'images et sont notamment utilisées dans le cadre de la confrontation des algorithmes d'apprentissage et de classification. Pour prendre en compte la spécificité du traitement d'images, nous avons utilisé différentes bases d'images pour lesquelles l'extraction des informations utiles est au cœur du problème. Dés lors en amont des étages de classification puis de décision, il faut intégrer les étapes de numérisation, de prétraitement puis d'extraction des attributs censés porter les informations caractéristiques de l'image. Afin d'intégrer dans notre validation ces problématiques, nous avons conçu ou utilisé différentes bases d'images liées à l'écrit. Dans le cadre du projet COROC notamment, la société RCSOFT s'est constitué une base d'images de caractères et de symboles manuscrits, dont un extrait est visible sur la figure 2.2.

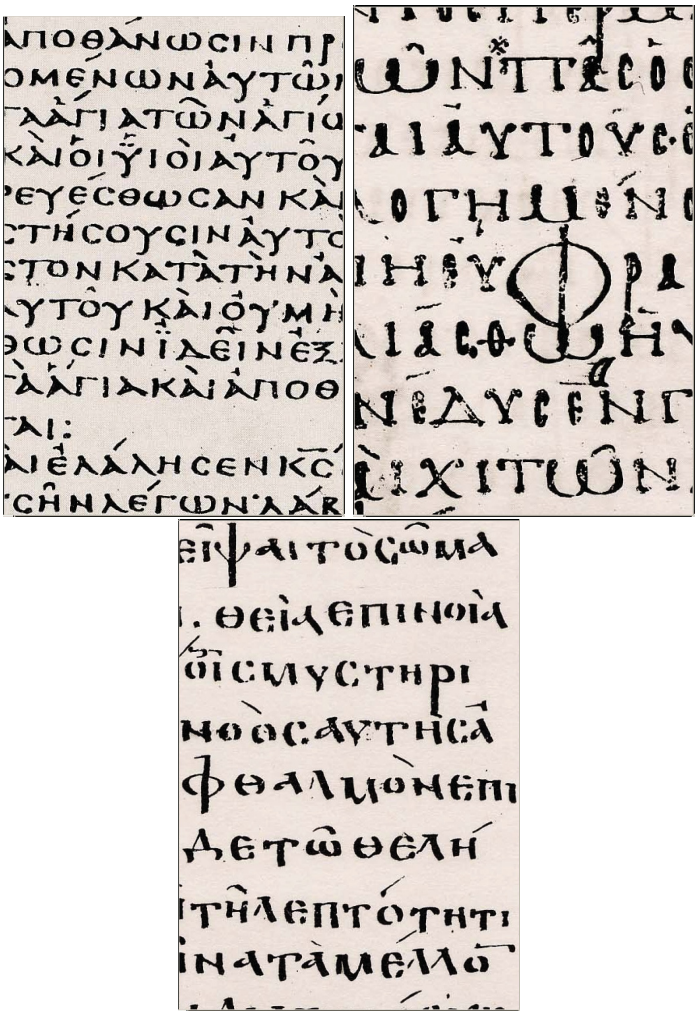

FIGURE 2.2 – Extrait de la base RCSOFT originale

2.3.1 Notre base de caractères

Le projet COROC concernant en partie la reconnaissance de l'écriture manuscrite en grec ancien, il nous fallait constituer une base d'apprentissage et de tests pour entraîner et valider le système. Une partie de cette base comprend donc des caractères minuscules grecs manuscrits. Or, et comme nous pouvons le constater sur la figure 2.2, les documents numérisés disponibles sont très complexes, du fait de l'agencement des caractères, de la façon parfois incomplète ou détachée dont ils sont dessinés, ou bien encore de la présence d'accents ou de ponctuation. Pour ramener le problème au cadre de la Thèse en se consacrant uniquement aux taches de calcul d'attributs et de classification, il nous a fallu nous affranchir des étapes de débruitage, de reconstitution des caractères etc.

Pour cela, une nouvelle base de caractères a été créée de toutes pièces. Ces caractères ont été écrits sur des pages blanches avec un feutre noir par quatre scripteurs différents et l'acquisition a été effectuée avec un scanner grand public à 300 dpi. Les images étant de bonne qualité, les prétraitements se sont résumés à une binarisation manuelle, la segmentation ayant été réalisée par une analyse en composantes connexes avec calcul de boites englobantes [2]. La population de cette base de données est décrite en table 2.4. Un exemple de cette base est illustré en figure 2.3.

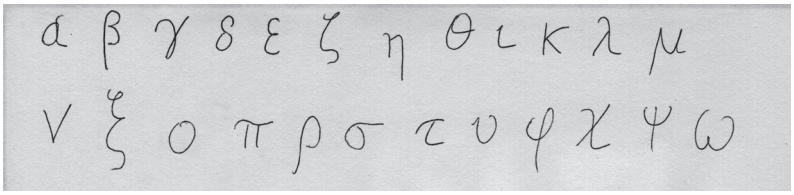

FIGURE 2.3 – Extrait de la base RCSOFT - caractères manuscrits

En plus de posséder une version "naturelle" de chaque caractère, nous devons

2. Une zone rectangulaire de l'image englobant l'objet en cours d'analyse

Caractère	Nb individus	Caractère	Nb individus
alpha	176	nu	62
beta	94	omega	84
chi	47	omicron	87
delta	121	phi	69
dzeta	99	pi	75
epsilon	103	psi	66
eta	107	rho	84
gamma	117	sigma	100
iota	78	tau	66
kappa	63	theta	98
lambda	59	upsilon	67
mu	59	xi	60

TABLE 2.4 – Population de la base de caractères RCSOFT

en avoir une version ”artificielle”, afin d'identifier les comportement des opérateurs en fonction de caractéristiques issues de la variabilité des scripteurs. Chaque caractère manuscrit voit donc son équivalent généré artificiellement via un logiciel de traitement de texte, les différences entre individus étant simulées par un changement de type et de taille de fonte. Un exemple de caractères obtenus de cette façon est illustré en figure 2.4.

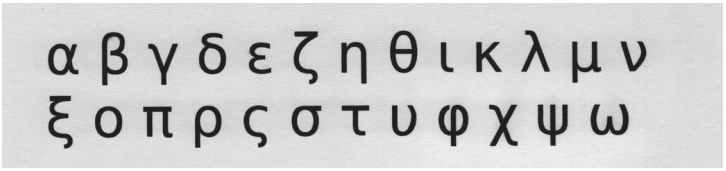

FIGURE 2.4 – Extrait de la base RCSOFT - caractères artificiels

Ainsi, par rapport au contexte du projet COROC, en constituant une base de caractères et de symboles tantôt manuscrits, tantôt générés, nous possédons suffisamment d'éléments pour simuler une utilisation réelle, et également analyser le comportement des opérateurs en fonction d'informations telles que l'origine du caractère (manuscrit ou artificiel), la forme du caractère (plutôt ”arrondi” ou plutôt ”carré”) ou encore la taille du trait. Il nous manque cependant des données

issues de bases utilisées par la communauté, nous permettant de confronter nos résultats.

2.3.2 GREC'05

La base d'images du projet COROC correspond pleinement aux objectifs visés. En revanche, elle ne permet pas de comparer les résultats obtenus avec ceux de la littérature. Nous avons donc également dû exploiter des bases particulières à cette fin de comparaison.

Parmi les bases correspondant à nos critères de sélection, la base GREC'05 [3] s'oriente vers la détection de symboles plus que de la détection de caractères. Si le choix peut paraître éloigné au premier abord, il doit être mis au regard des objectifs annoncés de RCSOFT face aux caractères anciens et notamment les hiéroglyphes.

Dans un travail amont à cette thèse, Denis Arrivault[2] a montré les difficultés induites par ces caractères à tous les niveaux de la chaîne de traitement. Ainsi, la base GREC, tout en étant orientée symbole, correspond à nos attentes en terme d'objectif applicatif et de validation du propos.

La base de données GREC contient 150 modèles de symboles, séparés en deux groupes : ceux issus du domaine architectural et ceux issus du monde de l'électronique. Dans les deux cas, les symboles sont composés uniquement d'arcs et de lignes droite.

Comme l'ont montré les travaux sur les hiéroglyphes, la notion d'arc est difficilement séparable de celle de ligne droite dans le cas de symboles manuscrits.

3. http ://symbcontestgrec05.loria.fr/sampletest.php

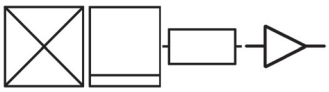

FIGURE 2.5 – Extrait de la base GREC'05

Pour étudier cette problématique, nous avons construit une extension à la base GREC, pour laquelle différents symboles ont été construits à la main (figure 2.6).

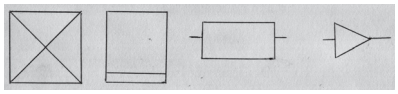

FIGURE 2.6 – Extrait de la base GREC/RCSOFT - symboles manuscrits

Au final, les deux bases GREC peuvent être exploitées de façon conjointe pour analyser le comportement des opérateurs en fonction d'informations telles que l'origine du caractère (manuscrit ou artificiel), la forme du caractère (plutôt "arrondi" ou plutôt "carré") ou encore la taille du trait.

2.3.3 MNIST

Puisque notre travail s'inscrit en partie dans le domaine de la reconnaissance de caractères (OCR), la base MNIST [4] est incontournable pour la confrontation des résultats. A l'heure actuelle, les taux de reconnaissance sur cette base atteignent les 100% avec des tailles d'attributs très importantes[51][50].

Ce travail s'est concentré sur la chaîne de traitement et s'est appuyée sur des attributs simples et bien connus pour des raisons didactiques dans la démarche. Le propos tenu s'appuie donc sur l'exploitation optimale des informations extraites et devra être comparé aux résultats obtenus à partir d'attributs équivalents. Néanmoins, en perspective de ce travail, nous reviendrons sur l'intérêt de l'approche

4. http ://yann.lecun.com/exdb/mnist/

face à de tels niveaux de résultats.

La base MNIST a été constituée pour l'évaluation des algorithmes de reconnaissance de chiffres. Elle comporte par conséquent dix classes pour chacun des chiffres arabes. En évaluation, l'apprentissage s'appuie sur une base de 60 000 individus pour une validation sur un ensemble de 10 000 individus. C'est un extrait d'une base beaucoup plus importante disponible près de NIST [5].

Les chiffres ont été normalisés en taille et centrés dans une image de taille fixe. L'image originale noir et blanc (deux niveaux) a d'abord été normalisée pour entrer dans une boîte de 20x20 pixels afin de conserver les proportions. L'image finale est en niveaux de gris à cause de la technique d'*anti-aliasing* employée par l'algorithme de normalisation. Les images sont toutes centrées dans une boîte de 28x28 par calcul du centre de gravité des pixels et translation de ce point jusqu'au centre de la boîte. La table 2.5 représente le nombre d'individus utilisés en apprentissage et en reconnaissance pour chacune des classes. Un extrait de cette base est illustré par la figure 2.7.

Caractère	Nb individus en apprentissage	Nb individus en reconnaissance
0	5923	980
1	6742	1135
2	5958	1032
3	6131	1010
4	5842	982
5	5421	892
6	5918	958
7	6265	1028
8	5851	974
9	5949	1009

TABLE 2.5 – Population de la base de données MNIST

5. National Institute of Standards and Technology

FIGURE 2.7 – Extrait de la base MNIST

2.4 Résumé sur les choix des différentes bases de validation

Notre propos intègre plusieurs niveaux de questions scientifiques, allant de l'extraction d'informations à partir d'images à la combinaison d'informations ou d'étiquettes pour produire une décision. Face à ces différentes problématiques, une démarche de validation méthodique a été mise en œuvre. Pour cela, différentes bases de test ont été conçues ou sélectionnées.

Un premier groupe de bases de données a été choisi pour valider les étages de classification et de décision indépendamment des étapes d'extraction d'informations de l'image. Deux bases sont exploitées, l'une dite artificielle contenant des données issues de générateurs aléatoires dont la complexité est contrôlée. L'autre, dite naturelle, contient des données issues de problèmes liés au vivant ou à la variabilité des systèmes naturels, donc de complexité différente de celle de la base artificielle.

Le second groupe de bases de données intègre l'aspect image du propos et s'appuie sur des bases de référence pour la communauté, les bases GREC et MNIST ainsi que des bases créées en fonction de l'objectif de RCSOFT telle la base COROC et la version manuscrite de la base GREC.

Ayant présenté les données utilisées pour la validation des différentes questions, nous allons maintenant présenter le cœur de notre propos, c'est à dire l'aspect prise de décision en fin de chaîne de traitement.

"*All that is necessary for the triumph of evil is that good men do nothing.*"
Edmund Burke

Chapitre 3

Décider en présence d'imprécision

Sommaire

3.1 Introduction

Comme énoncé plus haut, ce travail se situe dans le cadre d'un projet industriel porté par la société RCSOFT. Ce projet, appelé COROC, consiste en un système de reconnaissance de caractères manuscrits anciens, ouvert à l'apprentissage de nouvelles écritures, ou de nouvelles langues. COROC c'est aussi un système demandant à l'utilisateur de choisir son propre compromis entre la qualité du résultat (le taux de reconnaissance), et le coût en temps de calcul. D'une façon conceptuelle, cela ramène à voir une application de traitement d'images comme une chaîne de traitements (voir figure 3.1), où chaque traitement serait choisi dynamiquement par le système.

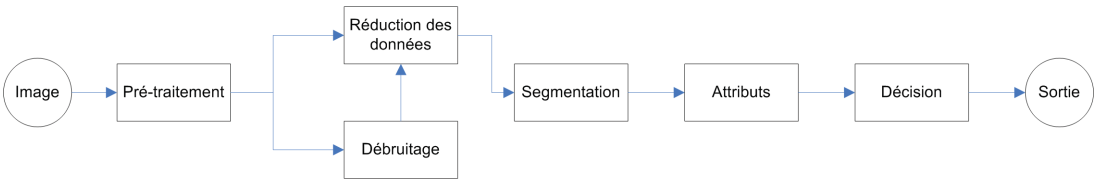

FIGURE 3.1 – Une application de traitement d'images vue sous forme de chaîne de traitements

Or, une question à laquelle nous devons répondre est : comment, en fin de chaîne, choisir le comportement à adopter ? Le système doit-il boucler pour essayer une nouvelle combinaison d'opérateurs ? Ou alors le système doit-il décider d'affecter le caractère en cours d'analyse à telle ou telle classe ? Ou encore, le système doit-il rejeter ce caractère, c'est à dire le déclarer comme n'appartenant à aucune classe connue ? Répondre à ces questions équivaut à décider en fonction d'un certain nombre d'informations, notamment la précision d'un résultat donné par un opérateur.

Mais comment décider en informatique ? Sur quoi se base un système automatique pour prendre une décision ? Bien souvent, ce genre de système est basé sur un principe de règles de décision (souvent probabilistes), ou encore sur une

approche pignistique (un niveau entièrement dédié à la prise de décision et clairement séparé de la modélisation des données).

Bien qu'une très grande partie des applications traitement d'images grand public soit tournée vers l'amélioration, la restauration ou le traitement des photographies ; dans le cadre industriel, les applications visées portent sur la prise de décision à partir des images. Si les classes d'information utilisées à cette fin sont connues (texture, forme, couleur), elles ne présentent pas toutes le même intérêt dans l'aide à la prise de décision. Dés lors, différentes formes d'apprentissage ont été développées pour identifier le meilleur assemblage possible des informations. Des dizaines de solutions existent pour répondre à cette question, certaines même combinent plusieurs de ces approches pour optimiser au maximum la prise de décision.

Néanmoins, si toutes les classes d'information, et de façon plus fine tous les opérateurs d'extraction d'information (moments de Zernike, attributs D'Haralick, pour n'en citer que deux), ne présentent pas le même intérêt pour la prise de décision, les informations extraites sont quant à elles jugées toutes de même qualité. Or, le traitement d'images provient pour partie du traitement du signal et par là même, du monde de la physique et plus particulièrement de la métrologie. Nous sommes donc bien loin d'un monde binaire, où l'information est vraie ou fausse mais certaine quant à son affectation. En métrologie, la mesure n'existe que liée à un critère de précision ou tolérance. Ce critère quantifie la plage de valeurs au sein de laquelle la mesure se situe avec une bonne certitude. Ce point de vue typique de la métrologie conduite avec des appareils à aiguille est bien souvent oublié lors du passage au numérique. Pourtant, tout scientifique reste sceptique lors de l'annonce d'un résultat avec plus de 3 chiffres derrière la virgule (lorsque cette précision n'a pas été justifiée au préalable).

A quoi peut ressembler une mesure de précision associée à un opérateur de

traitement d'images ? C'est la question à laquelle nous répondrons lors du chapitre 5. Néanmoins, pour comprendre la logique du présent chapitre, nous pouvons établir les attendus de cette mesure. Parce que la construction d'une telle mesure est dépendante de son usage, il apparaît que c'est au crible de la prise de décision qu'il nous faut spécifier les contraintes de la construction de la mesure de la précision. Pour comprendre le propos, prenons deux exemples.

Le premier, choisi dans le cadre de l'étude et repris par la suite concerne la mesure de précision associée à un attribut de forme, tel qu'un moment de Zernike. Par le choix de l'ordre des moments, l'expert en traitement d'images peut définir le niveau de détail choisi dans la caractérisation de la forme de l'objet. De façon évidente, la mesure de précision sera liée à ce choix, mais pas uniquement. Selon le type d'attribut choisi, pour un même ordre de moment, sa capacité discriminatoire sera plus ou moins forte selon que l'objectif considéré sera construit à partir de parties sphériques ou géométriques. Or, ce que nous cherchons à prendre en compte exactement, c'est cette capacité à discriminer l'objet en fonction de l'attribut, capacité dépendante de l'attribut d'une part et de l'objet d'autre part. Cet aspect de la mesure est lui purement dynamique puisqu'évalué pour chaque objet.

Pour le second exemple, hors du cadre de validation de notre propos, nous pouvons étendre l'étude sur les attributs de texture. Quelque soit l'attribut considéré, l'information texture est estimée sur un ensemble de pixels connexes, mais sans considération sur la taille de cet ensemble ou même sur la forme (sauf pour les approches séquentielles). Dans un contexte d'accroissement régulier de la taille des images avec son corollaire sur la résolution des images acquises, comment exploiter cette mesure de texture ? De façon évidente, cette information n'a pas le même sens pour une région représentant moins d'un pour cent de l'image que pour une région représentant 20% de l'image.

Le sujet de la mesure de précision associée à chaque attribut, ou extraction

d'information, paraît ainsi générique et extensible à tout opérateur, et s'intègre naturellement dans la chaîne de traitements comme l'illustre la figure 3.2. Bien évidement, l'adaptation du propos peut être plus ou moins délicate selon le formalisme mathématique utilisé pour spécifier l'attribut ou selon que la mesure embarque des notions psychovisuelles. Néanmoins, nous pouvons poursuivre ce chapitre en considérant cette notion, pour au final étudier comment la prendre en compte.

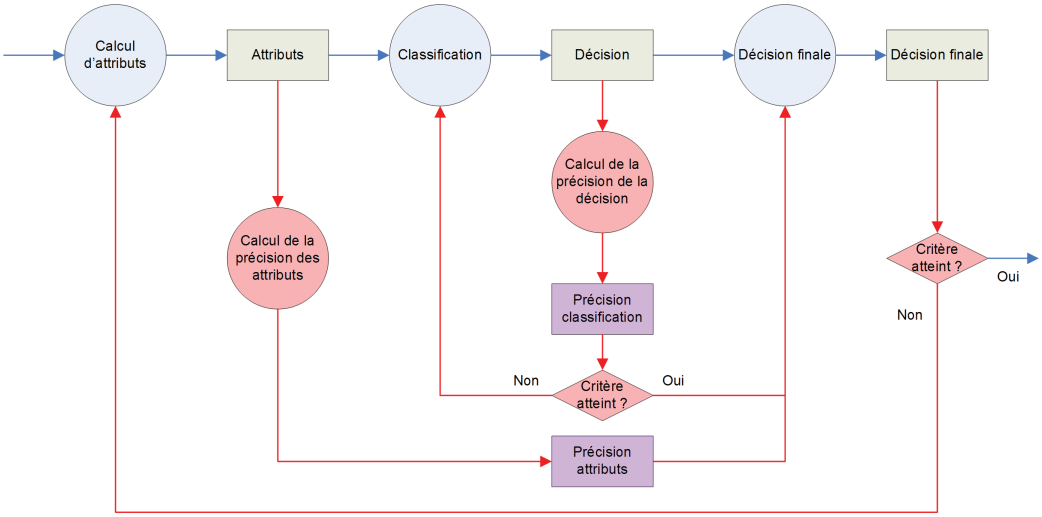

FIGURE 3.2 – Chaîne de traitements intégrant la précision des attributs et de la classification

Cependant, maintenant que nous avons intégré la notion de précision associée à la prise de mesure pour un attribut, il serait ridicule de ne pas considérer que l'outil devant prendre la décision pourrait affecter ladite décision sans quantifier son degré de certitude. A l'image de ce que nous avons appris en métrologie[19], les mesures de tolérance/précision se combinent dans la chaîne de traitements pour aboutir à une mesure finale dotée de sa propre tolérance/précision[82]. Donc, au cours du chapitre qui commence ici, la question posée est double. Premièrement, comment prendre en compte le couple (*mesure*, *précision*) dans la prise de décision ? Et deuxièmement, comment établir la mesure de précision associée à la

décision ?

Cette seconde partie de la décision est terriblement nécessaire car elle est au cœur de notre vision d'un système bouclé de traitement d'images. Cette mesure doit servir de critère de bouclage ou d'arrêt du traitement, ou bien comme mesure dans un couple (décision, précision) pour un étage supplémentaire de combinaison de décisions.

Au sein de ce chapitre, nous nous intéressons aux méthodes de base de la prise de décision, en commençant par les approches probabilistes pour aboutir aux approches liées à la théorie de l'évidence, proposant directement au sein du formalisme une possibilité d'intégration de la capacité discriminatoire de l'attribut. Nous proposerons ensuite une modification du formalisme pour prendre en compte la mesure de précision associée à la mesure proprement dite, issue du calcul de l'attribut, et enfin, nous analyserons quelques résultats.

Dans ce chapitre, nous considérerons comme acquise la mesure de précision, étant entendu qu'elle sera précisée plus tard dans le chapitre 5. L'objectif est ici de spécifier l'intérêt de cette mesure avant de la décrire, mais surtout d'en faire l'objectif principal plutôt qu'une application.

3.2 Approche probabiliste

Puisque nous cherchons à proposer une nouvelle extension aux outils de prise de décision, il nous apparaissait important de rester dans le cadre des outils à fond mathématique, par opposition aux outils construits sur des bases de règles. Comme les probabilités sont au cœur de ces approches, nous avons choisi de démarrer ce chapitre à partir de ce formalisme.

Classiquement, nous appelons probabilité, l'évaluation du caractère probable d'un évènement. Elle est représentée par un nombre réel compris entre 0 et 1, indiquant le degré de risque (ou de chance) que l'évènement se produise. Dans l'exemple typique du lancer de pièce de monnaie, l'évènement "coté pile" présente une probabilité de 1/2, ce qui signifie qu'après avoir lancé la pièce un très grand nombre de fois, la fréquence d'apparition de "coté pile" sera de 1/2.

En probabilité, un évènement peut être à peu près tout ce que l'on imagine, pouvant se produire ou non. Cela peut être le fait qu'il fasse beau demain ou le fait d'obtenir le chiffre 6 en lançant un dé. La seule contrainte est de pouvoir vérifier l'évènement ; il est tout à fait possible de vérifier s'il fera beau demain (évènement que l'on mesure de façon personnelle, le lendemain, en fonction de ses goûts en matière d'ensoleillement, de température etc.), ou si on arrive à obtenir un chiffre 6 avec un dé.

Nous dirons qu'un évènement ayant une probabilité de 0 est impossible, tandis qu'un évènement ayant une probabilité de 1 est certain.

Notons que la probabilité est un outil de prédiction d'évènements du monde réel, et non un outil d'explication de celui-ci. Il s'agit d'étudier des ensembles en les mesurant.

3.2.1 Définitions

Une expérience est dite aléatoire quand son résultat n'est pas prévisible. Définissons alors $\mathcal{E}$, une **expérience aléatoire** ayant pour résultat ω un élément de Ω, l'ensemble de tous les résultats possibles, appelé **univers des possibles** ou **référentiel**. Soit $\mathcal{P}(\Omega)$ l'**ensemble des parties** de Ω. Un événement est alors une proposition logique liée à une expérience, relative au résultat de celle-ci. Notons

$\mathcal{A}$ l'ensemble des événements et sous-ensemble de $\mathcal{P}$. Pour tout $A \in \mathcal{A}$ et $B \in \mathcal{A}$:

- $A \cup B$ désigne la réalisation de A **ou** B
- $A \cap B$ désigne la réalisation de A **et** B
- $\bar{A} = \Omega \backslash A$ désigne le contraire de A
- Ω est l'événement certain
- $\emptyset$ est l'événement impossible

Quand le référentiel Ω est fini, $\mathcal{A}$ désigne toutes les parties de Ω, noté habituellement 2^{Ω}. Quand le référentiel est • (ou un intervalle de •) la notion de **tribu** permet de définir $\mathcal{A}$:

- $\mathcal{A} \subseteq \mathcal{P}(\Omega)$
- $\emptyset \in \mathcal{A}$
- $\forall \{A_1, \ldots, A_n\} \in \mathcal{A}$, $\bigcup_i A_i \in \mathcal{A}$ et $\bigcap_i A_i \in \mathcal{A}$
- $\bar{A} \in \mathcal{A}$ $\forall A \in \mathcal{A}$

Une **mesure de probabilité** sur un espace mesurable $(\Omega, \mathcal{A})$ est une fonction de $\mathcal{A}$ dans $[0,1]$ telle que :

- $P(\Omega) = 1$
- Pour tous les éléments A et B **incompatibles** (i.e. $A \cap B = \emptyset$), $P(A \cup B) = P(A) + P(B)$

Le nombre $P(A)$ quantifie dans quelle mesure l'événement $A \subseteq \Omega$ est probable.

Dans le cas du lancer de dé, nous pouvons affirmer que $\Omega = \{1,2,3,4,5,6\}$ représente l'ensemble des 6 faces du dé. Ainsi, la probabilité d'obtenir un chiffre compris entre 1 et 6 est certaine, donc égale à 1. Nous écrirons donc $P(1 \cup 2 \cup 3 \cup 4 \cup 5 \cup 6) = P(\Omega) = 1$.
De la même façon, la probabilité d'obtenir 1 (ou 2, ou 3, ou 4 etc.) est de 1/6. Donc, $P(1) = P(2) = ... = P(6) = \frac{1}{6}$.

En lien avec P, dans le cas où Ω est fini, une **distribution de probabilité** est définie comme une fonction p de Ω dans $[0,1]$, telle que :

$$P(A) = \sum_{\omega \in A} p(\omega) \quad \forall A \subseteq \Omega \tag{3.1}$$

avec la condition de normalisation :

$$\sum_{\omega \in \Omega} p(\omega) = 1 \tag{3.2}$$

Notons que $P(A) + P(\bar{A}) = 1$

3.2.2 Règle de Bayes

L'introduction d'un modèle statistique conduit à une théorie de la décision directement applicable, en principe, au problème de la classification. Cette application repose essentiellement sur la règle de Bayes qui permet d'évaluer des probabilités *a posteriori* (c'est à dire après l'observation effective de certaines grandeurs) connaissant les distributions de probabilité conditionnelles *a priori* (c'est-à-dire indépendantes de toute contrainte sur les variables observées). Les probabilités nécessaires à la prise de décision selon la règle de Bayes sont apportées par un (ou plusieurs) classifieur statistique qui indique la probabilité d'appartenance d'un individu donné à chaque classe du problème.

En théorie des probabilités, le théorème de Bayes énonce des probabilités conditionnelles : étant donné deux évènements A et B, le théorème de Bayes permet de déterminer la probabilité de A sachant B, si l'on connaît les probabilités : de A, de B et de B sachant A. Pour aboutir au théorème de Bayes, on part d'une des définitions de la probabilité conditionnelle : $P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$, en notant $P(A \cap B)$ la probabilité que A et B aient lieu tous les deux. En divisant de par et d'autre par $P(B)$, on obtient $P(A|B) = P(B|A)P(A)/P(B)$ soit le théorème de Bayes. Chaque terme du théorème de Bayes a une dénomination usuelle. Le terme $P(A)$ est la **probabilité *a priori*** de A. Elle est « antérieure » au

sens qu'elle précède toute information sur B. $P(A)$ est aussi appelée la **probabilité marginale** de A. Le terme $P(A|B)$ est appelée la **probabilité a posteriori** de A sachant B (ou encore de A sous condition B) . Elle est "postérieure", au sens qu'elle dépend directement de B. Le terme $P(B|A)$, pour un B connu, est appelé la **fonction de vraisemblance** de A. De même, le terme $P(B)$ est appelé la probabilité marginale ou a priori de B.

Mais, comme énoncé précédemment, les valeurs des probabilités utilisées par la règle de Bayes sont apportées par des classifieurs statistiques, voire tout simplement par estimation des lois de probabilités de chacune des classes.

3.2.3 Prise de décision statistique

Depuis l'hypothèse sur les lois de probabilités des classes du problème (et ses tests de validation) jusqu'à la mesure de la distance au centre de chaque classe en passant par différentes formes d'estimation de modèle, il existe dans le domaine des statistiques un très grand nombre de méthodes et notre propos n'est pas d'en faire une étude exhaustive mais de comparer les théories (statistiques, théorie de la décision...) sur le principe.

Sachons seulement qu'il existe notamment deux types d'approches statistiques : les classifieurs linéaires et les classifieurs quadratiques. Les premiers sont appelés ainsi car leur approche consiste à partitionner l'espace des attributs de façon simple, à partir de droites, chaque droite symbolisant la frontière entre deux classes, comme l'illustre la frontière en vert sur la figure 3.3. La seconde approche est plus complexe car elle tente d'estimer la forme des frontières entre les régions, c'est ainsi que l'on obtient la séparation circulaire entre les classes visible sur la figure 3.3.

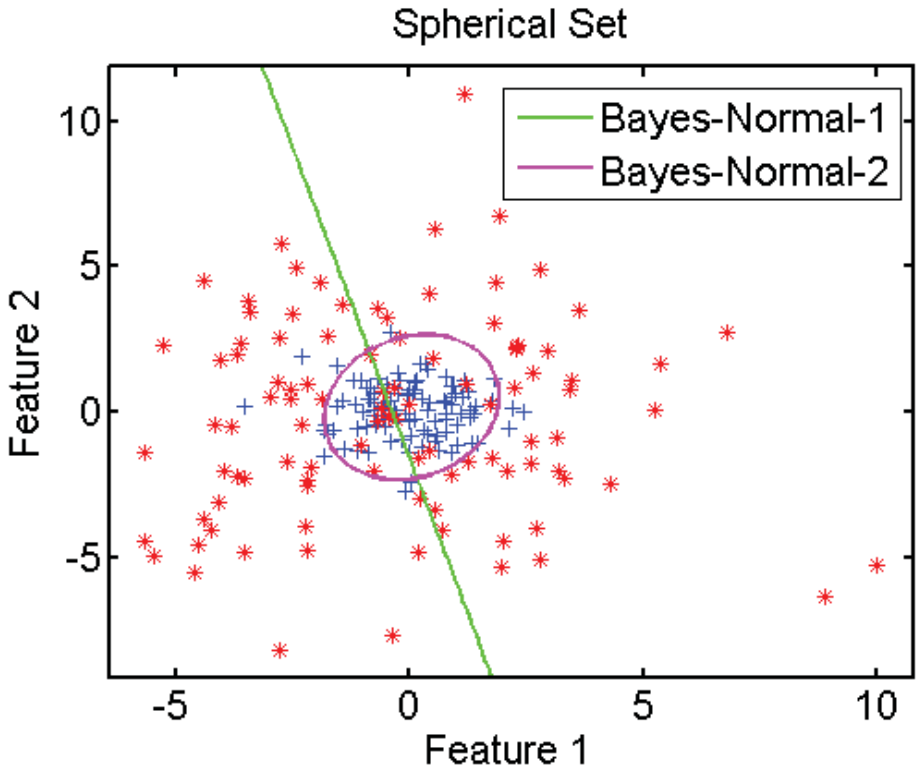

FIGURE 3.3 – Frontières selon un classifieur linéaire (Bayes-Normal-1) et un classifieur quadratique (Bayes-Normal-2)

Afin d'illustrer notre propos sur les approches statistiques, nous étudierons des résultats obtenus sur nos données par un classifieur linéaire et un classifieur quadratique.

Classifieur Bayesien[93][9][63]

La théorie de la décision de Bayes est une approche statistique du problème. Elle est basée sur les hypothèses suivantes :

- Le problème de la décision est exprimé en terme de probabilités
- La valeur des probabilités *a priori* du tirage d'une classe est connue

Soit ω, une variable aléatoire représentant l'état de l'individu (ω_j si l'individu appartient à la classe j).
Soit $\Omega = \{\omega_1, \ldots, \omega_S\}$ ensemble des S états possibles ou classes d'un problème.
L'ensemble des probabilités ***a priori*** $p(\omega_1), p(\omega_2), \ldots$ est calculé à partir des d'observations réalisées sur le système étudié.
En l'absence d'autres informations, il ne reste d'autre choix que d'écrire $p(\text{individu} = \omega_i) = p(\omega_i)$.
Mais s'il est possible d'obtenir d'autres informations (mesures), alors $p(x|\omega_j)$, densité de probabilité de x **sachant** que l'individu est ω_j, ou densité de probabilité conditionnelle à son appartenance à ω_j, est calculable.
$p(\omega_j)$ est donc la probabilité d'apparition *a priori* de la classe ω_j et $p(\omega_j|x)$ est la probabilité que la forme caractérisée par le vecteur x appartienne à ω_j.
Dans ce cas, nous disposons de $x = \{x_1 \ldots x_d\}$, le vecteur aléatoire (mesures ou situation) sur $\mathcal{R}^d$.
Le problème devient alors : *soit un individu et sa mesure x. Comment décider à quelle catégorie (ω_j) affecter cet individu ?*.
La règle de Bayes donne la réponse :

$$p(\omega_j|x) = \frac{p(x|\omega_j)p(\omega_j)}{p(x)} \ , \ j \in \{1, S\} \tag{3.3}$$

avec

$$p(x) = \sum_{j=1}^{S} p(x|\omega_j)p(\omega_j) \tag{3.4}$$

où S est le nombre de classes différentes et $p(x)$ la distribution de probabilité pour l'ensemble des individus (densité de probabilité du vecteur x), dans laquelle les échantillons de chaque classe ω_j apparaissent *proportionnellement* aux probabilités ***a priori*** $p(\omega_j)$; c'est le facteur de normalisation.
La règle de décision est alors la suivante :

$$\text{décider } \omega_j \text{ si } p(\omega_j|x) = max_{i=1,n}(p(\omega_i|x))$$

Ainsi, il devient possible d'énumérer $A = \{\alpha_1, \ldots, \alpha_k\}$ l'ensemble des a décisions possibles.

Il reste à présent à exprimer les densités de probabilité conditionnelle. Pour cela, rappelons l'*inégalité de Bienaymé-Tchebicheff* :
Soient $X_1, X_2, ..., X_n$ des variables aléatoires indépendantes, toutes d'espérance μ et de variance σ^2. Soit $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ alors pour tout $d > 0$:

$$P(|\bar{X} - \mu| \geq d) \leq \frac{\sigma^2}{nd^2} \tag{3.5}$$

La probabilité qu'une variable aléatoire s'écarte de plus de d *de sa valeur moyenne est d'autant plus faible que sa variance est petite et que d est grand.*
En particulier (**Loi des grands nombres**)[7] :

$$P(|\bar{X} - \mu| \geq d) \longrightarrow 0 \ quand \ n \to +\infty \tag{3.6}$$

Si $n \geq 30$, il est admis, en appliquant le théorème central-limite (qui affirme intuitivement que toute somme de variables aléatoires indépendantes et identiquement distribuées tend vers une variable aléatoire gaussienne), que $X_1 + X_2 + ... + X_n$ suit pratiquement une loi Normale. Rappelons que la densité d'une loi Normale est de la forme (sur une dimension) :

$$x \sim N(\mu, \sigma^2), \ f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{3.7}$$

Enfin, nous savons que l'espérance mathématique d'une variable aléatoire peut être estimée ponctuellement par :

$$\mu \simeq \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.8}$$

Donc, pour notre mesure x, les échantillons sont supposés suivre une loi Normale $N(\bar{x}, \sigma^2)$ dont la densité de probabilité sera exprimée sous forme matricielle :

$$P(x) = \frac{1}{2\pi^{\frac{n}{2}}|\Phi|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\bar{x})^t[\Phi]^{-1}(x-\bar{x})\} \tag{3.9}$$

avec $\bar{x}$ moyenne d'une classe, $[\Phi_k]$ matrice de covariance de la classe k et $|\Phi_k|$ déterminant de la matrice de covariance de la classe k. La probabilité conditionnelle est donc exprimée par l'équation suivante :

$$P(x|\omega_k) = \frac{1}{2\pi^{\frac{n}{2}}|\Phi_k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\bar{x}_k)^t[\Phi_k]^{-1}(x-\bar{x}_k)\} \tag{3.10}$$

L'opération de classement consistera à calculer pour chaque mesure x sa probabilité d'appartenance à toutes les classes, et de l'affecter à la classe qui vérifie la probabilité a posteriori maximum.

Les k-plus proches voisins (KPPV)

Cette approche consiste à prendre dans l'espace des attributs un nombre donné (k) d'individus connus, individus choisis dans le voisinage de l'individu dont on cherche à estimer la classe d'appartenance. C'est à la connaissance des classes représentées dans ce voisinage que se fait l'étiquetage de l'individu inconnu.

La règle des k-plus proches voisins a été formalisée par Fix et Hodges[14] dans les années 50. Cette règle consiste à baser le choix du voisinage $\mathcal{R}$ sur le nombre d'observations qu'il contient fixé à k parmi le nombre total d'observations m. Autrement dit, il s'agit de sélectionner dans l'ensemble des observations,

un sous-ensemble de k individus. C'est le volume V_k de $\mathcal{R}$ qui est variable. L'estimateur de la densité de probabilité s'exprime alors de la façon suivante :

$$\hat{P}(x) = \frac{k}{m \times V_k} \tag{3.11}$$

Cette estimation n'est valable que dans les conditions suivantes :

$$\begin{cases} \lim_{m\to+\infty} V_k &= 0 \\ \lim_{m\to+\infty} \frac{k}{V_k} &= 0 \end{cases}$$

La première hypothèse correspond à un espace complètement occupé par les observations : quand le nombre d'observations est infini, ces observations occupent l'espace de façon uniforme. La seconde hypothèse considère que, pour un nombre infini d'observations, tout sous-ensemble de cardinal k est occupé par un nombre infini d'éléments répartis uniformément.

Dans le cadre de la classification supervisée, la loi de probabilité conditionnelle peut être approximée par :

$$\hat{P}(x|\omega_q) = \frac{k_q}{m_q \times A} \tag{3.12}$$

où k_q est le nombre d'observations contenues dans le volume V_k appartenant à la classe d'étiquette ω_q et m_q le nombre total d'observations appartenant à la classe ω_q. La loi *a posteriori* d'observation d'une étiquette conditionnellement à une observation s'obtient avec la règle de Bayes :

$$\hat{P}(\omega_q|y) = \frac{P(\omega_q)\hat{P}(y|\omega_q)}{\sum_k P(\omega_k)\hat{P}(y|\omega_k)} = \frac{P(\omega_q)\frac{k_q}{m_q \times V_k}}{\hat{P}(y)} = \frac{P(\omega_q) \times m}{m_q}\frac{k_q}{k} \tag{3.13}$$

Le choix de k est lié à la base d'apprentissage. Prendre k élevé permet d'exploiter au mieux les propriétés locales mais nécessite un grand nombre d'échantillons pour contraindre le volume du voisinage à rester petit. Bien souvent, k est choisi comme la racine carrée du nombre moyen d'élément par classe soit $\sqrt{\frac{m}{c}}$ [2].

3.2.4 Synthèse

Nous avons implémenté les méthodes de classification décrite ci-dessus et les avons testées avec les données présentées au chapitre 2 afin d'illustrer notre propos sur les approches statistiques de la prise de décision. Les résultats obtenus par le classifieur Bayesien sont présentées en figure 3.4, et ceux obtenus avec la méthodes des kppv en figure 3.5.

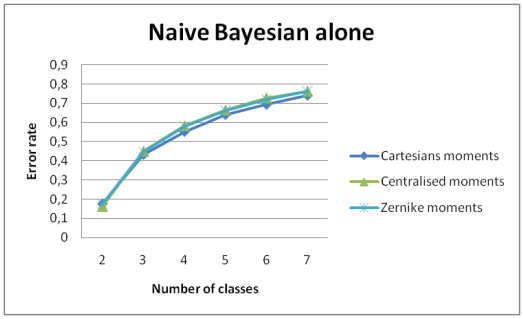

FIGURE 3.4 – Taux d'erreur pour le Bayesien seul

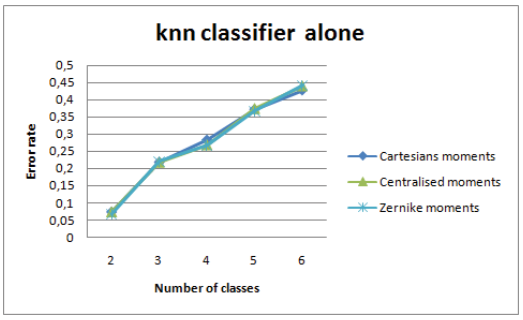

FIGURE 3.5 – Taux d'erreur pour le KPPV seul

Le principal constat que nous pouvons faire avec les résultats obtenus par le classifieur Bayesien est que le taux d'erreur est proche de 20% pour un simple problème à deux classes, et dépasse les 50% au delà de quatre classes, ce qui n'est absolument pas une bonne performance, mais qui s'explique très bien par le principe même de fonctionnement d'un classifieur linéaire : comment séparer la répartition complexe des individus de deux classes différentes dans l'espace des attributs au moyen d'une seule droite ?

Les résultats obtenus avec la méthode des KPPV ont un taux d'erreur globalement plus faible qu'avec le Bayesien, et ces résultats deviennent même très acceptables pour le problème à deux classes. Nous retrouvons, de façon logique, l'intérêt d'une séparation complexe des classes dans l'espace des attributs.

En présence de choix, la théorie des probabilités propose de calculer les espérances mathématiques de gain et d'opter pour le choix qui maximise cette espérance de gain. Cependant, par rapport à notre propos, ce procédé présente plusieurs limites. En effet, le formalisme figé des probabilités ne permet pas d'intégrer les notions que nous étudions, à savoir la précision ou la pertinence de l'information.

D'autre part, un apprentissage s'avère nécessaire pour intégrer les statistiques (il s'agit pour nous des probabilités des enchaînements entre les opérateurs). Se pose dans ce cas le problème de la construction dynamique de cet apprentissage.

Supposons à présent que nous décidions de construire une chaîne d'opérateurs non linéaire, avec bouclages. Le système devra décider à un moment donné que les enchaînements passés n'ont pas donné de résultat satisfaisant, et qu'il s'avère nécessaire de tenter d'autres combinaisons d'opérateurs. Mais comment alors choisir autre chose que ce qui est statistiquement le plus probable ?

Enfin, si nous voulons construire un système contraint par l'utilisateur (temps, précision), comment intégrer la gestion de tels critères lors d'une prise de décision statistique ?

Se pose alors le problème de la mesure d'une précision d'une décision accompagnée de la production d'une mesure de la précision, même si le problème peut être cette mesure.

3.3 Théorie de l'évidence

Le modèle des croyances transférables MCT (TBM : transferable belief model) est un cadre formel générique développé par Ph. Smets[77] pour la représen-

tation et la combinaison des connaissances. Le TBM est basé sur la définition de fonctions de croyance fournies par des sources d'information pouvant être complémentaires, redondantes et éventuellement non-indépendantes. Il propose un ensemble d'opérateurs permettant de combiner ces fonctions. Il est donc naturellement employé dans le cadre de la fusion d'informations pour améliorer l'analyse et l'interprétation de données issues de sources d'informations multiples. Ce cadre correspond naturellement à celui qui nous préoccupe par son aspect de prise de décision grâce à différentes informations issues de différents attributs.

L'un des points fondamentaux qui caractérisent le TBM est la différenciation des niveaux de représentation des connaissances et de décision. Cette différenciation est beaucoup moins prépondérante pour d'autres approches, particulièrement pour le modèle probabiliste pour lequel la décision est souvent le seul objectif visé. Les mécanismes de raisonnement du TBM sont donc regroupés en deux niveaux comme illustré sur la figure 3.6 :

- Le niveau **crédal** : siège de la représentation des connaissances (partie statique), des combinaisons et du raisonnement sur ces connaissances (partie dynamique)
- Le niveau **pignistique** indiquant une prise de décision en prenant éventuellement en compte le risque et/ou le gain associés à cette décision.

Dans notre chaîne de traitements, nous nous retrouvons confrontés au problème de l'observation du même événement par plusieurs sources, et de la combinaison de l'information apportée par ces sources (voir figure 3.7), imposant les questions suivantes :

- Comment représenter la connaissance d'une source d'information sous forme de fonctions de croyance et quels avantages peut-on tirer d'une telle représentation ?
- Comment combiner plusieurs sources de fonctions de croyance afin de résumer l'information et d'améliorer la prise de décision ?
- Comment prendre une décision à partir des fonctions de croyance ?

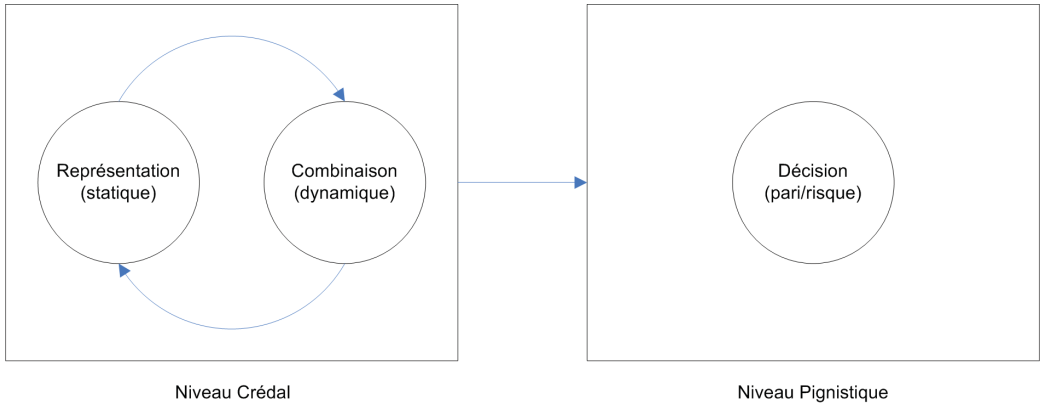

FIGURE 3.6 – Mécanismes du modèle des croyances transférables

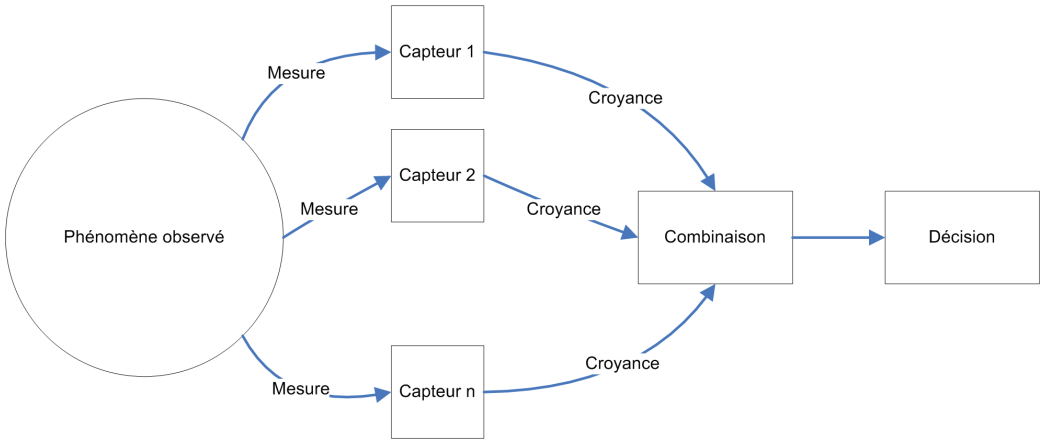

FIGURE 3.7 – Chaîne de traitements

3.3.1 Présentation

Considérons un exemple proche du notre, comme une reconnaissance de chiffres manuscrits. Les chiffres possibles pouvant être écrits sont appelés **hypothèses** et l'ensemble des hypothèses forment le **cadre de discernement** généralement noté Ω. Pour notre exemple, $\Omega = \{\omega_0, \omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8, \omega_9\}$ où ω_i est le symbole représentant le chiffre i. Deux hypothèses ne pouvant être vraies simul-

tanément (on ne peut écrire un symbole représentant à la fois plusieurs chiffres), elles sont supposées être **exclusives**, ce qui se traduit par une intersection nulle entre elles (i.e. $\{\omega_1\} \cap \{\omega_3\} = \emptyset$). Supposons maintenant que le document que nous sommes en train d'étudier soit dégradé, ou bien que le chiffre soit très mal écrit, un lecteur donnerait alors un avis pondéré sur ce chiffre, faisant de lui une **source** d'information. Avec sa connaissance disponible, le lecteur affirme qu'il s'agit soit du chiffre ω_1, soit ω_7 ou soit ω_9. La question est de savoir comment représenter cette information.

Dans le cadre de la théorie des probabilités, la réponse est donnée par le principe **d'équiprobabilité**, c'est à dire que chaque chiffre se voit attribué la même probabilité ($\frac{1}{3}$) de sorte que la somme de toutes les probabilités soit égale à 1. Ainsi, la distribution des probabilités pour ce chiffre est égale à :

$$P(\omega_1) = P(\omega_7) = P(\omega_9) = \frac{1}{3}$$

Cette probabilité n'est affectée qu'à des hypothèses prises seules (appelées également *singletons*).

En théorie de l'évidence, la notion principale est celle de **fonctions de croyance** au sein de laquelle la connaissance est modélisée par une distribution de **masses de croyance**. Dans notre exemple, toute la masse de croyance (une unité entière comme avec les probabilités) est affectée à l'union de toutes les hypothèses, soit la **proposition** $\{\omega_1, \omega_7, \omega_9\}$:

$$m(\{\omega_1, \omega_7, \omega_9\}) = 1$$

Comme nous pouvons le constater, cette masse ne nous donne aucune information concernant la croyance de chacune des hypothèses (ω_1, ω_7 et ω_9) qui la compose. Nous pouvons dire alors que la masse de croyance modélise explicitement le doute (ou l'ignorance) sur cet ensemble d'hypothèses. Pour revenir à la théorie des probabilités, la valeur de la probabilité sur une union d'hypothèses découle

implicitement de la probabilité sur les singletons, probabilité établie à partir de la somme de toutes les probabilités privée de la partie commune, c'est à dire la probabilité de l'intersection. Soit par exemple :

$$P(\{\omega_1, \omega_7\}) = P(\omega_1) + P(\omega_7) - P(\{(\omega_1, \omega_7)\})$$

Ici, la probabilité sur $P(\{(\omega_1, \omega_7)\})$ représente la partie commune des évènements "*Le chiffre est 1*" et "*Le chiffre est 7*", que nous devons retirer afin de ne pas compter les évènements deux fois. Dans le cas où les hypothèses sont exclusives (cas que nous retrouverons tout au long du manuscrit), on obtient :

$$\{\omega_1\} \cap \{\omega_7\} = \emptyset \Rightarrow P(\{(\omega_1, \omega_7)\}) = 0 \Rightarrow P(\{\omega_1, \omega_7\}) = P(\omega_1) + P(\omega_7)$$

en prenant en compte le fait que $P(\emptyset) = 0$ par définition.

Contrairement aux probabilités, une fonction de masse sur une union d'hypothèses n'est pas égale à la somme des masses des hypothèses composant l'union[44], et c'est justement grâce à cette propriété que nous représentons le doute entre des hypothèses. De plus, les fonctions de croyance ont pour avantage d'être génériques, car elles sont la représentation mathématique d'un sur-ensemble de fonctions de probabilités. En effet, une distribution de masses, où seules les hypothèses singletons ont une masse non nulle, est interprétable comme une distribution de probabilités. Klir et Wierman expliquent dans [44] que les fonctions de croyance généralisent mathématiquement les fonctions de possibilités et que chacune de ces fonctions (probabilités, possibilités et croyances) possède des caractéristiques particulières et permettent de modéliser la connaissance différemment.

Voyons à présent le formalisme mathématique des fonctions de croyance, basé sur les modèles de Dempster[15][16], Shafer[67], des croyances transférables[77] et sur le modèle des Hints[41]. Pour ce travail, nous nous sommes restreints au cadre axiomatisé et formalisé du modèle des croyances transférables développé par P. Smets.

3.3.2 Formalisme

Introduction

Si nous revenons au problème de la gestion de sources multiples d'informations (comme illustré par la figure 3.7), ce que nous cherchons à faire est de déterminer l'état réel ω_0 du système étudié en utilisant plusieurs observations issues de plusieurs observateurs. Dans notre cas, ω_0 sera la décision à produire et les différents attributs fournissent les observations sur le système.

Nous supposons que cet état ω_0 prend des valeurs discrètes ω_k (les **hypothèses**) et que l'ensemble des N hypothèses possibles est appelé **cadre de discernement** (ou **univers de discours**), noté généralement :

$$\Omega = \{\omega_1, ..., \omega_k, ..., \omega_N\} = \bigcup_{k=1}^{N} \{\omega_k\} \tag{3.14}$$

Le nombre total d'hypothèses composant Ω est appelé cardinal et s'écrit comme $|\Omega| = N$. Comme nous l'avons vu précédemment, toutes les hypothèses sont considérées exclusives.

L'ensemble des observateurs du système (les capteurs de la figure 3.7) fournit un avis pondéré, représenté grâce à une fonction de croyance, à propos de son état réel $\omega_0 \subseteq \Omega$. Nous appelons alors ces capteurs, des **sources** de croyance.

Fonctions de masses

Notons 2^Ω l'espace formé de toutes les parties de Ω, c'est à dire l'espace rassemblant tous les sous-ensembles possibles formés des hypothèses et unions d'hypothèses, soit :

$$2^\Omega = \{\emptyset, \{\omega_1\}, \{\omega 2\}, \{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_1, \omega_3\}, \{\omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_3\}, ..., \{\omega_1, ..., \omega_N\}\}$$

Soit A une **proposition** telle que $A \subseteq \Omega$, par exemple $A = \{\omega_1, \omega_2\}$. A représente explicitement le doute entre les hypothèses qui la compose (ω_1 et ω_2).

Comme nous l'avons vu plus haut, les masses de croyances sont non-additives, ce qui est une différence fondamentale avec la théorie des probabilités. Ainsi, la masse de croyance $m(A)$ allouée à A ne donne aucune information à propos des hypothèses et sous ensembles composant A[44].

Nous pouvons définir une **distribution de masses de croyance** (BBA : basic belief assignment) comme un ensemble de masses de croyance concernant des propositions quelconques $A \subseteq \Omega$ vérifiant :

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{3.15}$$

Chacun des capteurs observant le système apportera sa propre distribution de masse.

La masse $m(A)$ fournie par un capteur est la part de croyance de ce capteur en la proposition "$\omega_0 \in B$" où ω_0 est l'état réel du système observé. Tout élément B de masse non nulle (soit $m(B) > 0$) est appelé **élément focal** de la distribution de masses de croyance. Ramasso[60] présente dans son mémoire de Thèse un ensemble notable de distributions de masses de croyance.

Quand l'univers du discours contient la totalité des hypothèses possibles (c'est à dire lorsqu'il est **exhaustif**), il contient forcément l'hypothèse permettant de décrire le système observé. Dans le cas contraire, le cadre de discernement n'est pas adapté au problème traité.

A partir du moment où Ω est exhaustif, la masse de l'ensemble vide est nulle ($m(\emptyset) = 0$), la masse est dite **normale**, et la problématique entre alors dans le cadre d'un **monde fermé** comme dans le cas du modèle de Shafer[67].

Dans le cas du modèle des croyances transférables, l'univers du discours peut être non exhaustif ; la masse de l'ensemble est non nulle ($m(\emptyset) \neq 0$) et elle est dite

sous normale et dans ce cas, la problématique s'inscrit dans le cadre d'un **monde ouvert**. Ce type de construction permet de redéfinir, par exemple, le cadre de discernement en ajoutant une hypothèse en fonction de la masse de l'ensemble vide, comme une **hypothèse de rejet**. Les hypothèses de rejet correspondent à des états du genre "*je ne sais pas*", permettant en fin de chaîne de rejeter un individu pour lequel le doute est trop grand au lieu de prendre le risque de le classer malgré tout. Pour passer d'un monde ouvert à un monde fermé, il faut redistribuer la masse de l'ensemble vide sur les autres sous-ensembles de l'espace 2^{Ω}[52].

Pour résumer ce que nous avons vu jusqu'à présent, la fonction de masse est déterminée *a priori*, sans nécessiter de considérations probabilistes. Elle peut correspondre à l'attribution, de façon subjective, de degrés de crédibilité en la réalisation des différents évènements envisageables sur Ω, par un observateur.

L'attribution des valeurs de m sert à exprimer le degré avec lequel l'expert ou l'observateur juge que chaque hypothèse est susceptible de se réaliser. Ainsi, $m(\Omega)$ décrit son imprécision relativement à ces degrés.

A partir de la fonction de masse m, nous pouvons écrire d'autres fonctions représentant la même information mais sous une forme différente. Ce sont des outils mathématiques portant des interprétations différentes, mais elles sont également utiles pour simplifier les calculs de combinaisons. Nous retrouverons plus bas les fonctions suivantes :

1. **Crédibilité** $bel(A)$. La fonction de crédibilité (ou fonction de **croyance**) résume dans quelle mesure la réalisation de chaque événement est crédible, étant donnée l'incertitude exprimée. Pour un élément A, c'est la part de croyance spécifiquement allouée à A, soit :

$$\begin{aligned} bel(A) = &\quad \sum_{B\subseteq A, B\neq\emptyset} m(B), \quad \forall A \subseteq \Omega \\ m(A) = &\quad \sum_{B\subseteq A} (-1)^{|A|-|B|} bel(B), \quad \forall A \subseteq \Omega \end{aligned} \tag{3.16}$$

2. **Plausibilité** $pl(A)$. Du fait de l'imprécision de l'expert, ni la fonction de masse, ni la fonction de croyance ne suffisent à indiquer précisément dans quelle mesure un événement est susceptible de se réaliser, comme ce serait le cas en probabilités. Si la fonction de croyance indique la borne inférieure de la probabilité qu'il est possible d'attribuer à un événement, il est intéressant de renforcer notre connaissance sur celle-ci par l'intermédiaire de sa borne supérieure. Pour cela, la plausibilité d'un élément A est définie comme la part maximale de croyance qui pourrait soutenir A :

$$\begin{aligned} pl(A) = & \quad \sum_{B\cap A\neq\emptyset} m(B), & \forall A \subseteq \Omega \\ m(A) = & \quad \sum_{B\subseteq A} (-1)^{|A|-|B|+1} pl(\overline{B}), & \forall A \subseteq \Omega \end{aligned} \tag{3.17}$$

3. **Communalité** $q(A)$. Un autre coefficient permet de synthétiser les informations fournies par un corps d'évidence en jouant le rôle inverse de la fonction de masse par rapport à la fonction de croyance. C'est la somme des masses allouées aux sous ensembles de A et donc qui ont A en commun :

$$\begin{aligned} q(A) = & \quad \sum_{B\supseteq A} m(B), & \forall A \subseteq \Omega \\ m(A) = & \quad \sum_{A\subseteq B} (-1)^{|B|-|A|} q(\overline{B}), & \forall A \subseteq \Omega \end{aligned} \tag{3.18}$$

4. **Implicabilité** $b(A)$. C'est la somme des masses allouées aux sous ensembles de A tels que leur véracité implique la véracité de A :

$$\begin{aligned} b(A) = & \quad \sum_{B\subseteq A} m(B) = bel(A) + m(\emptyset), & \forall A \subseteq \Omega \\ m(A) = & \quad \sum_{B\subseteq A} (-1)^{|A|-|B|} b(\overline{B}), & \forall A \subseteq \Omega \end{aligned} \tag{3.19}$$

Prenons l'exemple d'un problème quelconque à 3 évènements. Les observations sur le système permettent de construire les masses de croyances m pour la réalisation des évènements seuls, en couple ou même la réalisation des 3. A partir des valeurs de m, en fonction des hypothèses, les résultats obtenus pour chacune des fonctions détaillées précédemment sont présentés en table 3.1.

	m	pl	bel	b	q
$\emptyset$	0.1	0	0	0.1	1
$\{\omega_1\}$	0.07	0.46	0.08	0.17	0.46
$\{\omega_2\}$	0.12	0.44	0.13	0.22	0.44
$\{\omega_1, \omega_2\}$	0.22	0.59	0.45	0.51	0.31
$\{\omega_3\}$	0.31	0.49	0.34	0.41	0.49
$\{\omega_1, \omega_3\}$	0.08	0.78	0.51	0.56	0.17
$\{\omega_2, \omega_3\}$	0.01	0.83	0.48	0.54	0.1
$\{\omega_1, \omega_2, \omega_3\}$	0.09	0.9	1	1	0.09

TABLE 3.1 – Exemple d'informations disponibles sur un système à 3 évènements

Le principe de minimum d'information

Le formalisme de la théorie de l'évidence introduit la notion de fonction de masse de croyance (m). Or, il se peut que cette information soit donnée par plusieurs sources (comme par exemple un four qui mesure sa température grace à plusieurs capteurs). Afin de gérer correctement ces différentes sources d'information, il peut être nécessaire de choisir une de ces sources, en fonction de la quantité ou de la qualité de l'information qu'elle apporte. En statistique, la quantité d'informations portée par un système peut-être associée au principe physique de l'**entropie**.

En thermodynamique, l'entropie caractérise le degré d'organisation ou de désordre d'un système. Par exemple, une tasse intacte sur une table est en état élevé d'ordre (entropie faible), alors qu'une tasse brisée sur le plancher est en désordre (entropie forte). Par analogie, nous pouvons associer la notion d'entropie à la quantité d'information contenue dans notre système. Si celui-ci est fortement informatif, son entropie sera très faible (forte cohérence de l'information) et inversement. Dans le domaine du traitement du signal, l'entropie de Shannon caractérise par exemple la distribution d'un ensemble de probabilités discrètes $P = \{p_1, ..., p_n\}$, étant donné les connaissances disponibles, à partir de l'entropie de Shannon :

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i \tag{3.20}$$

Le principe de minimum d'information (PMI) est au cœur d'un grand nombre de mécanismes développés dans le modèle des croyances transférables. L'idée est la suivante : lorsqu'une distribution de masse de croyance doit être choisie parmi un ensemble de distributions possibles (dans le cas par exemple de plusieurs sources d'information), le principe de minimum d'information impose de choisir celle qui est la moins engagée ou, autrement dit, la moins informative (donc, celle dont l'entropie sera la plus forte).

Une approche quantitative consiste à calculer le taux d'incertitude des distributions de masse de croyance et de choisir celle qui maximise ce taux d'incertitude, minimisant ainsi le parti pris. Un grand nombre de méthodes ont été proposées pour mesurer le taux d'incertitude à partir d'informations imprécises[44][34]. La plus répandue des mesures est la non-spécificité[24] basée sur la cardinalité des éléments focaux pondérée par la valeur des masses :

$$N(m) = \sum_{A \subseteq \Omega, A \neq \emptyset} m(A).log_2(|A|) \tag{3.21}$$

Cette mesure est bornée sur $[0, log_2(|A|)]$.

Plus la mesure de non-spécificité est grande, moins la distribution de masse de croyance est informative. Par exemple, la distribution de masse de croyance vide (l'élément focal est Ω i.e. $m(\Omega) = 1$) est totalement non-spécifique tandis qu'une autre distribution précise (telle qu'une distribution Bayésienne, i.e. les éléments focaux sont des singletons, donc m est une distribution de probabilités avec $m(\{\omega_q\}) > 0$, $\omega_q \in \Omega$) est totalement spécifique. Le principe de minimum d'information basé sur la non-spécificité maximise donc la valeur de ce critère jouant un rôle similaire au maximum d'entropie en théorie de l'information.

Règles de combinaison de base

Si nous revenons à la figure 3.7, nous constatons que le problème qui se pose est la fusion d'information, consistant à *combiner des informations hétérogènes issues de plusieurs sources afin d'améliorer la prise de décision*[5].

Voyons à présent plusieurs outils formels développés dans le cadre du modèle des croyances transférables permettant la fusion d'information. Pour ces méthodes, nous posons l'hypothèse que les sources sont **distinctes**[74][89][90][91] par analogie avec l'**indépendance** des variables aléatoires en statistique.

1. La **règle de combinaison conjonctive**. Soient m_1 et m_2, deux sources de masse de croyance distinctes et définies sur le même cadre de discernement Ω. Les deux sources sont supposées fiables. La règle de combinaison conjonctive (CRC : conjunctive rule of combination), notée $\textcircled{\cap}$, des fonctions de masse m_1 et m_2 est définie $\forall A,B,C \subseteq \Omega$ par les équations suivantes :

$$m_{1\textcircled{\cap}2} = (m_1 \textcircled{\cap} m_2)(A) = \sum_{B \cap C = A} m_1(B).m_2(C) \tag{3.22}$$

   L'intersection intervenant ici joue un effet de spécialisation en transférant la masse sur des sous-ensembles de cardinalité plus faible. Ainsi, lorsque l'une des fonctions de masse de croyance est Bayesienne (cas particulier où éléments focaux sont des singletons), alors le résultat est une **fonction de masse de croyance Bayesienne**.

2. **Le conditionnement**. La règle de combinaison conjonctive permet de définir l'opération de conditionnement pour combiner des informations incertaines. Conditionner une fonction de masse de croyance m par un élément $B \subseteq \Omega$ consiste à restreindre le cadre des propositions possibles 2^{Ω} à celles ayant une intersection non nulle avec B. Ainsi, lors d'un conditionnement, les masses affectées à $C \subseteq \Omega$ sont transférées sur $C \cap B$. La fonction de

masse de croyance conditionnelle qui en résulte est notée $m[B]$ (les crochets traduisent le conditionnement) et ces éléments sont donnés par la règle de conditionnement de Dempster non normalisée :

$$m[B](C) = \sum_{A \subseteq \bar{B}} m(C \cup A), \quad \forall C \subseteq B \tag{3.23}$$

L'opération de conditionnement est particulièrement adaptée dans les cas où l'on dispose d'une information certaine (mais qui peut être imprécise). Le conditionnement revient à effectuer une combinaison conjonctive avec une fonction de masse de croyance catégorique sur B (i.e. $m(B) = 1$) et donc la fonction de masse de croyance issue d'un conditionnement est souvent sous normale.

3. La **règle de combinaison disjonctive** (DRC : disjunctive rule of combination), notée $\bigcirc$[24][74], a été proposée pour combiner les fonctions de masse de croyance dont l'une au moins est fiable, sans pouvoir quantifier la fiabilité ni savoir quelle source est fiable. Elle est définie $\forall D \subseteq \Omega$ par l'équation suivante :

$$m_{1 \bigcirc 2} = (m_1 \bigcirc m_2)(A) = \sum_{B \cup C = A} m_1(B).m_2(C) \tag{3.24}$$

La règle de combinaison disjonctive est une règle de combinaison associative, commutative mais non idempotente dont l'élément neutre est la fonction de masse de croyance $m(\emptyset) = 1$ et l'élément absorbant la fonction de masse de croyance vide. Cette règle est conservatrice car elle transfère les masses sur des sur-ensembles des éléments focaux. Par conséquent, les éléments focaux résultants peuvent être de cardinal élevé. On dit que la fonction de masse de croyance résultante est moins spécialisée que celles d'origines ou encore que la DRC est un processus de généralisation. Notons que la dualité des règles $\cap$ et $\cup$ est mise en valeur dans les règles de De Morgan s'appliquant aux fonctions de croyance[24].

4. La **règle de combinaison de Dempster**.

   Il peut arriver que pour la même hypothèse, deux capteurs apportent une information complètement différente. D'un point de vue formel, lors de la combinaison de deux fonctions de masse de croyance m_1 et m_2, l'intersection entre les éléments focaux peut être vide. La fonction de masse de croyance résultat dans cette combinaison sera non nulle sur l'ensemble vide ($m_{1\textcircled{\cap}2}(\emptyset) > 0$). Cette valeur quantifie alors la discordance entre les sources de croyance, et est appelée **conflit**[73]. Une normalisation peut être effectuée, ramenant la règle de combinaison conjonctive à la **règle orthogonale de Dempster**[15][16] (notée $\oplus$) utilisée dans le modèle de Shafer[67][75] :

$$m_{1\oplus 2}(A) = (m_1 \oplus m_2)(A) = \begin{cases} \frac{m_{1\textcircled{\cap}2}(A)}{1-m_{1\textcircled{\cap}2}(\emptyset)}, & \forall A \subseteq \Omega \quad \text{si} \quad A \neq \emptyset \\ 0 & \text{si} \quad A = \emptyset \end{cases} \tag{3.25}$$

   Une autre façon d'écrire cette règle de combinaison est :

$$m(A) = \frac{\sum_{B\cap C=A} m_1(B)m_2(C)}{\sum_{B\cap C\neq 0} m_1(B)m_2(C)} \quad \forall A \subseteq \Omega \tag{3.26}$$

   Avec, pour mesure du conflit entre les sources :

$$K = \sum_{B\cap C=0} m_1(B)m_2(C) \tag{3.27}$$

5. La **règle d'affaiblissement simple** pour intégrer la fiabilité des sources. L'application de la règle de combinaison disjonctive peut s'avérer trop conservatrice lorsque les sources de fonctions de masse de croyance ne sont pas fiables. Elle peut ainsi mener à une fonction de masse de croyance complètement non-informative (vide). Dans certaines applications, il est cependant possible de quantifier la fiabilité ce qui permet d'appliquer une règle de combinaison moins conservatrice dans le cas de sources non fiables.

   La prise en compte de la fiabilité dans le cadre des fonctions de croyance porte le nom d'**affaiblissement** car le processus consiste à pondérer la masse des éléments d'une fonction de masse de croyance. Les premiers travaux sur

l'affaiblissement dans le cadre des fonctions de croyance ont été développés par Shafer[67], axiomatisés par Smets[74] et généralisés par Mercier et Denœux[56]. L'affaiblissement généralisé a été appelé affaiblissement contextuel.

L'affaiblissement simple de Shafer d'une fonction de masse de croyance m est défini comme la pondération de chaque masse $m(B)$ de la distribution par un coefficient $(1-\alpha)$ appelé fiabilité et où $\alpha \in [0,1]$ est le taux d'affaiblissement. Formellement :

$$m^{\alpha}(B) = \quad (1-\alpha).m(B), \qquad \forall B \subset \Omega \tag{3.28}$$

$$m^{\alpha}(\Omega) = \quad (1-\alpha).m(\Omega)+\alpha \tag{3.29}$$

Par le principe du minimum d'information, le reste de la masse (après pondération) est transféré sur l'élément d'ignorance Ω. Plus la source est fiable (et plus le taux d'affaiblissement α est faible), moins la fonction de masse de croyance est modifiée alors que, plus la fiabilité diminue, plus la fonction de masse de croyance issue de l'affaiblissement tend vers la fonction de masse de croyance vide.

La décision pignistique

A partir de l'ensemble des fonctions de croyance fournies par tous les capteurs, et suite à leur combinaison par la règle de Dempster, nous obtenons une fonction de croyance unique pour chacune des hypothèses possibles modélisant la connaissance du système. Il faut ensuite, à partir de cette connaissance, prendre une décision sur l'une des hypothèses.
Notons que les hypothèses du système ne représentent pas forcément les seules classes possibles (i.e. les 26 lettres de notre alphabet). Il est également possible de créer une nouvelle hypothèse qui correspondrait à l'évènement "aucune des hypothèses connues". On parle alors de **rejet**, c'est à dire que l'individu inconnu n'est

pas réinjecté dans la chaîne, il est tout simplement classé comme ”impossible à reconnaître”. La mesure du conflit des capteurs lors de la fusion de Dempster (voir l'equation 3.27) est aussi une mesure sur laquelle le système peut se baser pour rejeter l'individu.

Les systèmes de fusion d'information, qu'ils soient basés sur la théorie des probabilités, des possibilités ou des fonctions de croyance, ont pour finalité la prise de décision et l'analyse de cette décision. Prendre une décision consiste à choisir une hypothèse sur un **cadre de pari**, généralement le cadre de discernement Ω. La prise de décision peut être réalisée de façon automatique ou laissée à la responsabilité de l'utilisateur final (par exemple dans le cas de de l'aide au diagnostique dans le domaine de la Médecine).

Dans le cadre du modèle des croyances transférables, la phase de décision s'appuie sur la distribution de probabilités pignistiques notée $\mathrm{BetP}\{m\}$ à partir de la distribution de masse m[76]. La transformée pignistique consiste à répartir de manière équiprobable la masse d'une proposition B sur les hypothèses contenues dans B. Formellement :

$$\mathrm{Bet}\{m\} : \quad \Omega \rightarrow [0,1] \tag{3.30}$$

$$\omega_k \mapsto \mathrm{BetP}\{m\}(\omega_k) \tag{3.31}$$

où $\mathrm{BetP}\{m\}(\omega_k), \forall \omega_k \in \Omega$ est donnée par :

$$\mathrm{BetP}\{m\}(\omega_k) = \frac{1}{(1-m(\emptyset))} \sum_{B \subseteq \Omega, \omega_k \in B} \frac{m(B)}{|B|} \tag{3.32}$$

La décision est généralement prise en choisissant l'élément ω_k possédant la plus grande probabilité pignistique :

$$\omega_k = \arg\max_{\omega_k \in \Omega} \mathrm{BetP}\{m\}(\omega_k) \tag{3.33}$$

L'agent qui se base sur la transformée pignistique lors de la phase de prise de décision présente un comportement rationnel en maximisant l'utilité espérée.

Alternativement, les plausibilités et les crédibilités peuvent être utilisées lorsque l'agent présente une attitude plutôt optimiste ou pessimiste.

Dans le cas des plausibilités, on dit que l'agent a une attitude optimiste dans le sens où la plausibilité d'un élément représente la valeur maximale de la masse de croyance sur cet élément qui pourrait être atteinte si des informations supplémentaires parvenaient au système de fusion (ce dernier utilisant la combinaison conjonctive pour intégrer cette nouvelle information). La distribution de probabilités sur les singletons à partir de laquelle une décision peut être prise est alors donnée $\forall \omega_k \in \Omega$ par :

$$\mathrm{PIP}\{m\}(\omega_k) = \frac{pl(\omega_k)}{\sum_{\omega_k \in \Omega} pl(\omega_k)} \tag{3.34}$$

Notons que la plausibilité d'un singleton $pl(\omega_k)$ est égale à sa communalité $q(\omega_k)$. De la même manière, un critère adapté à un agent "pessimiste" peut être obtenu en utilisant les crédibilités. La crédibilité d'un singleton $bel(\omega_k)$ est simplement égale à sa masse après normalisation de la fonction de masse de croyance m. Remarquons que pour tout élément $A \in 2^{\Omega}$:

$$bel(A) \leq \mathrm{BetP(A)} \leq pl(A) \tag{3.35}$$

et en particulier lorsque les plausibilités et les crédibilités sont calculées sur les singletons.

Pour le problème de prise de décision, nous supposons avoir une fonction de croyance m sur Ω qui résume l'ensemble des informations apportées sur la valeur de la variable y. La décision consiste à choisir une action a parmi un ensemble fini d'actions $\mathcal{A}$. Une fonction de perte $\lambda : \mathcal{A} \times \Omega \rightarrow \bullet$ est supposée définie de telle manière que $\lambda(a, \omega)$ représente la perte encourue si l'action a est choisie lorsque $y = \omega$. A partir de la probabilité pignistique, chaque action $a \in \mathcal{A}$ peut être associée à un risque défini comme le coût espéré relatif à p_m si a est choisie :

$$R(a) = \sum_{\omega \in \Omega} \lambda(a, \omega) p_m(\omega) \tag{3.36}$$

L'idée consiste ensuite à choisir l'action qui minimise ce risque, généralement appelé risque pignistique. En reconnaissance de formes, $\Omega = \{\omega_1, ..., \omega_Q\}$ est l'ensemble des classes et les éléments de $\mathcal{A}$ sont généralement les actions a_q qui consistent à assigner le vecteur inconnu à la classe ω_q. Il est démontré que la minimisation du risque pignistique R conduit à choisir la classe ω_0 de plus grand probabilité pignistique[17]. Si une action supplémentaire de rejet a_0 de coût constant λ_0 est possible, alors le vecteur est rejeté si $p_m(\omega_0) < 1 - \lambda_0$.

Prise de décision à partir de fonctions de croyance

Nous avons vu jusqu'à présent la construction de la croyance et la décision pignistique. Or, par rapport à la chaîne de traitements, le système dispose pour décider de la précision des attributs, de la précision des classifieurs et des données en sortie de l'étage de classification. Voyons à présent comment construire, en vue de la prise de décision, les fonctions de croyances sur les différentes hypothèses à partir de ces données.

Au niveau crédal du modèle des croyances transférables, afin d'obtenir les fonctions de croyance à partir des données d'apprentissage, deux familles de techniques sont généralement utilisées : les méthodes basées sur la vraisemblance qui utilisent l'estimation des densités et une méthode basée sur la distance dans laquelle les jeux de masses sont construits directement à partir des distances aux vecteurs d'apprentissage.

1. Méthodes basées sur la vraisemblance.
   Les densités de probabilités $F(x|\omega_q)$ conditionnellement aux classes ω_q sont supposée connues. En ayant observé x, la fonction de vraisemblance $L(\omega_q|x)$ est une fonction de Ω dans $[0, +\infty[$ définie par $L(w_q|x) = f(x|\omega_q)$, pour tout $q \in [1, ..., Q]$. A partir de L, Shafer[67] a proposé de construire une

fonction de croyance sur Ω définie par sa fonction de plausibilité comme :

$$pl(A) = \frac{\max_{\omega_q \in A}[L(\omega_q|X)]}{\max_q[L(\omega_q|X)]} \quad \forall A \subseteq \Omega \tag{3.37}$$

A partir de considérations axiomatiques, Appriou[1] a proposé une autre méthode basée sur la construction de Q fonctions de croyances $m_q(.)$. L'idée consiste à prendre en compte de manière séparée chaque classe et à évaluer le degré de croyance accordé à chacune d'entre elles. Dans ce cas, les éléments focaux de chacune des fonctions de croyance m_q sont les singletons $\{\omega_q\}$, les sous ensembles complémentaires $\bar{\omega_q}$ et Ω. Appriou obtient ainsi deux modèles différents :

– Modèle 1 :

$$m_q(\{\omega_q\}) = \alpha_q \frac{R.L(\omega_q|x)}{1 + R.L(\omega_q|x)} \tag{3.38}$$

$$m_q(\overline{\omega_q}) = \alpha_q \frac{1}{1 + R.L(\omega_q|x)} \tag{3.39}$$

$$m_q(\Omega) = 1 - \alpha_q \tag{3.40}$$

– Modèle 2 :

$$m_q(\{\omega_q\}) = 0 \tag{3.41}$$

$$m_q(\overline{\omega_q}) = \alpha_q(1 - R.L(\omega_q|x)) \tag{3.42}$$

$$m_q(\Omega) = 1 - \alpha_q(1 - R.L(\omega_q|x)) \tag{3.43}$$

Ici, α_q est un coefficient qui peut-être utilisé pour modéliser une information complémentaire (comme par exemple la fiabilité d'un capteur), et R est une constante de normalisation qui est choisie dans $]0, (\sup_x \max_q(L(\omega_q|x)))^{-1}]$. En pratique, les performances de ces deux modèles semblent être équivalentes. Cependant, Appriou recommande l'utilisation du modèle 2 qui a l'avantage d'être consistant avec le "théorème de Bayes généralisé" proposé par Smets dans [74]. A partir de ces Q fonctions de croyance et en utilisant la règle de combinaison de Dempster, une fonction de croyance

unique m est obtenue par $m = \bigoplus_q m_q$.

2. Méthode basée sur la distance.

   La seconde famille de modèles se base sur des informations de distance. Dans cette dernière, citons comme exemple l'extension de l'algorithme des k plus proches voisins qui a été introduit par T. Denœux dans [18]. Dans cette méthode, une fonction de croyance m_i est directement construite en utilisant les informations apportées par les vecteurs x^i situés dans le voisinage du vecteur inconnu x par :

$$m^i(\{\omega_k\}) = \alpha^i \phi^i(d^i) \tag{3.44}$$

$$m^i(\Omega) = 1 - \alpha^i \phi^i(d^i) \tag{3.45}$$

$$m^i(A) = 0, \forall A \in 2^\Omega \backslash \{\{\omega_k\}, \Omega\} \tag{3.46}$$

   où d^i est la distance euclidienne du vecteur inconnu au vecteur x^i, α^i un paramètre associé au i-ème voisin et $\phi^i(d^i) = e^{-\gamma^i (d^i)^2}$ avec γ^i un paramètre positif (rappelons que l'opérateur $\backslash$ signifie "privé de"). La méthode des k plus proches voisin permet d'obtenir k fonctions de croyance à agréger par la règle de combinaison pour la prise de décision.

3.4 Proposition

3.4.1 Intégration de la précision

Principe

La théorie de l'évidence présente plusieurs éléments d'intérêt dans notre quête d'une mesure de la précision associée à la décision, ainsi que dans l'intégration d'une telle mesure dans cette prise de décision. Le premier d'entre eux est certainement celui de pouvoir proposer un bornage pessimiste et optimiste de la décision, ce qui peut en première instance être rapproché de la tolérance souhaitée en association avec la décision. Le second correspond ensuite à l'introduction du facteur d'affaiblissement dans le propos. Par son entremise, il est possible d'intégrer

un facteur prenant en compte la fiabilité de l'opérateur fournissant l'information. Cette fiabilité, ou confiance, est déduite d'un apprentissage effectué au préalable, c'est à dire hors ligne. Or tel qu'est construite l'approche, quelque soit la valeur établie, quelque soit l'objet à partir duquel est établie la mesure, celle ci sera considérée au même niveau d'importance que toutes celles issues de l'opérateur.

Il ne faut pas voir ici une critique de la construction du facteur d'affaiblissement qui a été introduit pour permettre la combinaison de différentes sources d'information, tels que des capteurs ou des bandes de fréquences dans des IRM. Derrière cette proposition se place l'hypothèse de linéarité de comportement de la source d'information face au bruit et/ou à l'information à acquérir. Cette hypothèse certainement valide dans le contexte où a été développée cette théorie n'est en revanche pas valide dans notre contexte plus générique.

Deux solutions s'offrent à nous pour développer notre proposition, la première serait d'intégrer un second facteur dans la formulation qui ne serait dépendant que de l'objet à mesurer au travers de l'opérateur j. La seconde solution consiste à ne conserver qu'un seul facteur d'affaiblissement intégrant les deux aspects à prendre en compte :

- L'apport du type d'information considéré dans la prise de décision. Ce qui correspond à une **pertinence *a priori***, c'est à dire uniquement liée à l'opérateur et considérée à la suite de l'apprentissage avant la phase de prise de décision proprement dite (phase dite en ligne).

- La qualité de la mesure, que nous appelons précision. Cette mesure de précision est directement dépendante de l'objet à mesurer vu au travers de l'opérateur, mesure effectuée pour prendre en compte les non linéarités de comportement de l'opérateur ou du capteur (i.e. cas des réponses en fonction des différent tissus dans l'imagerie médicale). Cette mesure de précision

sera donc appelée **mesure de précision *a posteriori*** car établie à l'issu de la mesure.

Dans les deux cas, nous proposons de conserver l'écriture d'Appriou. La combinaison des différents éléments d'affaiblissement étant écrite sous forme multiplicative. Pour des raisons de temps nécessaire, nous n'avons pas associé dans ce travail ces deux parties liées à l'estimation de la précision totale(*a priori* et *a posteriori*). Entre autre raison à cette décision intervient notamment la problématique de l'estimation de la précision *a priori*, fortement dépendante de la construction de la base d'apprentissage.

Le chapitre suivant abordera une partie des pistes envisagées pour cette question dans le cadre des classifieurs. Néanmoins, la littérature sur le sujet étant déjà importante, nous avons estimé qu'il existait déjà de nombreuses bases de réponse à cette question.

Mise en œuvre

Les fonctions de croyances par classe/classifieur/attribut sont combinées (Dempster) afin de construire une fonction de croyance par classe. La prise de décision est ensuite faite selon l'équation 3.33. Le mécanisme de décision, est schématisé par la figure 3.9. Cette figure représente l'enchaînement des processus ainsi que les échanges entre eux. Nous avons représenté ici chacune des variables du formalisme, son origine et sa destination. La source du flux illustrant l'individu à analyser (x^*) n'est pas représentée car nous ne traitons pas de cette partie. x^* peut être un objet injecté directement dans le processus, mais il peut être issu de traitements postérieurs.

L'exploitation de l'expression d'Appriou, et notre version modifiée de l'expression d'Appriou exploite des étiquettes discrètes basées sur la fonction de vrai-

semblance $L(\omega_q|x)$ liant un individu x à son appartenance à la classe ω_q.

Afin d'obtenir ces fonctions et selon le schéma de la figure 3.8 ces étiquettes sont dans notre cas issues d'un étage de classification liant l'individu à la classe ω_q. Classiquement dans ce cadre, la fonction de vraisemblance $L(\omega_q|x)$ est alors liée à la distance $C_i(\omega_q|x)$ de l'individu x à la classe ω_q selon le classifieur i (suivant le classifieur, la distance peut être remplacée par un degré d'appartenance ou une probabilité). Cependant, comme nous l'avons exprimé en introduction, la capacité d'associer un individu à une étiquette par un classifieur est dépendante du jeu d'attributs utilisé. De la même façon, chaque classifieur propose un partitionnement différent de l'espace des attributs en fonction du modèle de classe utilisé ou des métriques mises en œuvre. Notre écriture reliera donc la fonction de vraisemblance $L(\omega_q|x)$ à la distance $C_i(\omega_q|A_j(x))$, où $A_j(x)$ représente le vecteur d'attributs fourni par l'opérateur j (moments de Zernike à l'ordre 15 par exemple ou histogramme couleur sur 32 bins obtenu par fuzzy C-means etc.) et C_i faisant référence au classifieur i.

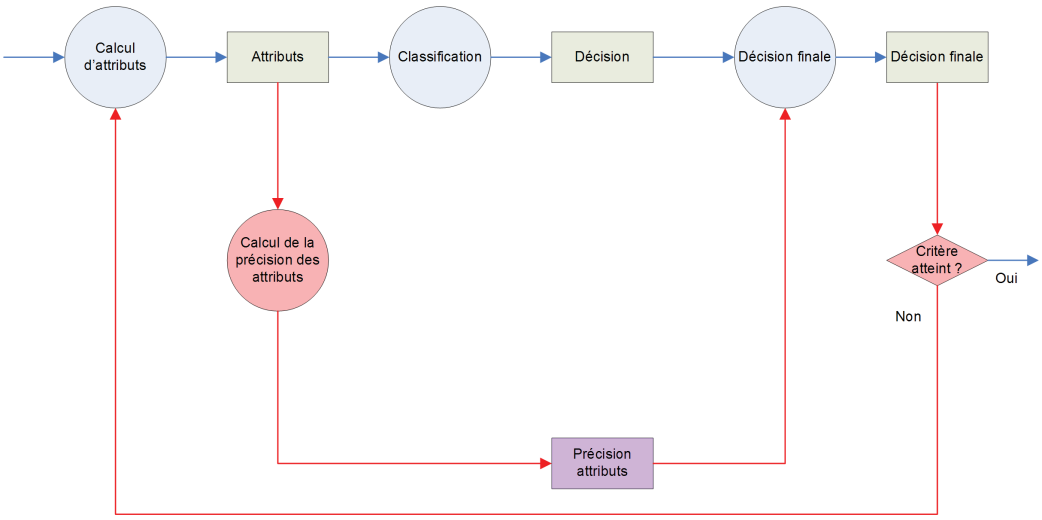

FIGURE 3.8 – Chaîne de traitements intégrant la précision des attributs

Le choix de cette écriture nous permet dés lors de combiner plusieurs éti-

quettes issues de classifieurs différents et/ou d'attributs différents. Selon le protocole choisi, les différents classifieurs seront mis en compétition directement ou une étape de sélection dynamique du meilleur d'entre eux sera opérée avant cette étape de fusion. Dans ce dernier cas, la compétition fusionne les étiquettes issues de différents attributs vus au travers du meilleur classifieur pour chacun d'entre eux. L'expression à laquelle nous aboutissons est directement déduite de l'équation 3.33 :

$$\left\{\begin{array}{ll} m_{ijq}(\{\omega_q\}) & = 0 \\ m_{ijq}(\overline{\omega_q}) & = \alpha_{jq}(1 - R_{ijq}.C_i(\omega_q|A_j(x))) \\ m_{ijq}(\Omega) & = 1 - m_{ijq}(\overline{\omega_q}) \end{array}\right. \quad (3.47)$$
$$\forall i \in [1,n_c], \forall j \in [1,n_a], \forall q \in [1,n_q]$$

Appriou intègre R, un paramètre de normalisation de la formule établit pour que la masse de croyance $m_{ijq}(\overline{\omega_q})$ soit comprise entre 0 et 1. R est donc une valeur comprise entre 0 et la valeur maximale qui peut être estimée pour $C_i(\omega_q|A_i(x))$. Cette valeur de R est unique pour le classifieur i et correspond à la plus grande distance identifiée ou possible entre un individu et les classes définies pour ce classifieur, d'où :

$$\begin{array}{c} \text{pour que } 0 \leq 1 - \dfrac{x}{y} \leq 1 \text{ il faut } y \geq x \\ \text{d'où } y = \sup\limits_x(\max\limits_q(C_i(\omega_q|A_j(x)))) \\ 0 < R \leq \dfrac{1}{\sup_x(\max_q(C_i(\omega_q|A_j(x))))} \end{array} \quad (3.48)$$

La masse de croyance que nous venons d'établir est $m_{ijq}(\{\omega_q\})$, c'est à dire la masse de croyance associée au fait que la classe q soit attribuée à l'individu x par le classifieur i pour le vecteur d'attributs j. Or, ce qui nous intéresse c'est la décision d'affectation finale de la classe q à l'individu x et par voie de conséquence la masse de croyance en la classe q, $m_q(\{\omega_q\})$, et les masses associée $m_q(\overline{\omega_q})$ et $m_q(\Omega)$:

$$\begin{cases} m_q(\{\omega_q\}) \\ m_q(\overline{\omega_q}) \\ m_q(\Omega) \end{cases} \quad \forall q \in [1,n] \tag{3.49}$$

A partir du modèle de combinaison de Dempster, nous formalisons cette prise de décision à partir des q décisions intermédiaires sur un ensemble de n étiquettes pour un jeu unique d'attributs :

$$\begin{cases} m(\{\omega_k\}) & = \bigoplus_{q=1}^{n_q} m_{iq}(\{\omega_k\}) \\ m(\overline{\omega_k}) & = \bigoplus_{q=1}^{n_q} m_{iq}(\overline{\omega_k}) \end{cases} \quad \forall q \in [1,n], \forall i \in [1,n_c] \tag{3.50}$$

et :

$$m(\Omega) = 1 - \sum_{i=1}^{n} (m(\{\omega_i\}) + m(\overline{\omega_i})) \tag{3.51}$$

De façon plus générale, en présence de plusieurs jeux d'attributs issus d'opérateurs différents (ou paramètres différents), la règle de combinaison de Dempster nous permet d'aboutir à la formulation générale :

$$\begin{cases} m(\{\omega_k\}) & = \bigoplus_{q=1}^{n_q} \bigoplus_{i=1}^{n_c} \bigoplus_{j=1}^{n_a} m_{ijq}(\{\omega_k\}) \\ m(\overline{\omega_k}) & = \bigoplus_{q=1}^{n_q} \bigoplus_{i=1}^{n_c} \bigoplus_{j=1}^{n_a} m_{ijq}(\overline{\omega_k}) \end{cases} \tag{3.52}$$

Le diagramme de flots de données de la figure 3.9 illustre le fonctionnement de cette écriture.

3.4.2 Bouclage

A présent que notre propos est écrit de façon formelle, voyons ce qu'il en est de son exploitation, et plus précisément du bouclage sur la chaîne de traitements. Sur quoi allons-nous nous baser pour décider de remettre en question l'enchaînement réalisé pour en tenter un autre ?

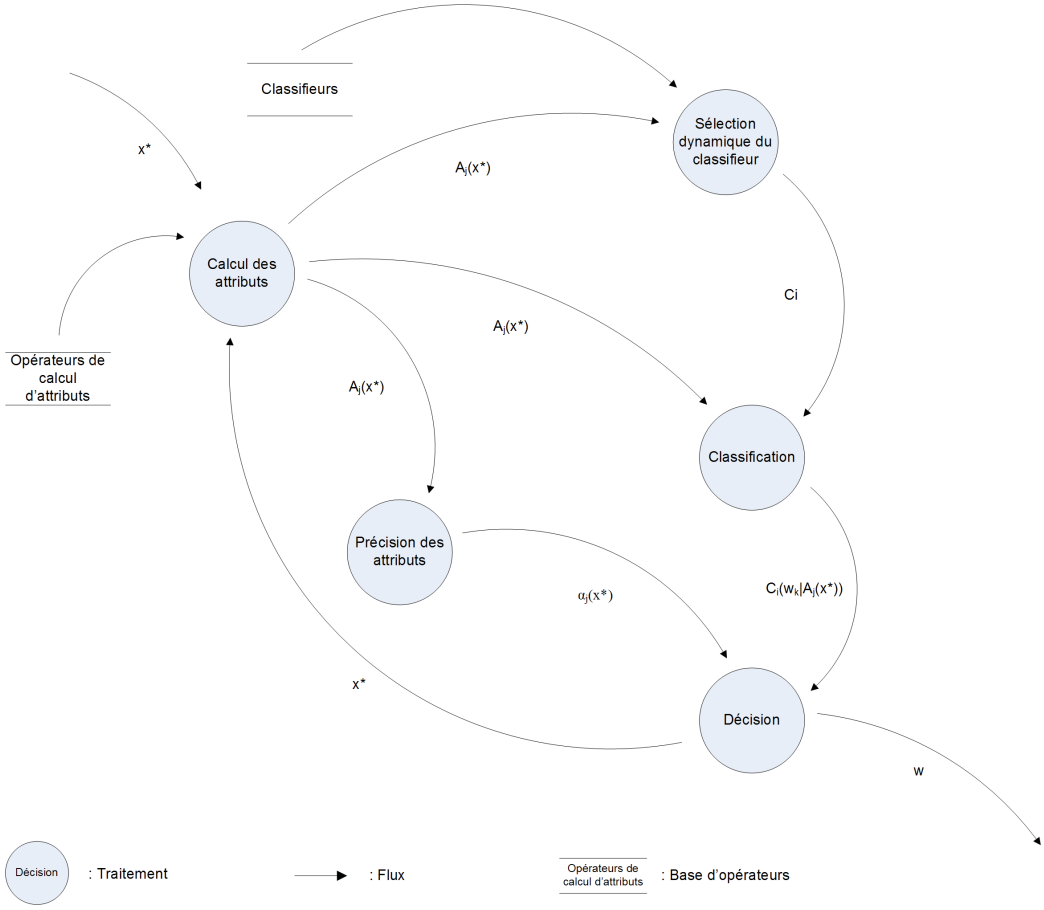

FIGURE 3.9 – Diagramme des flots de données de la chaîne de traitements

Comme nous l'avons vu plus haut, lors de la présentation des outils théoriques utilisés, il existe plusieurs manières de rejeter un individu, c'est à dire de décider que celui-ci ne peut être attribué à l'une des classes connues (i.e. les lettre de l'alphabet grec). Le rejet peut par exemple se décider en rajoutant une classe "inconnu". Nous avons également vu que l'opérateur de fusion des fonctions de croyance $\oplus$ nous indique un niveau de conflit K entre les sources. Ainsi, il devient possible de décider d'un rejet si le conflit est trop important, c'est à dire quand toutes les sources d'information du système tendent vers un désaccord global.

Nous avons fait le choix d'intégrer ces deux méthodes dans notre application. De façon séquentielle, le système décidera d'abord du niveau de conflit, et si celui-ci est acceptable, calculera les probabilités pignistiques de chaque hypothèse, y compris l'hypothèse "individu inconnu".
Ensuite, en cas de décision de rejet, l'individu sera systématiquement réinjecté dans la chaîne afin de tester d'autres enchaînements, jusqu'à ce qu'il soit attribué à une classe, ou bien qu'il n'y ait plus de combinaisons à tester.

En faisant cela, nous semblons nous éloigner de notre idée première qui est de prendre une décision selon la précision et le temps fixés par l'utilisateur en amont de la chaîne. Cependant, le propos est ici de valider l'écriture théorique proposée. L'intégration de données telles que le temps restant, le temps estimé pour chaque traitement ou encore la charge de la machine sont du domaine du traitement, c'est à dire qu'ils relèvent d'une application algorithmique pure.

Voyons à présent les résultats obtenus.

3.5 Résultats

Dans ce chapitre, et plus globalement dans ce travail, nous nous intéressons à la mesure de précision issue d'une mesure par un opérateur de type extraction d'attribut et à celle issue d'une décision de classification pour une prise de décision en fin de chaîne de traitement d'images.

Parmi les résultats de validation que nous allons présenter dans cette section, nous allons d'abord chercher à expliquer l'impact de cette mesure de précision dans la capacité du système à décider. Nous nous intéresserons ensuite au traitement d'un cas complexe (la base MNIST) pour étudier les résultats obtenus par cette nouvelle approche ainsi que les différentes possibilités induites, notamment les bouclages.

3.5.1 Protocole

La chaîne de traitements choisie se compose d'opérateurs de calcul d'attributs et d'opérateurs de classification (voir figure 3.10). Comme nous l'avons vu, nous ne faisons pas le choix d'étudier une chaîne de traitement d'image complète (avec les étapes d'amélioration, de débruitage, de segmentation etc.), notre propos se rapportant principalement à la prise de décision finale quant au bouclage.

Pour l'étage de calcul d'attributs, nous avons choisi des méthodes de calcul de moments statistiques, qui ont la caractéristique d'être inversibles, ce fait étant celui que nous exploitons pour le calcul de la précision des attributs générés (voir chapitre 5). Nous utiliserons les méthodes suivantes :

- moments cartésiens à l'ordre 15
- moments centrés à l'ordre 15
- moments de Zernike à l'ordre 15

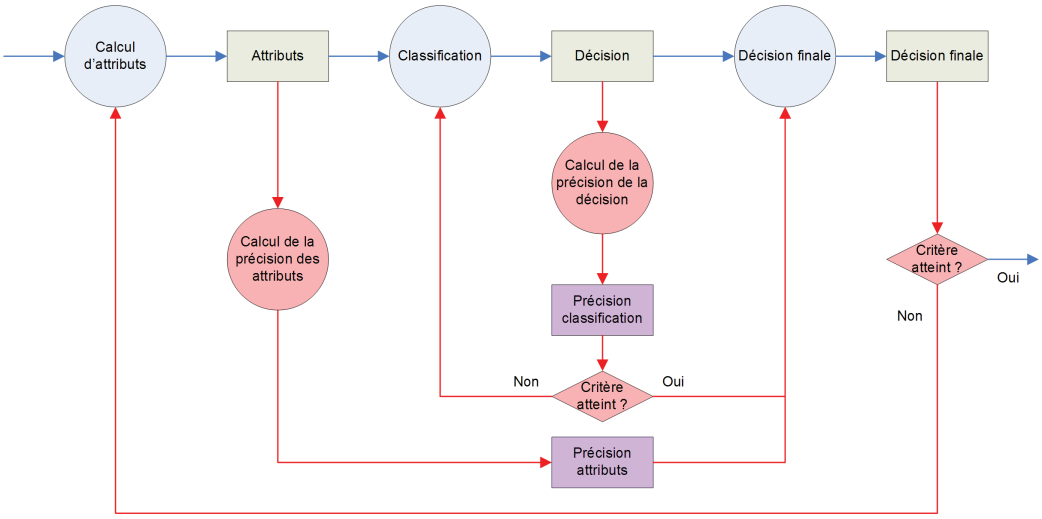

FIGURE 3.10 – Chaîne de traitements

Le choix de l'ordre des moments est arbitraire, décidé après analyse de l'évolution de la précision en fonction de l'ordre et arrêté sur la précision maximale globale (voir chapitre 5).

En ce qui concerne l'étage de classification, nous faisons le choix de 6 méthodes basiques, peu coûteuses en temps de calcul par rapport à certains classifieurs complexes (comme les Support Vector Machines ou SVM). Ce choix a été arrêté afin d'étudier le gain obtenu par collaboration d'opérateurs tant au niveau local (sur le résultat de la classification) qu'au niveau global (en fin de chaîne). Le travail réalisé sur le calcul de la précision des classifieurs ainsi que son exploitation pour la sélection dynamique d'opérateurs de classification est détaillé au chapitre 4. Les 6 méthodes choisies sont ici :

- QBNC : Quadratic Bayes Normal Classifier (Bayes-Normal-2)
- FLSLC : Fisher's Least Square Linear Classifier
- NMC : Nearest Mean Classifier
- KNNC : k Nearest Neighbors Classifier
- ParzenC : Parzen Classifier
- BayesN : Naive Bayesian Classifier

Afin de valider notre propos par rapport au contexte de la reconnaissance de caractères manuscrits, et de confronter nos résultats à ceux obtenus par la communauté, nous travaillons avec la base de caractères MNIST, présentée au chapitre 2.

L'idée de la combinaison est d'améliorer les résultats globaux de chaque méthode en leur faisant compenser leurs erreurs respectives. Nous utiliserons donc le taux d'erreur global lors de la décision finale comme critère de comparaison entre les méthodes. D'autre part, le fait d'intégrer le modèle des croyances transférables apporte la capacité à rejeter les individus pour lesquels l'information est jugée non satisfaisante. Le taux de rejet des individus lors de la prise de décision finale sera notre second critère de comparaison entre les différentes approches, après le taux d'erreur global.

3.5.2 Résultats théoriques

Influence de la précision

Notre propos est d'étudier l'influence de la précision de l'information sur le choix des opérateurs ainsi que sur la décision finale. Nous pouvons dans un premier temps simuler une partie des résultats en partant d'un cas simple. Supposons un problème à deux classes, pour lesquelles les masses de croyances seraient symétriques. Nous aurions donc : $m(\{\omega_2\}) = 1 - m(\{\omega_1\})$. A partir du second modèle d'Appriou (équation 3.43), nous générons l'évolution des deux masses de croyance, avec pour chaque simulation, une valeur différente de α comprise entre 0 et 1. Le résultat obtenu est présenté avec la figure 3.11.

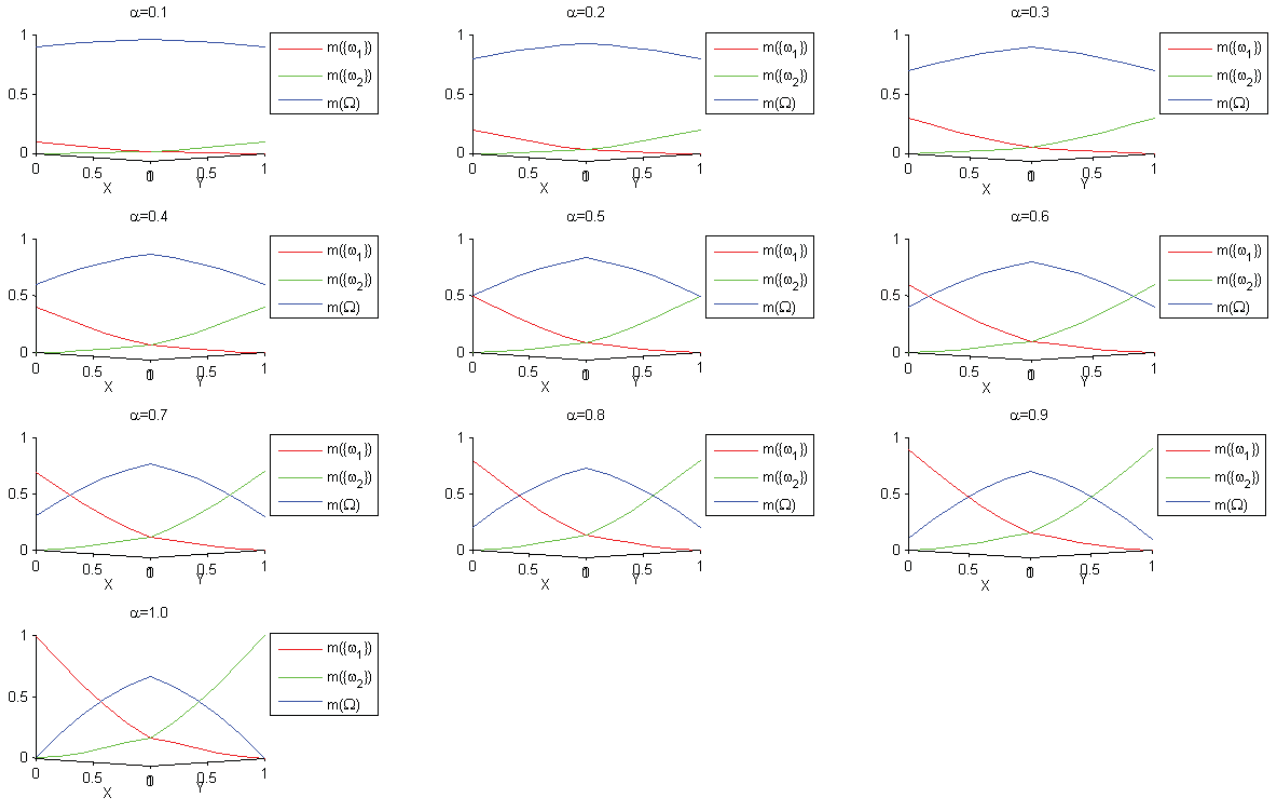

FIGURE 3.11 – Fonctions de masses pour un problème à 2 classes/Influence de la précision

	Moments cartésiens	Moments centrés	Moments de Zernike
Moyenne	0.5419	0.5419	0.5662
Ecart-type	0.1995	0.1995	0.0670

TABLE 3.2 – Précision moyenne et écart-type pour les différentes familles de moments

Nous constatons que pour une valeur de α inférieure à 0.5, la masse de l'univers des croyances Ω est supérieure aux masses de ω_1 et ω_2. Cela illustre une incertitude, une incapacité à décider entre les deux cas. En pratique, le résultat sera un rejet systématique de tous les individus ayant une mesure de précision inférieure à $0,5$. Nous démontrons donc bien que pour une précision trop faible de l'information, le système rejettera l'individu.

Le coefficient d'affaiblissement α étant issu de la combinaison entre la précision *a priori* et la précision *a posteriori* suivant une formulation de type multiplicative se trouvera comme ayant tendance à rejeter trop facilement l'attribut pour la prise de décision. Nous trouvons ici clairement une limite actuelle de la proposition, même si celle si se trouve cohérente du point de vue d'une décision prise en considérant la valeur de l'information proprement dite et la pertinence de celle ci dans cette prise de décision.

Intérêt de la sélection dynamique des opérateurs

Par rapport à la base MNIST, la précision moyenne (et l'écart-type de cette précision) calculée pour chaque famille de moments à l'ordre 15, est représentée par la table 3.2.

La table 3.2 et les résultats présentés à la section précédente indiquent que les attributs utilisés ne sont au final que très moyennement informatifs (les valeurs ne caractérisent que partiellement les objets concernés). Finalement, ces résultats

expliquent en partie les limites très rapidement atteintes par ces attributs dans une tache de décision directe (à l'issu du classifieur) pour cette base.

Avant de présenter l'utilisation du critère de précision en théorie des croyances, observons comment évolue ce critère sur nos données de test. Les figures 3.12 et 3.13 montrent pour chaque classe des jeux de données, les histogrammes cumulés de la précision pour les classes reconstruites avec les moments cartésiens d'ordre 15 (figure 3.12 et avec les moments de Zernike d'ordre 15 (figure 3.13. Ces courbes indiquent le taux d'individus pour chaque classe pour lesquels la précision est inférieure à un certain seuil. En particulier, nous montrons que 7% et 30% des attributs construits respectivement avec les moments cartésiens et les moments de Zernike ont une précision inférieure à 50% ! Bien sûr, en prenant les moments à un ordre supérieur, les taux seraient meilleurs mais de tels changements augmentent le coût de calcul.

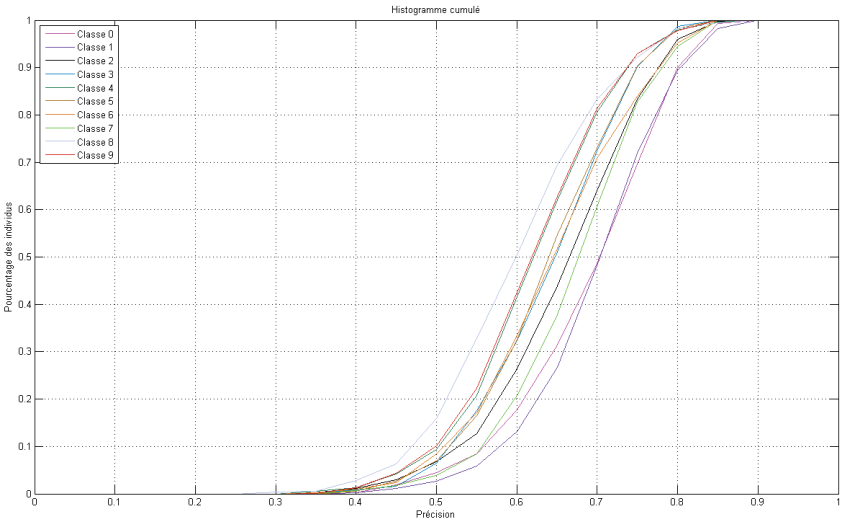

FIGURE 3.12 – Histogramme cumulé de la précision par classe après reconstruction avec les moments cartésiens

Ces premiers résultats montrent que tous les individus ne sont pas équiva-

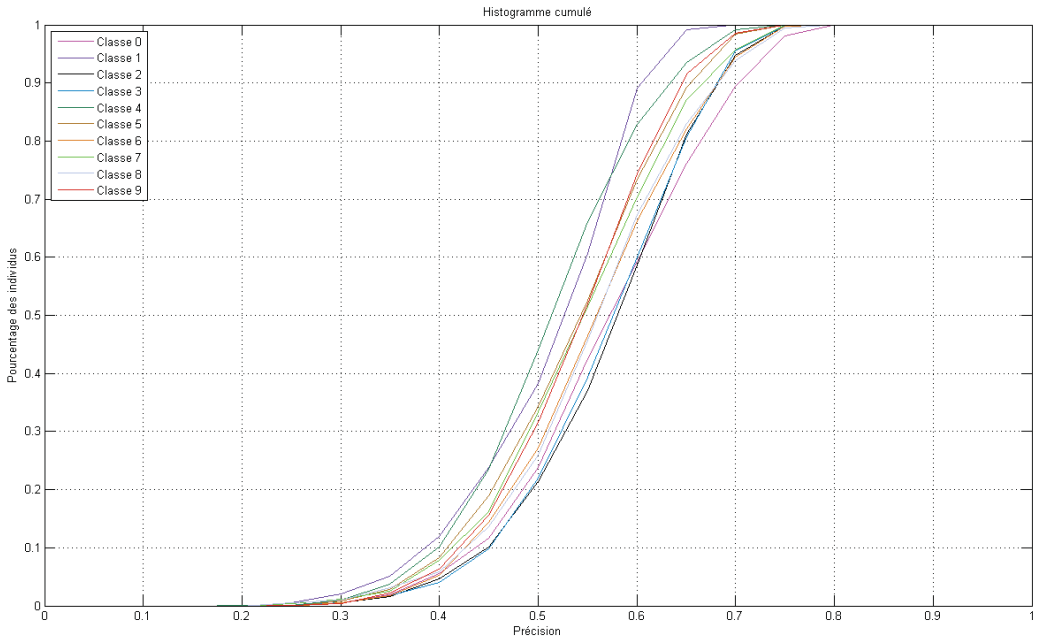

FIGURE 3.13 – Histogramme cumulé de la précision par classe après reconstruction avec les moments de Zernike

lents face aux moments. Comme tous les symboles sont des chiffres, les moments cartésiens semblent bien adaptés, mais 50% des individus ont une précision inférieure à 70%, ainsi la capacité à les classifier sera bonne mais non optimale. La disparité entre les classes est importante, environ 10% entre la classe des ”8” (la moins précise) et les classes de ”0” et des ”1” (les plus précises). Comme attendu, les classes les moins précises sont celles associées aux symboles les plus complexes (”8”,”9”,”4”) et les plus précises aux symboles les plus simples (”0”,”1”,”7”). Dans le cas des moments de Zernike, la disparité entre les classes est plus faible que pour les moments cartésiens mais globalement, la précision est également plus faible (la précision maximale est égale à 80% au lieu de 90%) et 50% des individus présentent une précision entre 50% et 60%, ce qui est très faible. Mais l'intérêt majeur est que l'ordre des classes de symboles n'est pas le même entre les deux schémas selon le critère de précision, ce qui amène à l'idée que la combinaison des méthodes peut aider à compenser leurs erreurs.

C'est là l'intérêt de la sélection dynamique des opérateurs : la compensation des erreurs (et rejets) des uns par les autres, chaque opérateur ne commettant pas systématiquement les mêmes erreurs que les autres, c'est ce que nous appelons la **diversité** (voir le chapitre 4.2.2). Suivant le protocole présenté ci-dessus, la figure 3.14 illustre l'évolution des erreurs communes à toutes les méthodes en fonction du nombre de classes. La figure 3.15 illustre quant à elle l'évolution des individus rejetés communs à toutes les méthodes, en fonction du nombre de classes. Dans ce type de figure, le traitement réalisé permet de tirer aléatoirement n classes parmi les 10 pour effectuer le calcul, puis de recommencer plusieurs fois ce procédé pour extraire les valeurs moyennes en ayant pris en compte les différentes combinaisons.

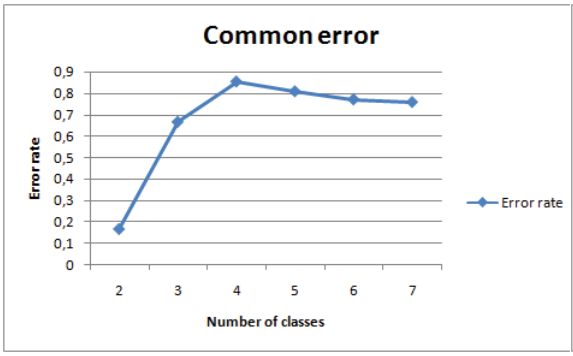

FIGURE 3.14 – Nombre d'individus mal reconnus par toutes les méthodes en fonction du nombre de classes

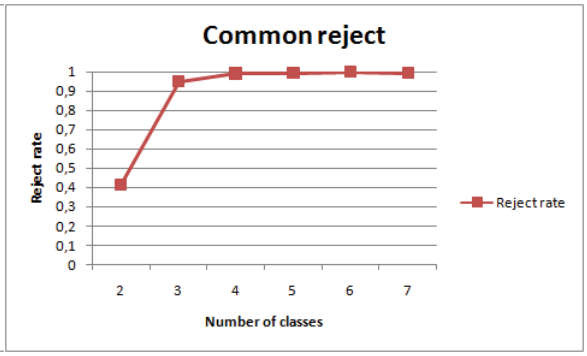

FIGURE 3.15 – Nombre d'individus rejetés par toutes les méthodes en fonction du nombre de classes

La première observation que nous pouvons faire, est que le nombre d'individus rejetés par toutes les méthodes est très faible pour un problème à deux classes. Ensuite, ce nombre croît au delà des 80% d'individus en commun rejetés pour un problème à quatre classes, pour redescendre doucement, autour de 80% ensuite. Ceci nous montre que pour un peu plus de 20% des individus, les méthodes compenseront leurs erreurs entre elles.

En ce qui concerne les rejets, les trois méthodes rejettent quasiment toutes les mêmes individus à partir de quatre classes. Nous n'avons donc manifestement pas de compensation des rejets entre ces méthodes.

Par rapport à ces résultats préliminaires, nous pouvons nous attendre à certains comportements lors des tests grandeur nature. En premier lieu, l'augmentation du nombre de classes devrait répartir équitablement la croyance en toutes les classes, ce qui impliquera une baisse du nombre des rejets. Ces rejets devraient d'ailleurs se compenser entre les enchaînements, jusqu'au maximum de quatre classes.

Dans la boucle complète, à partir du moment où l'on réalise un bouclage simple, l'erreur finale ne peut pas être inférieure à l'erreur de la première méthode testée car un individu mal classé la première fois ne sera pas remis en cause à la fin pour être réévalué à l'aide d'une autre approche.

3.5.3 Comportement expérimental

Voyons à présent les résultats obtenus par expérimentation selon le protocole indiqué plus haut. La figure 3.16 représente le taux d'erreur obtenu par le classifieur k plus-proches voisins (KPPV) seul, pour chaque famille d'attributs. Nous retrouvons en figure 3.17, le taux d'erreur du même KPPV seul, suivi d'une prise de décision selon le modèle des croyances transférables (MCT) sans utilisation de la précision. Pour cette même méthode, nous pouvons observer l'évolution des rejets sur la figure 3.18. Les résultats pour ce même classifieur accompagné d'une décision selon le modèle des croyances transférable avec cette fois l'intégration de la précision des attributs sont représentés par les figures 3.21 et 3.22.

Résultat initial avec un seul classifieur

La figure 3.16 illustre le comportement d'un k plus-proche voisin sur la base MNIST à partir des attributs construits avec les moments cartésiens, les moments centralisés et les moments de Zernike.

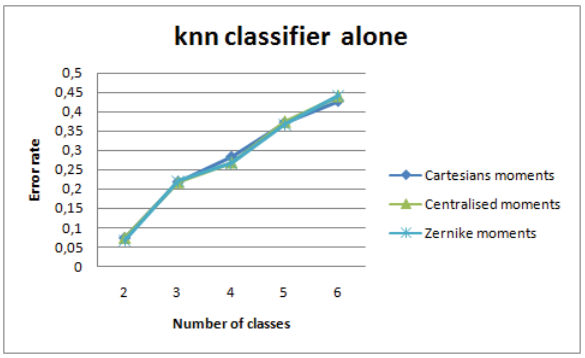

FIGURE 3.16 – Taux d'erreur pour le KPPV seul

Les résultats obtenus sont classiques et indiquent que plus le nombre de classe à discriminer est grand, plus l'erreur de classification croît pour atteindre de façon asymptotique une valeur de l'ordre de 42%. Même si les moments cartésiens présentent une inflexion de l'ordre de 5% supérieure pour une décision à 3 ou 4 classes, les comportements sont assez proches.

La courbe de la figure 3.16 nous servira de référence pour les résultats suivant.

Résultats expérimentaux sans utiliser l'apport de la précision

Avec intégration du Modèle de Croyance Transférable

La figure 3.17 illustre le comportement du même classifieur (KPPV) que précédemment, suivi d'une prise de décision finale selon le modèle des croyances transférables (MCT). Ce calcul est effectué sans utiliser la précision des attributs calculés à partir des individus à reconnaître, et uniquement à partir des probabilités *a posteriori* issues de l'étage de classification. La décision porte sur l'acceptation

ou le rejet de la décision prise par le classifieur.

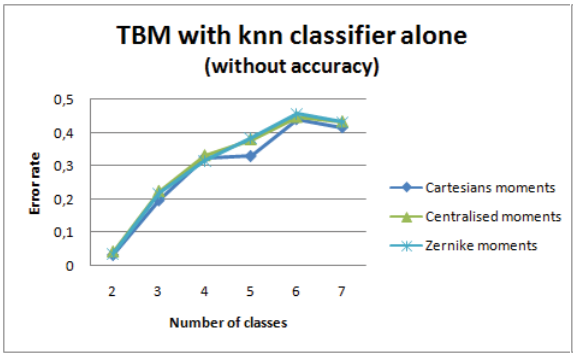

FIGURE 3.17 – Taux d'erreur pour KPPV+MCT sans utilisation de la précision

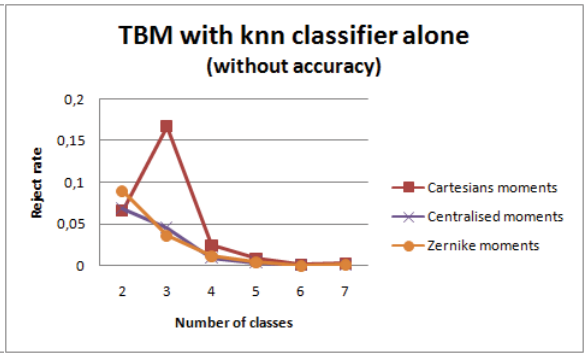

FIGURE 3.18 – Taux de rejet pour KPPV+MCT sans utilisation de la précision

L'apport intéressant du modèle des croyances transférables par rapport à un classifieur seul (KPPV) est la robustesse du système. On rajoute à la chaîne de traitements un étage de décision capable de produire un rejet au profit de la décision finale. L'évolution de ce rejet en fonction du nombre de classes est présenté en figure 3.18. Rappelons que la notion de rejet correspond au cas où la possibilité (ou probabilité) qu'un individu appartienne à une des classes connues est trop faible pour décider d'une quelconque affectation. C'est typiquement le genre de cas où un système industriel demandera l'intervention d'un opérateur humain.

Tandis que cette méthode ne semble pas apporter d'amélioration pour l'utilisation des moments cartésiens, nous notons un gain de 10% de bonne reconnaissance pour les moments centrés et les moments de Zernike. Cela s'explique, comme vu précédemment, par l'invariance de ces deux dernières familles à certaines transformations géométriques (contrairement aux moments cartésiens) ; les individus sont mieux répartis dans l'espace des attributs, facilitant leur discrimination et améliorant la compensation des erreurs pour les régions "difficiles" de cet espace grâce à la combinaison de classifieurs. Le facteur d'affaiblissement étant

lié aux distances entre l'individu et les classes d'affectation lorsque le rapport des distances n'est pas suffisant, le système refuse de prendre une décision.

Avec intégration du Modèle de Croyance Transférable et sélection dynamique du meilleur classifieur

En complément de l'étage de prise de décision que nous venons de rajouter, nous ne travaillons plus cette fois avec un unique classifieur, mais avec un groupe de classifieurs, dont un seul est choisi dynamiquement (dynamic classifier selection ou DCS) en fonction des caractéristiques des attributs de l'individu en cours d'analyse. Les modalités de sélection dynamique du meilleur classifieur sont présentées dans le chapitre suivant. Les résultats obtenus avec ce système sont présentés en figures 3.19 et 3.20.

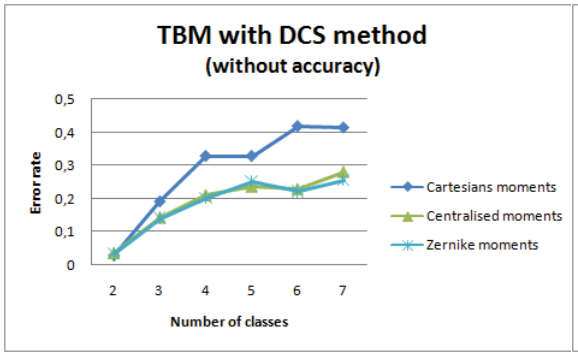

FIGURE 3.19 – Taux d'erreur pour DCS+MCT sans utilisation de la précision

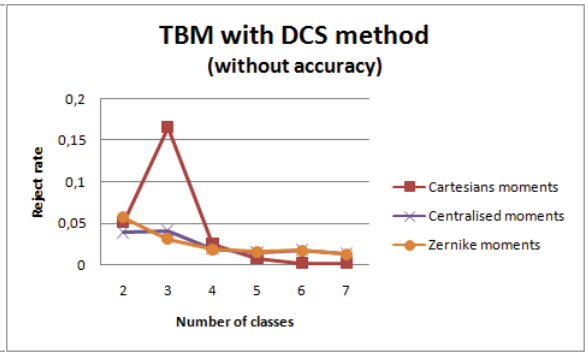

FIGURE 3.20 – Taux de rejet pour DCS+MCT sans utilisation de la précision

Ces deux figures montrent l'influence du choix du classifieur pour la prise de décision. Comme nous le verrons dans le chapitre 4, **le classifieur choisi dépend réellement de l'individu et n'est donc pas toujours le même sur la base de travail**.

Comparaison expérimentale de l'apport de la précision de la mesure dans la prise de décision

L'objectif de cette partie expérimentale est de montrer l'influence d'une mesure de précision, associée à chaque attribut établi pour un individu à classer ou étiqueter. Nous cherchons à mettre en évidence que dans un système simple, cette mesure de précision permet de ne pas prendre de décision et donc de réduire le taux d'erreur.

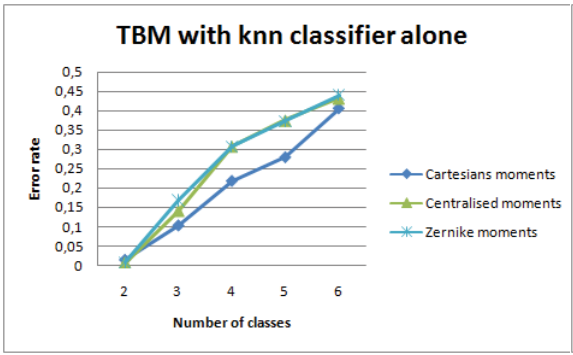

FIGURE 3.21 – Taux d'erreur pour KPPV+MCT

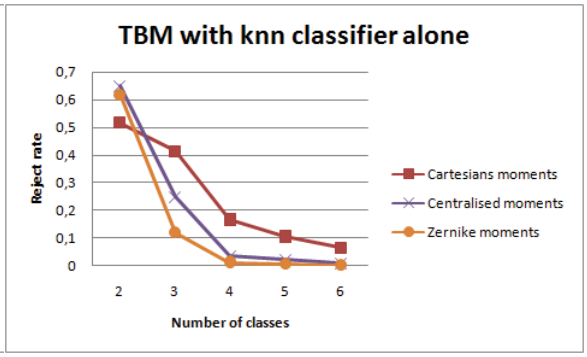

FIGURE 3.22 – Taux de rejet pour KPPV+MCT

Pour cette phase, nous allons associer à chaque valeur d'attribut une mesure de précision déterminée lors du calcul de l'attribut. Cette mesure sera intégrée dans la décision selon la formule 3.48.

Nous rappelons qu'à partir de $L(\omega_q|x) = C_i(\omega_q|A_j(x))$ alors :

$$\begin{cases} m_{ijq}(\{\omega_q\}) & = 0 \\ m_{ijq}(\overline{\omega_q}) & = \alpha_{jq}(1 - R.C_i(\omega_q|A_j(x))) \\ m_{ijq}(\Omega) & = 1 - m_{ijq}(\overline{\omega_q}) \end{cases}$$
$$\forall i \in [1,n_c], \forall j \in [1,n_a], \forall q \in [1,n_q]$$

avec $R \in]0, (\sup_x \max_q(L(\omega_q|x)))^{-1}]$.

Nous considérons le système de prise de décision suivant :

$$\begin{cases} m(\{\omega_k\}) & = \bigoplus_{q=1}^{n_q} m_{ijq}(\{\omega_k\}) \\ m(\overline{\omega_k}) & = \bigoplus_{q=1}^{n_q} m_{ijq}(\overline{\omega_k}) \end{cases} \quad \forall q \in [1,n], \forall i \in [1,n_c], \forall j \in [1,n_a]$$

et :

$$m(\Omega) = 1 - \sum_{i=1}^{n} (m(\{\omega_i\}) + m(\overline{\omega_i}))$$

La première conséquence est que nous créons un ensemble de rejet, contenant tous les individus pour lesquels la décision d'affectation n'a pas été prise. La figure 3.22 présente l'évolution du cardinal de cet ensemble en fonction du nombre de classes à discriminer, alors que la figure 3.21 propose le taux d'erreur obtenu.

Première constatation, les courbes de taux d'erreur n'ont pas ou peu été modifiées. Seules les valeurs obtenues pour les décisions à deux ou trois classes gagnent légèrement quelques points de pourcentage. Seconde constatation, il apparaît un taux de rejet de l'ordre de 6% pour un problème à deux classes et 1% pour un problème à une classe. Ainsi et tel qu'attendu, le gain obtenu en terme de taux d'erreur s'est effectué en décidant de rejeter les individus dont les attributs n'étaient pas assez précis.

Ce résultat est partiellement satisfaisant. Il montre bien que l'hypothèse effectuée est vérifiée : à partir de la confiance que nous avons sur la mesure d'un individu nous prenons une meilleure décision. En revanche, globalement les résultats ne sont pas nettement meilleurs et notamment lors de l'accroissement du nombre de classes à discriminer pour lesquelles le gain devient nul au delà de cinq classes.

Avec intégration du Modèle de Croyance Transférable et sélection dynamique du meilleur classifieur

Notons d'abord le réel apport de la sélection dynamique des classifieurs, avec un taux d'erreur d'environ 10% sur un problème à 3 classes pour les moments

centrés et les moments de Zernike, taux d'erreur qui semble se stabiliser autour des 20%, alors que ce dernier s'accroît avec le nombre de classes pour le classifieur *k* plus-proches voisins utilisé seul.

La sélection dynamique des classifieurs n'a pas la même influence sur les performances selon le type de moments utilisé : étant donné que celle ci se base sur la répartition des individus dans l'espace des attributs (calcul de la précision locale dans le voisinage de l'individu inconnu), on peut supposer que la sensibilité aux transformations géométriques influence la répartition des individus dans l'espace des attributs. Nous rappelons que les familles de moments utilisées ont les caractéristiques suivantes :

- Moments cartésiens : pas d'invariance
- Moments centrés : invariance aux changements d'échelle et aux homothéties
- Moments de Zernike : invariance aux changement d'échelle, aux homothéties et aux rotations

Notons également que théoriquement les translations et homothéties sont normalement gérées dans les prétraitements ayant permis la construction de la base MNIST.

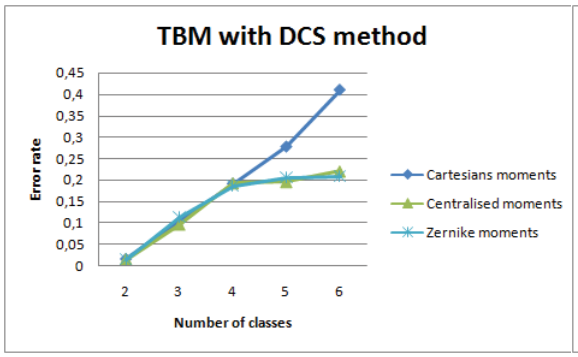

FIGURE 3.23 – Taux d'erreur pour DCS+MCT

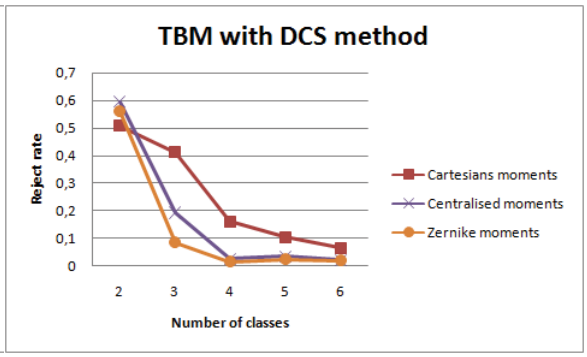

FIGURE 3.24 – Taux de rejet pour DCS+MCT

Utilisation de la précision dans une chaîne de décision bouclée

L'objectif de cette expérience est de montrer la réduction du rejet qui peut-être effectuée par un bouclage de la chaîne de traitements. Ce bouclage ayant pour conséquence d'enchaîner les traitements en utilisant un nouvel attribut à chaque itération. Nous cherchons donc à mettre en évidence la réduction du volume d'individus rejetés.

Avec intégration du Modèle de Croyance Transférable

Cette partie illustre les résultats obtenus avec notre chaîne de traitements complète, exécutée avec bouclages. Pour ces tests, nous simulons un choix de l'utilisateur favorisant le temps de calcul au détriment de la précision. Pour cela, notre chaîne exécute une première passe avec les moments cartésiens (l'attribut le moins coûteux en temps de calcul), puis réinjecte les individus rejetés pour une deuxième passe avec cette fois les moments centrés, puis effectue une dernière passe avec les individus rejetés par la seconde en utilisant les moments de Zernike comme attributs. Les résultats obtenus sont représentés en figure 3.25 pour le taux d'erreur, et 3.26 pour le taux de rejet.

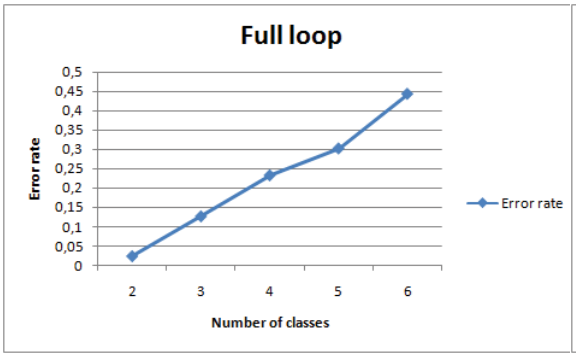

FIGURE 3.25 – Taux d'erreur pour une boucle à 3 étapes

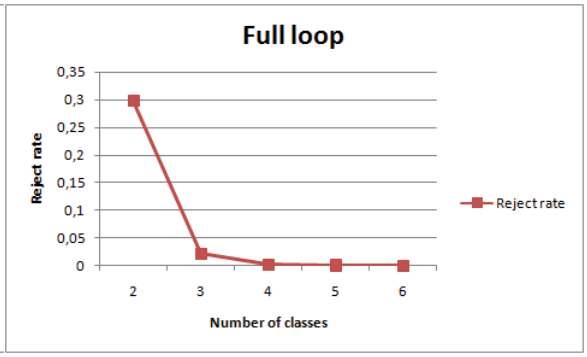

FIGURE 3.26 – Taux de rejet pour une boucle à 3 étapes

Comme nous l'avons expliqué plus haut, nous ne nous attendons pas à ce que le taux d'erreur global soit supérieur à celui obtenu avec la méthode choisie pour la première passe de la chaîne, les individus mal classés n'étant pas rejetés, ils ne sont pas analysés avec une autre méthode. C'est ce que nous pouvons constater en comparant avec la figure 3.23 ; le taux d'erreur obtenu avec les moments cartésiens est la limite vers laquelle tend le taux d'erreur obtenu avec notre chaîne complète, effectuant une première passe avec ces méthodes du fait de la recherche du temps de calcul minimum.

Ainsi, étant donné que la figure 3.23 montre des résultats nettement meilleurs, en terme de bonne reconnaissance, pour les moments centrés et les moments de Zernike, le système sera plus performant si lors de l'exécution nous privilégions la précision au détriment du temps de calcul.

D'autre part, et comme nous l'attendions, l'augmentation du nombre de classes implique effectivement une répartition équitable de la croyance en toutes les classes, provoquant une baisse du nombre de rejets. Notons que le système tend à ne plus rejeter aucun individu au delà de problèmes à quatre classes. Notre proposition d'extension du second modèle d'Appriou, consiste à calculer la masse de croyance d'une hypothèse comme $m_{ijq}(\overline{\omega_q}) = \alpha_{jq}(1 - R.C_i(\omega_q|A_j(x)))$. La précision ($\alpha_{jq}$) étant toujours la même qu'importe le nombre de classes, c'est l'information utile qu'elle porte qui se retrouve diluée en même temps qu'augmente la complexité du problème. D'ailleurs, la figure 3.11 illustre bien que pour une valeur de α_{jq} trop faible par rapport au nombre de classes, la possibilité d'un rejet est quasi inexistante.

Enfin, du fait que l'annulation des rejets coïncide avec la stabilisation des erreurs, nous pouvons supposer qu'à partir de problèmes à quatre classes, les erreurs entre les méthodes ne se compensent plus et que le système atteint un point d'équilibre.

Fusion

A ajouter à la liste des résultats montrés ci-dessus, nous avons expérimenté une version sans bouclage de la chaîne. Pour cela, chaque couple attributs-classifieur est exécuté, et l'ensemble des sorties des classifieurs est combiné en fin de chaîne de la manière suivante :

$$\begin{cases} m(\{\omega_k\}) & = \bigoplus_{q=1}^{n_q} \bigoplus_{i=1}^{n_c} \bigoplus_{j=1}^{n_a} m_{ijq}(\{\omega_k\}) \\ m(\overline{\omega_k}) & = \bigoplus_{q=1}^{n_q} \bigoplus_{i=1}^{n_c} \bigoplus_{j=1}^{n_a} m_{ijq}(\overline{\omega_k}) \end{cases} \tag{3.53}$$

Le taux d'erreur de cette chaîne non bouclée, mettant en coopération toutes les méthodes existantes dans le système est représenté par la figure 3.27. Le taux de rejet obtenu est visible dans la figure 3.28.

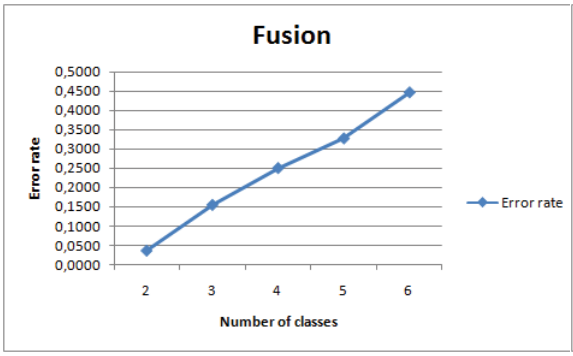

FIGURE 3.27 – Taux d'erreur pour un système coopératif

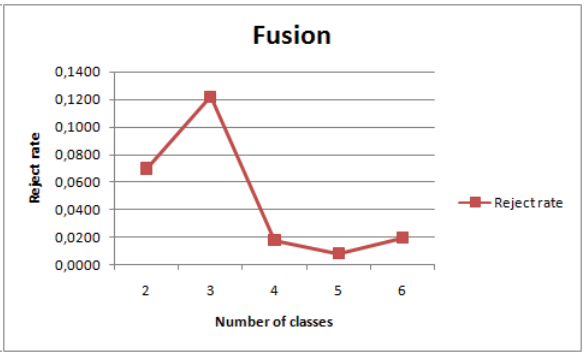

FIGURE 3.28 – Taux de rejet pour un système coopératif

Notre premier constat est que la mise en coopération de tous les opérateurs n'apporte pas forcément de meilleurs résultats que ceux obtenus avec un système de chaîne bouclée. Nous pourrions supposer qu'une mise en coopération de toutes les méthodes permettrait, à l'instar de la sélection dynamique des classifieurs, que les erreurs finissent par se compenser entre les méthodes. Or, n'oublions pas que l'étape finale de notre système consiste à combiner, fusionner, l'ensemble

des résultats obtenus. Comme notre étage de décision est basé sur le modèle des croyances transférables, le système choisira l'hypothèse qui possédera la masse de croyance la plus élevée. Mais dans ce cas, et contrairement à l'intérêt d'une entropie forte pour un système multi-classifieur, l'intégralité des croyances tend à être uniformément répartie sur toutes les hypothèses, les erreurs des différentes méthodes testées ne pourront donc pas se compenser.

3.5.4 Conclusion du chapitre

La première constatation que nous pouvons faire en synthèse de ce chapitre est que la sélection dynamique du classifieur (DCS), telle que nous la détaillons au chapitre 4, apporte un gain notable, en terme de taux de reconnaissance globale, par rapport à l'utilisation d'un seul classifieur. Nous montrons par là un aspect intéressant de la construction dynamique d'une chaîne d'opérateurs, par rapport à l'utilisation de chaînes statiques fixées *a priori*.

De plus, le modèle des croyances transférables (MCT) apporte la robustesse nécessaire à un système industriel, grâce notamment à la gestion des rejets et à la fusion des informations, ce qui manque aux classifieurs utilisés seuls.

Le formalisme du MCT, et de manière plus générale celui de la théorie de l'évidence, permet d'intégrer la précision d'un opérateur de calcul d'attributs, ceci répondant à l'une des contraintes posées par le projet COROC.

Malheureusement, la prise en compte de la précision, telle qu'elle l'a été proposée dans ce chapitre, ne joue bien son rôle que pour des problèmes "simples" (jusqu'à quatre classes). Cependant, l'information qu'elle porte par la précision finit par se diluer dans l'information totale pour des problèmes plus complexes. L'approche que nous proposons doit encore s'améliorer, avec par exemple une

réécriture du second modèle d'Appriou, intégrant la précision de façon autre que sous forme de simple facteur.

Enfin, et contrairement au principe de diversité pour les classifieurs (principe détaillé au chapitre 4), nous n'avons pas ici de système de prise finale de décision qui permette aux méthodes testées de compenser les erreurs l'une de l'autre, car les méthodes ne sont pas en coopération ; elles sont simplement enchaînées de façon séquentielle. Les outils de fusion d'information de la théorie de l'évidence devraient donc aider à améliorer notre approche.

Par rapport au schéma général d'une chaîne de traitements, comprenant un étage de calcul d'attributs, un étage de classification et un étage final de décision, c'est ce dernier que nous venons d'aborder, en montrant comment il est possible de choisir dynamiquement les opérateurs, de décider de tenter d'autres enchaînements que le premier utilisé, et de boucler sur cette chaîne. Ces opérations se décident en fonction de la précision des opérateurs. Nous détaillons dans les chapitres suivants comment calculer la précision des opérateurs de calcul des attributs et des classifieurs.

"*Les idées précises conduisent souvent à ne rien faire.*"
Paul Valéry

Chapitre 4

Classifieur et mesure de précision

Sommaire

4.1 Introduction

En introduction de ce manuscrit, nous avons exposé notre point de vue sur le fait que nous considérons toute application de traitement d'images comme une chaîne de traitements. Et dans ce contexte, nous avons choisi une application de reconnaissance de caractères comme support applicatif. La chaîne type que nous considérons est constituée de trois étages, un premier concernant l'extraction des attributs, un second pour associer à l'objet, vu au travers des attributs, une étiquette par le biais d'un classifieur et enfin l'étage final de décision par combinaison de plusieurs étiquettes (figure 4.1).

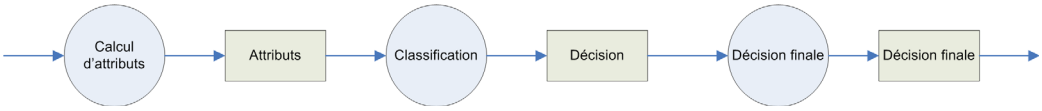

FIGURE 4.1 – Chaîne de traitement linéaire simple

Au cœur de ce point de vue, nous émettons comme hypothèse que si un opérateur est très bien adapté pour une partie des données à traiter, il peut ne pas être optimal pour une petite partie des données. En conséquence, nous considérons que pour chaque étage, plusieurs opérateurs en parallèle traitent les données pour que les résultats puissent ensuite être combinés (figure 4.2). Ce point de vue est notamment étayé par la notion de précision associée à une mesure de décision.

Nous admettons également qu'une méthode, un opérateur, ne sera pas optimal pour l'intégralité des individus qu'il traitera. En métrologie par exemple, nous savons que la précision d'un capteur n'est pas constante sur la bande des valeurs qu'il peut accepter, comme l'illustre la figure 4.3 représentant l'évolution de la précision d'un capteur d'humidité et de température utilisé en météorologie. Nous voyons très nettement que la précision des mesures est plus faible sur les extrêmes.

Nous avons vu au chapitre précédent notre approche concernant la prise de décision finale, intégrant les données en sortie de l'étage de classification ainsi

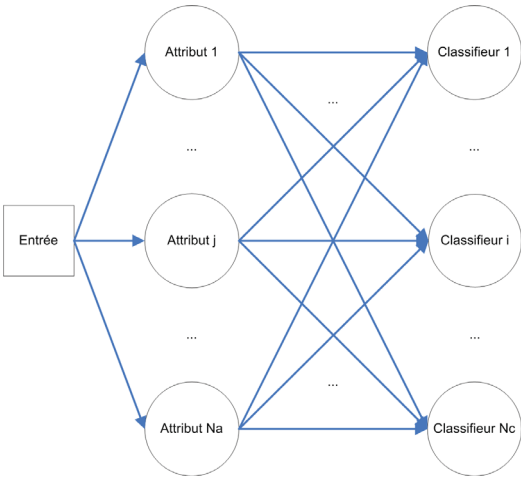

FIGURE 4.2 – Combinaison linéaire simple des opérateurs

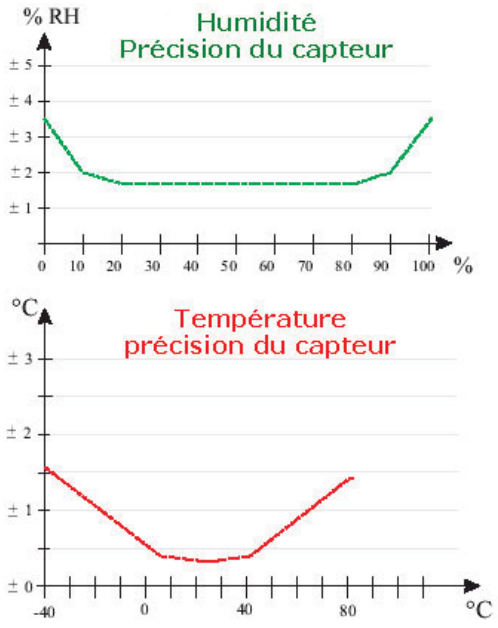

FIGURE 4.3 – Précision d'un capteur météorologique

que la précision mesurée en sortie des opérateurs de calculs d'attributs, comme illustré par la figure 4.4.

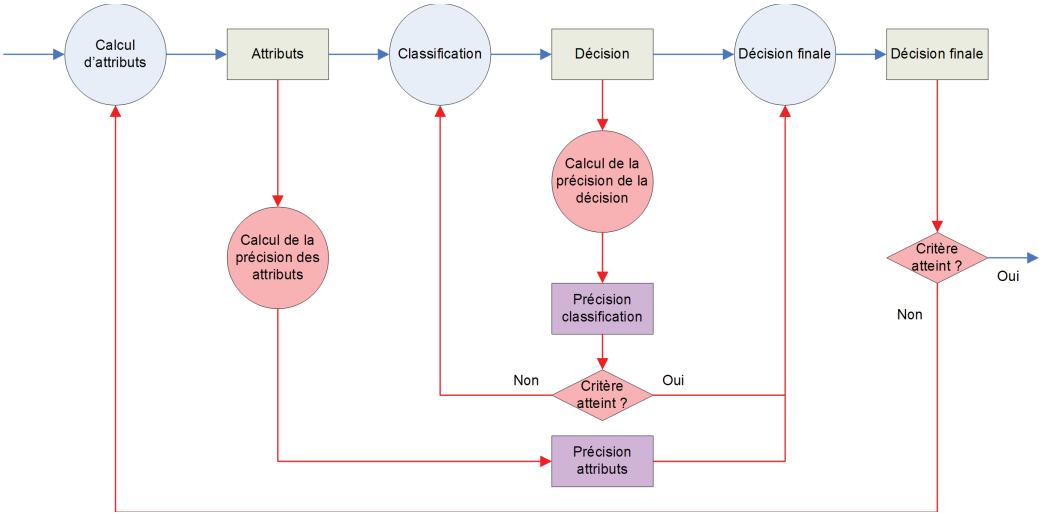

FIGURE 4.4 – Chaîne de traitement intégrant la précision des attributs et de la classification

Intéressons-nous à présent à la classification par elle même, et principalement aux questions que cette partie soulève quant à la mesure de précision. Qu'y a t'il de mesurable en sortie d'un classifieur ? Est-ce sa précision ? Si oui, pouvons-nous à partir de cette mesure, sélectionner dynamiquement un classifieur ?

4.2 Les classifieurs

Le Trésor de la Langue Française définit la classification comme la "répartition systématique en classes, en catégories, d'êtres, de choses ou de notions ayant des caractères communs notamment afin d'en faciliter l'étude". Nous définirons donc un **classifieur** comme un outil automatique, un opérateur, réalisant une classification.

D'un point de vue mathématique, un classifieur est une application d'un espace d'attributs X (discret ou continu) vers un ensemble d'étiquettes Y.

Les classifieurs peuvent-être fixes ou apprenants, et ceux-ci peuvent à leur tour être divisés en classifieurs apprenants supervisés ou non supervisés.

Les applications sont multiples. On en retrouve en médecine (analyse des tests pharmacologiques, analyse de données d'IRM), en finance (prédiction de cours), en téléphonie mobile (décodage de signal, correction d'erreurs), en vision artificielle (reconnaissance de visages, suivi de cible), en reconnaissance de la parole, en datamining (analyse des achats en supermarché) et encore dans bien d'autres domaines.

Nous pouvons citer en exemple un classifieur qui prend en entrée les salaires d'une personne, son âge, son statut marital, son adresse et ses relevés bancaires, et qui classifie une personne comme recevable/non-recevable à l'obtention d'un crédit.

Un classifieur se base sur une connaissance *a priori* (fixée par l'expert ou construite par apprentissage) décrivant les différentes classes dans l'espace des attributs. A partir de cette connaissance, le classifieur va attribuer, aux individus inconnus qui lui seront présentés, le label de la classe à laquelle l'individu à la plus forte probabilité d'appartenir.

Il faut savoir qu'un classifieur seul ne peut assurer des performances idéales pour tous les cas possibles (modèles de classes se recouvrant, individu aux frontières entre plusieurs classes etc.). C'est pour cela qu'il est intéressant de combiner plusieurs classifieurs afin qu'ils compensent leurs erreurs entre eux.

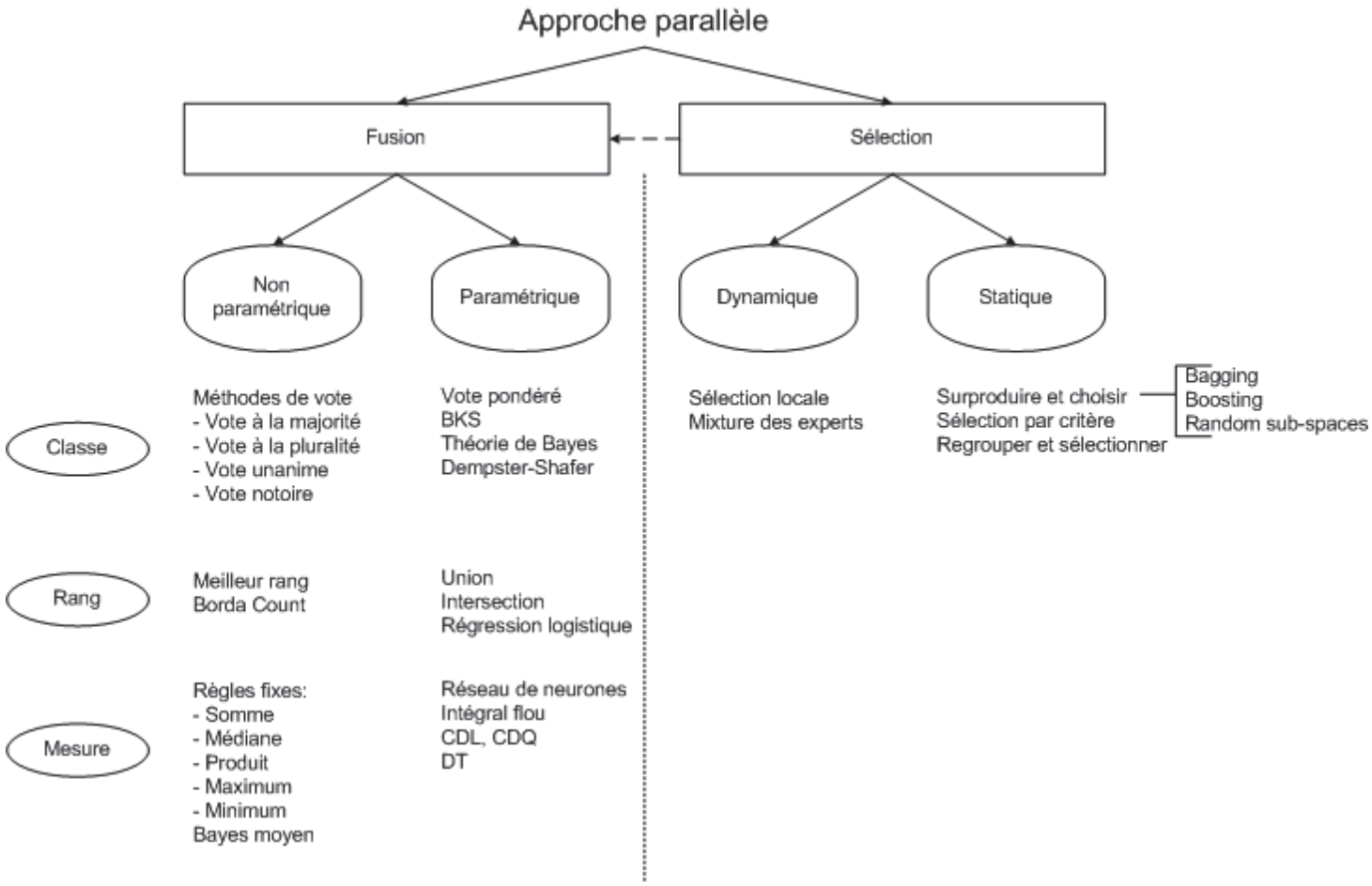

FIGURE 4.5 – Taxonomie des méthodes de combinaison parallèle de classifieurs[94]

Zouari propose une taxonomie [1] (voir figure 4.5) des différentes méthodes de combinaison dans [94].

Elle rapporte notamment les travaux de Parker[58] qui montrent que les méthodes de type **rang** peuvent être plus performantes que les méthodes de type classe et mesure. Elle explique aussi que les méthodes non-paramétriques sont les plus utilisées par les chercheurs car simples à mettre en œuvre et n'utilisant pas de traitements supplémentaires (apprentissage). Nous ajouterons à cela qu'elles sont souvent les plus rapidement codées et surtout indépendantes des structures de classes.

Le but de la classification est d'établir des relations d'ordre entre les individus, afin d'obtenir une information la plus concise possible, atteindre le niveau d'abstraction le plus important (i.e. décrire, de manière précise, un document de 3Mo en 100ko). Il existe un espace qui décrive les données connues mais comment le choisir ?

Un classifieur peut être vu d'une part comme une mesure de distance, et d'autre part comme une méthode d'organisation spatiale (avec en fond une méthode d'optimisation). Ces deux parties combinées font de l'opérateur de classification un opérateur de pavage de l'espace de représentation des données qui lui sont proposées. Ceci entraîne des questions telles que, compte tenu d'une distance, quelle est la meilleure organisation ?

Le rôle du classifieur est double :

- Analyser les données sans apprentissage (sans supervision) : créer des groupes, mettre en place une organisation ;
- Positionner les individus par rapport à des groupes avec apprentissage (avec

1. Classification d'éléments ; suite d'éléments formant des listes qui concernent un domaine, une science

supervision) : information locale autour de l'individu.

De façon plus générale, nous voulons créer des groupes répondant à des attentes d'usage ou d'interprétation, les créer artificiellement selon des caractères communs, le choix des critères contraignant le choix de l'espace. Mais comment choisit-on ces critères ?

Les techniques de classification automatique produisent des groupements d'objets (ou individus) décrits par un certain nombre de variables ou de caractères (les attributs). Le recours à de telles techniques est souvent lié à des suppositions sinon à des exigences sur les regroupements des individus. Il existe plusieurs familles d'algorithmes de classification, comme nous pouvons le trouver dans [49] :

- Les algorithmes de partitionnement
- Les algorithmes ascendants (agglomératifs)
- Les algorithmes descendants (divisifs)

Ces méthodes présentent chacune des avantages et des inconvénients différents qui sont à l'origine de leur sélection en fonction du contexte applicatif. Bien évidement, ces aspects justifient également l'utilisation combinée de classifieurs. Toute une communauté organisée autour du monde de l'apprentissage s'est ainsi constituée pour combiner plusieurs classifieurs dans une même application. Cependant, dans notre contexte, ces combinaisons sont définies comme statiques, car établies *a priori* de la décision finale, à l'issue de l'apprentissage.

Sans reprendre l'argumentaire induisant qu'un même opérateur peut avoir du mal à être adapté à l'ensemble des cas de figure posés par l'application, nous aboutissons ici également au même propos concernant les combinaisons statiques de classifieurs. Il est bien entendu que ceci n'est pas une vérité absolue, la qualité des résultats obtenus par ces techniques de combinaison statique de classifieurs pour certaines applications n'étant pas à remettre en cause. Cependant, dans le cadre

d'applications où la variabilité des données est très grande, l'assemblage statique peut avoir du mal à atteindre un haut niveau de résultat.

Dans l'hypothèse de la combinaison dynamique de classifieurs, nous savons que se posent à la fois les questions de sélection dynamique du meilleur classifieur et dans notre réflexion celles de l'estimation d'une mesure de précision à associer au résultat de la classification.

Pour démarrer l'étude de ces questions, nous nous appuierons sur les travaux de L. Kuncheva[47][45][46] qui a étudié et développé un ensemble d'outils de mesure de la qualité d'un classifieur par rapport à d'autres classifieurs, puis présente une mesure (la diversité) de l'apport d'un classifieur dans un groupe de classifieurs.

4.2.1 Mesure de la qualité dans la bibliographie

Qualité

L'estimation de la qualité des résultats issus d'un classifieur s'appuie bien souvent sur l'analyse *a posteriori* de la propension des classifieurs à se tromper. Lors d'analyse comparatives de classifieurs entre eux, la qualité est estimée en fonction du taux de bonne classification. Ces comparaisons peuvent se faire sur les mêmes critères que pour un seul classifieur (bonnes/mauvaises reconnaissances etc.), mais également utiliser des outils, inspirés des statistiques, de mesure de la qualité d'un modèle, d'une loi de probabilité etc.

Dans la recherche de ce qui pourrait constituer une estimation de la précision de décision ou des éléments entrant dans la combinaison dynamique des classifieurs, nous avons analysé différents critères associés à la qualité de la décision. Nous proposons ci-après quelques uns de ces critères :

1. Le **calcul de l'erreur** d'un classifieur, qui correspond au nombre d'individus mal rangés sur le nombre d'individus total. Exprimé sous forme de probabilité, il devient possible de sélectionner les classifieurs selon leur probabilité de faire une erreur dans un intervalle de confiance.

2. La comparaison entre 2 classifieurs : si 2 classifieurs donnent des taux d'erreurs différents sur une même base de test, sont ils vraiment différents ?

   – Soient 2 classifieurs, et leurs performances, N_{11} nombre d'individus correctement classés par les 2, N_{01} nombre d'individus mal classés par le premier mais bien classés par le second, N_{10} l'inverse et N_{00} le nombre d'individus mal classés par les 2. Il existe une variable statistique suivant une loi de Chi-deux, mesurant la qualité d'un classifieur par rapport à un autre, s'écrivant :

   $$x^2 = \frac{(|N_{01} - N_{10}| - 1)^2}{N_{01} - N_{10}} \tag{4.1}$$

   Sachant x^2 à peu près distribué selon χ^2 à 1 degré de liberté, on peut dire, avec un niveau de certitude de 0.05, que si x^2>3.841, alors les classifieurs présentent des performances fortement différentes. Notons qu'il existe aussi un test appelé "difference of two proportions" mais Dietterich prouve dans [22] que cette mesure est trop sensible à la violation de la règle d'indépendance des données (problème du jeu de données pour l'apprentissage et la reconnaissance restreint) et recommande d'utiliser la mesure ci-dessus.

3. Plusieurs auteurs ont étendu cette analyse comparative au cas où plusieurs classifieurs sont combinés à partir de la même base d'apprentissage [86] :
   – Le **Q Test de Cochran**[28][48] vérifie l'hypothèse : "tous les classifieurs

présentent les même performances". Si l'hypothèse est vérifiée alors

$$Q_c = (N_c - 1)\frac{N_c \sum_{i=1}^{N_c} G_i^2 - T^2}{N_c T - \sum_{j=1}^{N_x} (N_{cj})^2} \tag{4.2}$$

La variable Q_c suit un χ^2 à $N_c - 1$ degrés de liberté, avec G_i le nombre d'éléments de $\mathcal{L}$ (espace d'apprentissage) correctement classés par le classifieur $C_i (i = 1, ..., N_c)$, N_x étant le nombre total d'individus appris. N_{cj} est le nombre total de classifieurs de $\mathcal{C}$ qui ont correctement classé l'objet $x_j \in \mathcal{L}$ et T est le nombre total de bonnes décisions prises par l'ensemble des classifieurs.

Ainsi, pour un niveau de confiance donné, si Q_c est supérieur à la valeur attendue du χ^2 alors il existe des différences significatives entre les classifieurs justifiant leur combinaison.

– Le même principe peut être adopté en adoptant une loi de Fisher-Snedecor ayant $(N_c - 1)$ et $(N_c - 1) * (N_x - 1)$ degrés de liberté. C'est le **F-Test**[54]. En partant des performances des classifieurs estimées au cours de l'apprentissage $(\bar{p_1}, ..., \bar{p_{N_c}})$ et la performance moyenne globale $\bar{p}$ on obtient la somme des carrés pour les classifieurs :

$$SSA = N_x \sum_{i=1}^{N_c} \bar{p_i}^2 - N_x N_c \bar{p}^2 \tag{4.3}$$

Puis la somme des carrés pour les objets :

$$SSB = \frac{1}{N_c} \sum_{j=1}^{N_x} (N_{cj})^2 - N_c N_x \bar{p}^2 \tag{4.4}$$

La somme totale des carrés :

$$SST = N_x L \bar{p}(1 - \bar{p}) \tag{4.5}$$

Enfin, la somme totale des carrés pour l'interaction classification/objet :

$$SSAB = SST - SSA - SSB \tag{4.6}$$

Dés lors, le critère F est estimé comme le rapport entre le MSA et le MSAB définis comme suit :

$$MSA = \frac{SSA}{(N_c-1)}; \quad MSAB = \frac{SSAB}{(N_c-1)(N_x-1)}; \quad F = \frac{MSA}{MSAB} \tag{4.7}$$

– Il est également possible d'appliquer une **validation croisée**[22]. Il s'agit de répéter un certain nombre de fois (K) l'apprentissage/reconnaissance, en séparant à chaque fois le jeu de données à apprendre en 2 sous-jeux (habituellement 2/3 des données pour l'entraînement et 1/3 pour le test). Deux classifieurs C_1 et C_2 sont entraînés sur le jeu d'apprentissage et testés sur le jeu de test. A chaque tour, les précisions des deux classifieurs sont mesurées : P_{C_1} et P_{C_2}. Nous obtenons ainsi un ensemble de différences, de $P^{(1)} = P^{(1)}_{C_1} - P^{(1)}_{C_2}$ à $P^{(2)} = P^{(2)}_{C_1} - P^{(2)}_{C_2}$. En posant $\overline{P} = (1/K)\sum_{i=1}^{K} P^{(i)}$, nous mesurons :

$$t = \frac{\bar{P}\sqrt{K}}{\sqrt{\sum_{i=1}^{K}(P^{(i)} - \bar{P})^2/(K-1)}} \tag{4.8}$$

Si t suit bien une loi de Student à $K-1$ degrés de liberté (et pour le niveau de confiance choisi) alors les deux classifieurs présentent le même comportement.

Rappelons que l'optique de ce travail est d'aboutir à un système de sélection dynamique d'opérateur de classification, basé sur une mesure de qualité. Les méthodes présentées ci-dessus présentent un certain nombre d'inconvénients pour répondre à notre besoin :

– En ce qui concerne la mesure de l'erreur d'un classifieur, cette information est non uniforme sur l'espace des attributs. De plus, nous pouvons supposer que cette information variera si nous augmentons la taille de l'espace d'apprentissage dans le temps.

– Les méthodes de comparaison de classifieurs demandent un grand nombre de sessions d'entraînement et de test. Au delà des comparaisons, une fois la mesure établie, comment décider sur le classifieur à utiliser ? Dans le cas de la validation croisée, si nous réalisons 10 essais, nous construisons 10 classifieurs différents, construits sur 10 sous-ensembles différents. Ces méthodes ont pour seul but de donner une estimation de la précision d'un certain modèle construit uniquement sur le problème présent. Ainsi, Dietterich, dans [22], pose l'hypothèse selon laquelle la précision de la classification varie en fonction de la taille du jeu d'apprentissage.

Kuncheva propose pour cela une approche considérant un ensemble de classifieurs élémentaires comme un seul classifieur, construit et testé sur les mêmes jeux d'apprentissage et de test.

4.2.2 La diversité : une solution ?

Définition de la diversité :

Il faut admettre que chaque classifieur commet des erreurs, car sinon quelle serait l'utilité de vouloir les combiner ? A partir de là, nous cherchons donc à compenser les erreurs d'un classifieur en le combinant avec un autre, qui produira des erreurs mais sur des objets différents. Ainsi, la diversité des sorties d'un classifieur devient un élément crucial pour réussir la combinaison. De façon intuitive, nous pouvons dire que, cherchant à obtenir une sortie de la meilleure qualité possible, s'il y a des erreurs, il vaut mieux que ces erreurs tombent sur des objets différents d'un classifieur à l'autre. Kuncheva explique qu'en pratique il est très difficile de définir une seule mesure de diversité, et encore plus compliqué de lier cette mesure à la performance de l'ensemble en une et simple expression de dépendance. Il existe des mesures de diversité propres à certaines branches de la science (biologie, conception de logiciels etc.), nous nous pencherons sur celles

qui concernent les ensembles de classifieurs. La diversité peut être considérée de 3 points de vue différents :

- Comme caractéristique de l'ensemble des classifieurs : Nous avons un ensemble de classifieurs, mais nous ne savons pas quelle méthode de combinaison utiliser et nous ne savons même pas si les décisions données sont correctes. Dans ce cas, la diversité apporte des informations supplémentaires sur les taux d'erreurs individuels et sur l'erreur globale. La diversité permet de découvrir si un classifieur contribue ou non au succès de l'ensemble

- Comme caractéristique de l'ensemble des classifieurs et de leur combinaison : Dans ce cas, nous connaissons l'ensemble des sorties. Ainsi nous pouvons dire quel classifieur dévie le plus et celui qui dévie le moins par rapport au reste de l'ensemble et mesurer la diversité sur une base plus individuelle. Différentes combinaisons amèneront à différentes diversités pour le même ensemble de classifieurs

- Comme caractéristique de l'ensemble des classifieurs, de leur combinaison et des erreurs : nous pouvons enfin utiliser l'information connue *a priori* de la classe d'appartenance des individus du jeu d'apprentissage (cette information est aussi appelée *oracle*). La diversité est considérée comme un composant de l'ensemble des erreurs. C'est ce sur quoi portent actuellement les travaux de Kuncheva[46] : découvrir une relation entre la diversité et l'erreur de l'ensemble afin de construire de meilleures combinaisons

Avant de découvrir diverses mesures de la diversité, rappelons les notations suivantes :
$\mathcal{T}$: Jeu de données, $\mathcal{T} = \{x_1^*, x_2^*, ..., x_{N_{x^*}}^*\}$
$C_i(x)$: décision du classifieur i pour l'individu x

N_c : Nombre de classifieurs

γ_i : précision ou probabilité pour le classifieur i de donner la bonne réponse

C : espace des classifieurs

N_{x^*} : nombre d'individus inconnus

Mesure de diversité par paire de classifieurs :

- **Mesure de désaccord** : probabilité que deux classifieurs ne soient pas d'accord sur leurs décisions
- **Mesure de la double faute** : probabilité que les classifieurs se trompent tous les deux

Autres mesures de diversité :

- **Mesure de l'entropie** : considérant que l'ensemble est plus diversifié si pour un individu particulier, la moitié des votes est 0 et l'autre est 1, et de la même manière que la diversité est nulle si tous les classifieurs sont d'accord, une mesure possible de ce concept est :

$$E = \frac{1}{N_{x^*}} \frac{2}{N_c - 1} \sum_{j=1}^{N_{x^*}} \min[(\sum_{i=1}^{N_c} C_i(x_j^*)), (N_c - \sum_{i=1}^{N_c} C_i(x_j^*))] \tag{4.9}$$

  E variant de 0 à 1 selon si tous les classifieurs donnent la même réponse, ou si la diversité est maximale.
- Variance de **Kohavi-Wolpert** : si le même classifieur est entraîné plusieurs fois (avec des jeux de données différents) et que la variance de sa décision est étudiée pour un même individu, $Y(x_j^*)$ étant le nombre de votes corrects pour l'individu x_j^* alors :

$$KW = \frac{1}{N_{x^*} N_c^2} \sum_{j=1} N_{x^*} Y(x_j^*)(N_c - Y(x_j^*)) \tag{4.10}$$

  Notons que la mesure KW est proportionnelle à la moyenne des mesures de

désaccord entre les classifieurs $\overline{D}$:

$$KW = \frac{N_c - 1}{2N_c}\overline{D} \tag{4.11}$$

– Mesure du **taux d'accord** (Interrater Agreement) : connaissant la précision moyenne des classifieurs

$$\kappa = 1 - \frac{\frac{1}{N_c}\sum_{j=1}^{N_{x^*}} Y(x_j^*)(N_c - Y(x_j^*))}{N_{x^*}(N_c - 1)\overline{\gamma}(1-\overline{\gamma})} \tag{4.12}$$

Notons la relation avec KW :

$$\kappa = 1 - \frac{N_c}{(N_c-1)\overline{\gamma}(1-\overline{\gamma})}; \quad KW = 1 - \frac{1}{2\overline{\gamma}(1-\overline{\gamma})}\overline{D} \tag{4.13}$$

– Mesure de la **difficulté** : le principe est de définir une variable aléatoire discrète X ayant valeur dans $0/N_c, 1/N_c, ..., 1$ et de calculer la proportion de classifieurs de $\mathcal{C}$ qui classifient correctement une entrée x^* tirée aléatoirement de la distribution du problème. Les N_c classifieurs de $\mathcal{C}$ sont exécutés sur $\mathcal{T}$ afin d'estimer la fonction de probabilité de X. L'analyse de la distribution de X indiquera les individus qui posent le plus de difficultés à être reconnus. De plus, l'aspect général de cette distribution sera un indicateur de l'indépendance des classifieurs.

– **Diversité généralisée** : soit Y, une variable aléatoire exprimant la proportion des classifieurs qui "se trompent" sur un individu x^* tiré aléatoirement $(x \in \mathcal{T})$. Soit p_i, la probabilité que $Y = i/N_c$, soit la probabilité qu'il y ait exactement i sur les N_c classifieurs qui échouent sur un individu tiré au hasard. Soit $p(i)$ la probabilité que i classifieurs tirés au hasard échouent sur un x^* choisi aléatoirement. Supposons que nous choisissions 2 classifieurs au hasard dans $\mathcal{T}$. La diversité maximale est obtenue lorsqu'un des deux classifieurs échoue tandis que l'autre donne une bonne réponse. Dans ce cas, la probabilité que les deux classifieurs échouent est $p(2) = 0$. La diversité minimale est obtenue lorsque l'échec d'un classifieur est systématiquement accompagné par l'échec du deuxième classifieur. Ainsi, la probabilité que les deux classifieurs se trompent est la même que la probabilité

qu'un classifieur tiré aléatoirement échoue (soit $p(1)$). Ainsi, sachant que $p(1) = \sum_{i=1}^{N_c} \frac{i}{N_c} p_i$ et $p(2) = \sum_{i=1}^{N_c} \frac{i(i-1)}{N_c(N_c-1)}$ on obtient :

$$GD = 1 - \frac{p(2)}{p(1)} \tag{4.14}$$

– Erreur commune ou **Coincident Failure diversity** : cette mesure est simplement une modification de la diversité généralisée *GD*, afin d'obtenir une valeur minimale de 0 lorsque tous les classifieurs ont toujours raison, ou lorsqu'ils donnent tous simultanément soit une réponse juste, soit une réponse fausse :

$$CFD = \begin{cases} 0, & p_0 = 1.0; \\ \frac{1}{1-p_0} \sum_{i=1}^{N_c} \frac{N_c - i}{N_c - 1} p_i, & p_0 < 1. \end{cases} \tag{4.15}$$

Ces mesures indiquent la diversité d'un groupe de classifieurs construit *a priori*, en se basant sur les apports (ou pénalités) des classifieurs entre eux. Or, nous cherchons à établir une mesure de la qualité ou de la précision en sortie de l'étape de classification. Kuncheva[46] établit une liste de relations entre la diversité et la qualité de la classification.

Relation entre la diversité et la performance de la classification

– *"Plus la diversité est importante, plus l'erreur est faible"*[46]

– Chaque classifieur peut-être représenté graphiquement par un point dans un espace bidimensionnel généré par "Sammon mapping"[66], la distance entre chaque point étant la diversité par paire. Il est possible d'y ajouter chaque ensemble de classifieur comme un classifieur à part entière et même l'oracle lui-même. En considérant que chaque point est une paire de classifieur, κ peut jouer le rôle d'abscisse, l'erreur moyenne des deux classifieurs pouvant dans ce cas être l'ordonnée. Ceci nous donne ainsi $N_c(N_c-1)/2$ points, situant les meilleures paires dans la partie inférieur gauche (là où

l'erreur est minimale et la diversité maximale) (voir figure 4.6)

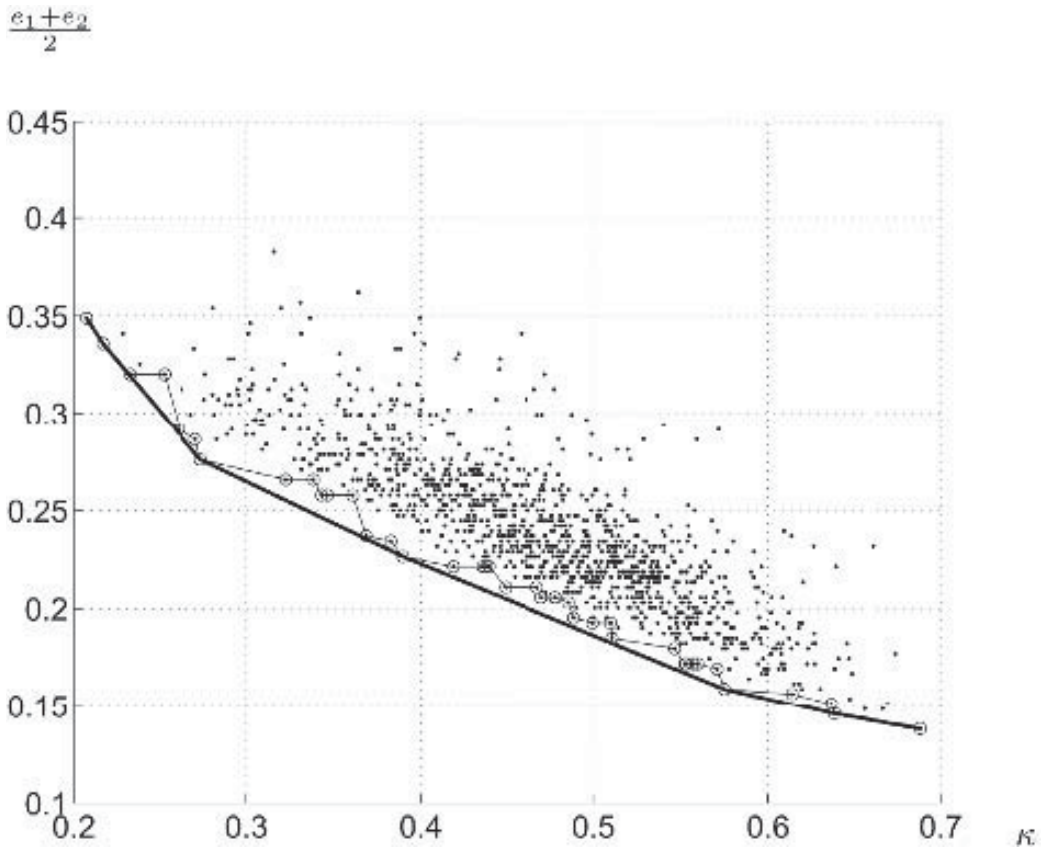

FIGURE 4.6 – Kappa-error de paires de classifieurs et son enveloppe convexe (trait gras)[46]

- Il est également possible de produire un grand ensemble de classifieurs pour n'en sélectionner que les plus diversifiés et les plus performants :
  - La matrice de diversité portant l'ensemble des valeurs du critère de la double faute est créée, et un nombre donné de classifieurs est choisi parmi les moins représentés. Une sorte de "pruning" analogue peut être réalisé sur la valeur de κ après utilisation de l'algorithme AdaBoost

  - Il est possible, en considérant la matrice des doubles fautes comme matrice de distance, de créer des clusters de classifieurs. Ainsi groupés, les

membres de chaque ensemble auront tendance à faire les mêmes erreurs alors que deux classifieurs choisis dans 2 groupes différents feront des erreurs sur des objets différents. A chaque étape de l'algorithme, un classifieur est sélectionné dans chaque cluster (i.e. celui possédant le plus fort taux de reconnaissance). Ainsi, au début, chaque classifieur est un groupe de lui-même (il y a donc N_c clusters), et ainsi l'ensemble est composé de tous les classifieurs. Ensuite, les deux classifieurs les moins diversifiés sont joints dans un cluster (l'ensemble se compose donc de $N_c - 1$ membres). De ce nouveau cluster, le classifieur le plus précis sera sélectionné et ainsi de suite. Le critère d'arrêt sera la meilleure performance possible de l'ensemble, performance qui sera évaluée avec un jeu de données différent de celui utilisé pour l'apprentissage

– En posant EDM (Ensemble Diversity Measure) comme la proportion d'individus pour lesquels la proportion de votes corrects est située entre 10 et 90 (ces points sont donc considérés comme incertains), il existe un algorithme d'amincissement de l'ensemble. Le but est d'enlever itérativement les classifieurs qui ont le plus souvent tord sur l'ensemble des points incertains. La valeur de la précision du classifieur doit être comprise entre $LB = m * EDM + \frac{1-EDM}{c}$ et $UB = \alpha * EDM + m * (1 - EDM)$ avec m la performance moyenne des classifieurs (ceux présents dans l'ensemble en cours), c le nombre de classes et α le paramètre de l'algorithme (valeur recommandée : 0.9)

– Comme cité plus haut, la répartition des classifieurs selon leur diversité par paire et leur erreur moyenne peut se faire de façon graphique. Ainsi, nous pouvons sélectionner les classifieurs permettant de former l'enveloppe convexe de l'ensemble

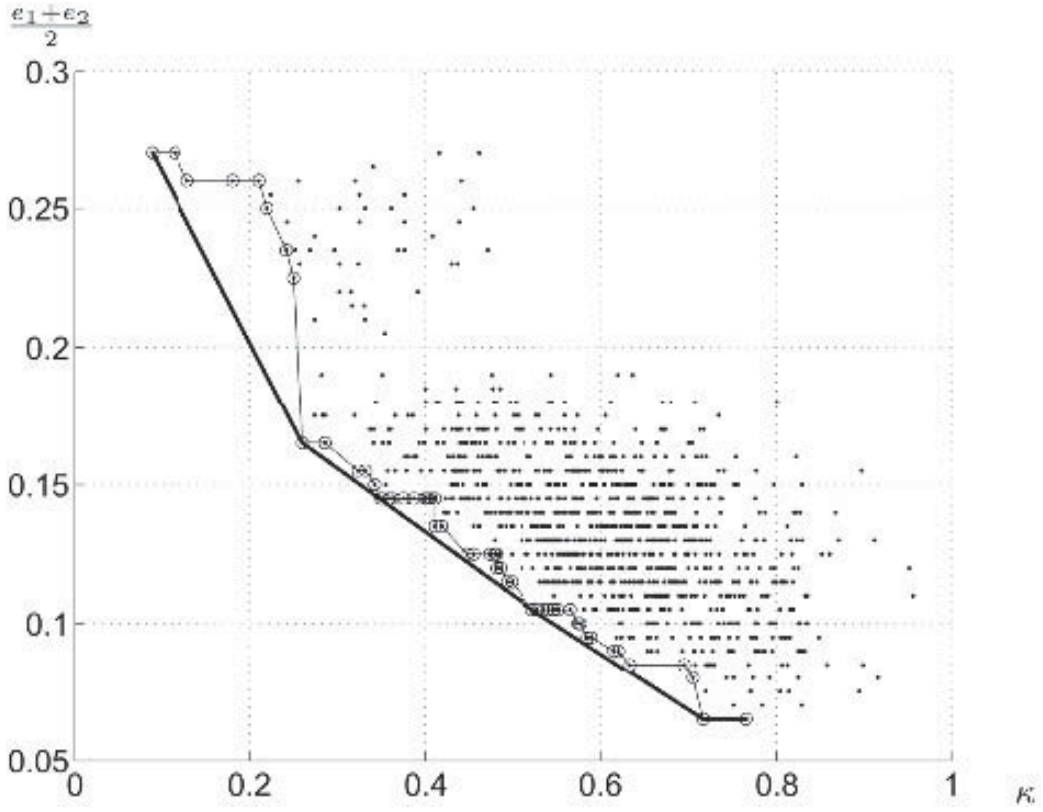

FIGURE 4.7 – Kappa-error de 50 classifieurs et son enveloppe convexe[46]

Toutes ces méthodes ne sont applicables que dans le cas de la combinaison statique de classifieurs ; même si la combinaison peut être construite dynamiquement (en choisissant par exemple les classifieurs apportant la plus grande diversité entre eux) elle passe obligatoirement par la mise en coopération de plusieurs classifieurs. Or, nous nous intéressons à la possibilité de choisir dynamiquement, pour un individu donné, un seul classifieur. Néanmoins, dans le contexte de notre travail, même si ces différentes mesures ne nous permettent pas de répondre au propos d'une combinaison dynamique, elles restent pertinentes pour choisir de façon préalable un jeu de classifieurs, ou pour raffiner ce choix de façon **hors-ligne** après un certain nombre d'itération du processus de décision.
Dans ce cas, comment peut-on mesurer la qualité d'un seul classifieur, sans se référencer à un groupe de classifieur mis en coopération ?
De plus, dans le cadre d'un problème d'OCR, le classifieur seul n'est pas tout, il faut s'intéresser au couple (classifieur, attributs) car la performance du classifieur dépend en partie de la répartition des attributs dans son espace de pavage.

4.3 Sélection dynamique d'un classifieur

4.3.1 Introduction

Notre propos concerne l'étude et la construction de systèmes de combinaison dynamique de classifieurs. L'objectif est la sélection locale de classifieurs dans l'espace des attributs. Les systèmes à classifieurs multiples (MCS) basés sur la combinaison de différents classifieurs sont actuellement très utilisés pour atteindre de hautes performances en reconnaissance de formes ou motifs. Typiquement pour chaque motif entrant, chaque classifieur calcule sa décision, puis celles ci sont fusionnées à partir de votes majoritaires, approches statistiques, théorie de l'évidence ou fonctions de croyance.

La littérature est essentiellement orientée vers l'idée de combiner *a priori* dif-

férentes décisions de classifieurs à la fin d'un apprentissage plus ou moins élaboré (voir les neuf workshops internationaux sur le problème du "Multiple Classifier Systems" [2]). Plusieurs études ont été réalisées sur le sujet [25, 29, 88]. Hand, dans [33], propose de séparer les méthodes en deux classes. La première regroupe les approches combinant les classifieurs sur chacune des entrées ou de leur combinaison [12, 55] ou bien sur les sorties des classifieurs [64, 81]. La seconde est basée sur une combinaison prenant en compte le contexte en utilisant une combinaison *a priori* [40, 46, 65, 72] ou une combinaison en cascade [13, 36, 85]. Mais l'idée de choisir le classifieur dynamiquement en fonction de l'individu requête est encore rare [21].

Selon Singh[70], la première idée de sélection dynamique d'un classifieur est véritablement apparue dans [80], où une description de l'espace des attributs est proposée en utilisant une fonction de dissension entre classifieurs. Intuitivement, les classifieurs tendent à être d'accord sur les cas "faciles" et inversement pour les cas "difficiles". L'espace est alors partitionné et, pour chaque partie, un modèle de régression est estimé. Dans un second temps, une fonction des décisions des classifieurs est construite dynamiquement, selon la position spatiale des individus à classer. Huang et Suen proposèrent dans [39] une première méthode pour résoudre l'hypothèse d'indépendance à l'erreur inhérente aux combinaisons de classifieurs. Puis, Woods soumet une méthode réelle de sélection dynamique de classifieurs utilisant l'analyse de la précision locale [87] et finalement, Giacinto et al. propose une étude générique pour combiner localement les classifieurs dans l'espace des attributs [32, 30].

La figure 4.8 représente les réponses d'un groupe de classifieur à un jeu de données artificielles (Type "Banana"). Les astérisques indiquent les erreurs d'attribution. Ces exemples montrent le comportement de chaque classifieur en fonction

2. http ://www.diee.unica.it/mcs/previous-mcs.html

des positions des individus dans l'espace des attributs. Le point intéressant est la différence de forme des frontières pour chaque classifieur ; les méthodes linéaires (règle de Bayes, Fisher, plus proche moyenne) séparent le problème en deux solutions simples, tandis que les approches non linéaires apportent une meilleure séparation grâce à une frontière complexe. Les méthodes de classification sont présentées brièvement une nouvelle fois à la section 4.3.3, ainsi que les jeux de données à la section 4.3.3.

Pour les méthodes non linéaires, les individus proches des centres des classes sont les mieux classés, les erreurs étant distribuées le long des frontières. Mais ces erreurs sont différentes selon les classifieurs. La sélection dynamique de classifieurs (DCS) améliore la robustesse du processus de classification pour les individus localisés dans les zones "difficiles" de l'espace des attributs (éloignement du centre des classes, frontières etc.). Ce gain est dû à la diversité des erreurs des classifieurs[46, 47, 45] ; excepté pour les cas où tous les classifieurs se trompent, cela améliore la probabilité de bonne classification pour chaque individu.

Dans ce qui suit, nous étudions dans un premier temps les différentes mesures de précision de la décision d'un classifieur dans un voisinage restreint. Une des problématique d'un tel développement a trait à la définition du voisinage et à la pondération des voisins dans la décision en fonction de leur distance à l'individu requête. Dans un second temps, nous développons cette question et proposons une approche simplifiée. Nous comparerons alors les différentes méthodes sur deux types de jeux de données : des données naturelles pour une vrai complexité, et des données artificielles pour une correspondance exacte. Dans cette partie, nous évaluons l'intérêt du DCS et concluons avec quelques remarques.

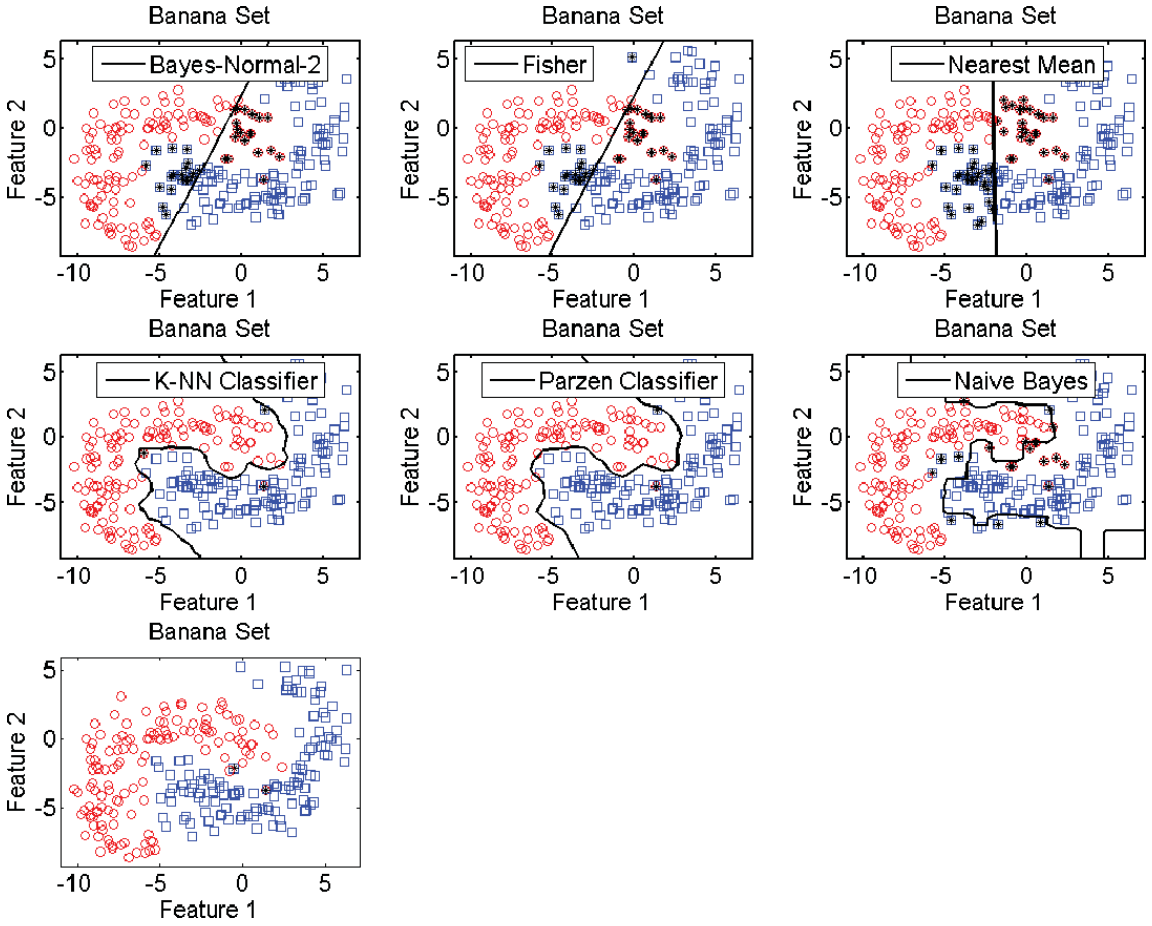

FIGURE 4.8 – Différents comportements de classifieurs dans l'espace des attributs

4.3.2 Décider dynamiquement

Au sein de cette partie, nous étudions les méthodes de combinaison dynamique de classifieurs basées sur la mesure de la précision locale, et l'évaluation de la décision. Voyons en premier lieu le formalisme associé ainsi que son utilisation.

Combiner les décisions

Soit $\mathcal{C}$ un ensemble de N_c classifieurs, tous entraînés sur un jeu de N_ω classes $\Omega = \{\omega_1, ..., \omega_p, ..., \omega_M\}$. L'ensemble $\mathcal{E}$ des individus x est exprimé dans l'espace des attributs $\mathcal{A}$. Soit $\mathcal{E} = \{\{x\}, \{x^*\}\}$ avec $\{x\}$ l'ensemble de tous les individus connus (i.e. l'ensemble d'apprentissage) et $\{x^*\}$ l'ensemble de tous les individus inconnus (i.e. l'ensemble de test). Chaque sortie de classifieur pour l'individu x est notée $C_j(x)$.

$$\begin{aligned} x : (\mathcal{E}, j) &\mapsto C = \{C_1, \ldots, C_j, \ldots, C_{N_c}\} \\ x &\mapsto C_j(x) \end{aligned} \tag{4.16}$$

Soit $B(x^*)$ la définition du voisinage de taille k (équation 4.17). Nous définissons la précision locale estimée $LA_{j,k}(x^*)$ pour le classifieur j pour un individu x^* dans le voisinage $B(x^*)$. Le voisinage considéré est restreint à l'ensemble d'apprentissage $\mathcal{L} = \{x\}$ pour lequel chaque classe d'appartenance est définie.

$$\begin{aligned} B(x) \quad : \quad & B(x) \subset \{x\} \text{ avec } \{x\} \subset \mathcal{E} \text{ et } card(B(x)) = k \\ & \text{ie } d(x,y) < d(x,z)\ \forall y \in B(x) \text{ et } \forall z \in \overline{B(x)} \end{aligned} \tag{4.17}$$

La fonction de décision permet alors de déterminer la meilleure décision d'assignation pour l'individu x^*, pour lequel la classe d'appartenance est inconnue. La décision est prise au sens de la précision locale $LA_{j,k}(x^*)$ dans le voisinage $B(x^*)$ de taille k. La décision est celle donnée par le classifieur présentant la plus grande précision dans x^* (algorithme DCS-LA) comme définie par l'équation 4.18.

$$C_j(x), \quad LA_{j,k}(x^*) = \max_{0<i\leq N_c}(LA_{i,k}(x^*)) \tag{4.18}$$

Prendre en compte le voisinage (MCB)

Afin de préciser la notion de voisinage local, Giacinto[31] propose une extension utilisant le concept de comportement de plusieurs classifieurs (multiple classifier behavior ou MCB)[39] pour le calcul des k plus proches voisins.

Le voisinage d'un individu donné x est défini par le vecteur $MCB(x) = \{C_1(x), \ldots, C_N(x)\}$ dont les éléments sont les décisions de classification pour l'individu x prises par les N_c classifieurs. Giacinto propose alors une mesure de similarité entre les individus x_1 et x_2 :

$$\mathcal{S}(x_1, x_2) = \frac{1}{N} \sum_{i=1}^{N} T_i(x_1, x_2) \tag{4.19}$$

où les fonctions $T_i(x_1, x_2)$, $i \in [1, N_c]$ sont définies par :

$$T_i(x_1, x_2) = \begin{cases} 1 & si \quad C_i(x_1) = C_i(x_2) \\ 0 & sinon \end{cases} \tag{4.20}$$

La fonction $S(x_1, x_2)$ prend valeurs dans $[0, 1]$, 1 si les N décisions de classification sont les mêmes pour x_1 et x_2, et 0 si toutes les decisions sont totalement différentes pour les deux individus. Ainsi, Giacinto propose de choisir les k plus proches voisins de l'individu à classifier, en terme de distance Euclidienne, mais satisfaisant une contrainte de seuil sur la similarité MCB. Dans [31], Giacinto note que, grâce à l'utilisation de l'information du MCB, les performances sont peu affectées par la valeur du paramètre k.

Mesurer la précision

La qualité de la décision finale dépend de la qualité de l'estimation de la précision locale. Elle est basée sur la capacité d'un classifieur à associer la décision

$C_j(x)$ à la bonne classe $L(x)$ pour les individus de l'ensemble d'apprentissage $\mathcal{L}$ autour de l'individu à classer x^*. Nous supposons alors que le classifieur j affecte x^* à la classe ω_p : $(C_j(x^*) = \omega_p)$ et proposons différentes mesure de la précision locale pour cette décision.

Soit S, l'ensemble des bonnes réponses dans le k-voisinage de x^*, c'est à dire l'ensemble des individus bien classés dans le voisinage de x^*.

$$S = \{x \in B_k(x^*), C_j(x) = L(x)\} \tag{4.21}$$

Trois propositions initiales de la mesure de la précision locale ont été initialement définies [87, 30, 21] :

– **Précision locale *a priori*** : $LAO_{j,k}$ pour le classifieur j selon un voisinage de k individus :

$$LAO_{j,k} = \frac{card(S)}{k}$$

Cette mesure est appelée "*a priori*" (ou *overall LA*) car elle ne prend pas en compte la classe attribuée à l'individu x^* par le classifieur j. Elle analyse la capacité du classifieur à prendre la bonne décision pour les classes existantes dans le voisinage de x^*.

– **Précision locale *a posteriori*** (ou *local class accuracy*) : $LCA_{j,k}$ pour le classifieur j selon un voisinage de k individus :

$$\begin{aligned} LCA_{j,k} \quad = \quad & \frac{card(S_{\omega_p})}{card(R)} \\ & \text{avec } S_{\omega_p} = \{x \in S, C_j(x) = \omega_p\} \\ & \text{et } R = \{x \in B_k(x^*), C_j(x) = \omega_p\} \end{aligned} \tag{4.22}$$

Cette mesure reprend le principe de la méthode classique de mesure de la précision particulièrement utilisée en indexation, et estime le ratio entre le nombre d'individus bien classés dans ω_p et ceux assignés à cette classe.

Elle est alors dédiée à la classe d'affection et limitée par l'hypothèse que la classification ne fasse pas d'erreur affectant la classe ω_p à un individu. D'autre part, cette méthode ne prend pas en compte les individus oubliés appartenant à ω_p mais non retrouvés ($C_j(x) \neq \omega_p$).

– **Précision locale probabiliste** : LAP_j pour le classifieur j :

$$LAP_j(x^*) = \frac{\sum_{x_n \in \omega_p} \left(\frac{\hat{P}(\omega_p|x_n)}{d_n} \right)}{\sum_{m=1}^{M} \sum_{x_n \in \omega_m} \left(\frac{\hat{P}(\omega_p|x_n)}{d_n} \right)} \quad \text{avec } d_n = d(x_n, x^*), \forall x_n \in B(x^*) \tag{4.23}$$

d_n est la distance entre l'individu connu x_n et x^*. Didaci n'indique pas le type de distance utilisé (L1, L2, etc.). Nous discuterons plus loin du choix de la métrique de la distance. $\hat{P}(\omega_p|x_n)$ est l'estimation de la probabilité *a posteriori* d'avoir la classe ω_p avec les attributs de x_n. Une telle formulation exprime un rapport du type densité intra-classe sur la densité inter-classes. L'aspect intra-classes apparaît au numérateur par la capacité des individus de la classe d'affectation à représenter cette classe avec leurs attributs. L'aspect inter-classes est d'autre part lié à l'incapacité des autres classes, dans le voisinage de x^*, à représenter la classe w_p. Pondérer avec la distance d_n permet de réduire l'impact de cette information par rapport à la distance dans le voisinage d'ordre k.

La limite de la proposition de Didaci[21] réside dans le fait que le dénominateur de sa formule intègre également l'estimation du numérateur, ce qui déforme le jugement. La raison d'une telle formulation est la nécessité de normaliser la valeur de la précision entre 0 et 1, ce qui fait perdre le pouvoir de discrimination quand deux (ou plus) classes ont presque les mêmes centres.

Toutes ces méthodes donnent une estimation de la précision locale, normalisée

entre 0 et 1, ce qui est nécessaire pour les comparer l'une à l'autre. Cependant, la problématique n'est pas de choisir dynamiquement la meilleure méthode de calcul de la précision locale mais le meilleur classifieur selon une métrique de précision choisie *a priori*. Ainsi, toute comparaison relative des mesures de précision est utilisable. Une fois la contrainte de normalisation rejetée, nous pouvons essayer de définir une nouvelle méthode de mesure de la précision à partir de celles définies précédemment.

Pour le calcul de la précision locale, il est nécessaire de définir les individus connus autour de l'individu inconnu à classer. Didaci définit la précision locale *a priori* comme étant la moyenne de toutes les précisions locales de tous les classifieurs, sur tous les individus connu du voisinage. L'inconvénient de cette méthode est qu'elle suppose le voisinage sphérique (donc défini par une norme euclidienne). Il semble nécessaire de considérer un individu connu en fonction de sa distance à l'individu inconnu. Comme nous n'avons pas d'information à propos de la forme du voisinage, nous proposons d'utiliser la distance entre l'individu inconnu et chacun des individus connus (toujours dans le voisinage) comme pondération de la précision locale (i.e. d_n). De plus, nous ne possédons aucune information à propos de l'espace des attributs, particulièrement à propos de son orthogonalité ou de sa normalité. C'est pourquoi nous nous restreignons à une distance L1 et nous définissons ainsi d_n comme :

$$d_n = |x_n - x^*|, x_n \in B(x^*) \tag{4.24}$$

Dans l'optique de choisir la meilleur fonction de pondération, nous avons d'abord besoin de savoir comment évolue la distance. Nous avons ainsi étudié son évolution pour chaque jeu de données dans la région la plus dense de l'espace de attributs ainsi que dans la plus clairsemée. Les figures 4.9 et 4.10 montrent le comportement de la distance (tracé avec l'écart-type), son inverse, log et exponentielle. Le paramètre ε est choisi comme constante de valeur faible pour éviter les opérations non définies (i.e. division par zéro). La distance tracée ici est la

moyenne de toutes les distances de chaque individu aux autres. Nous présentons aussi dans les tables 4.1, 4.2, 4.3 et 4.4 l'évolution de la distance dans le voisinage de l'individu moyennement plus proche des autres (l'individu dont la distance moyenne à tous les autres est la plus faible) et celui moyennement le plus éloigné (l'individu dont la distance moyenne à tous les autres est la plus forte).

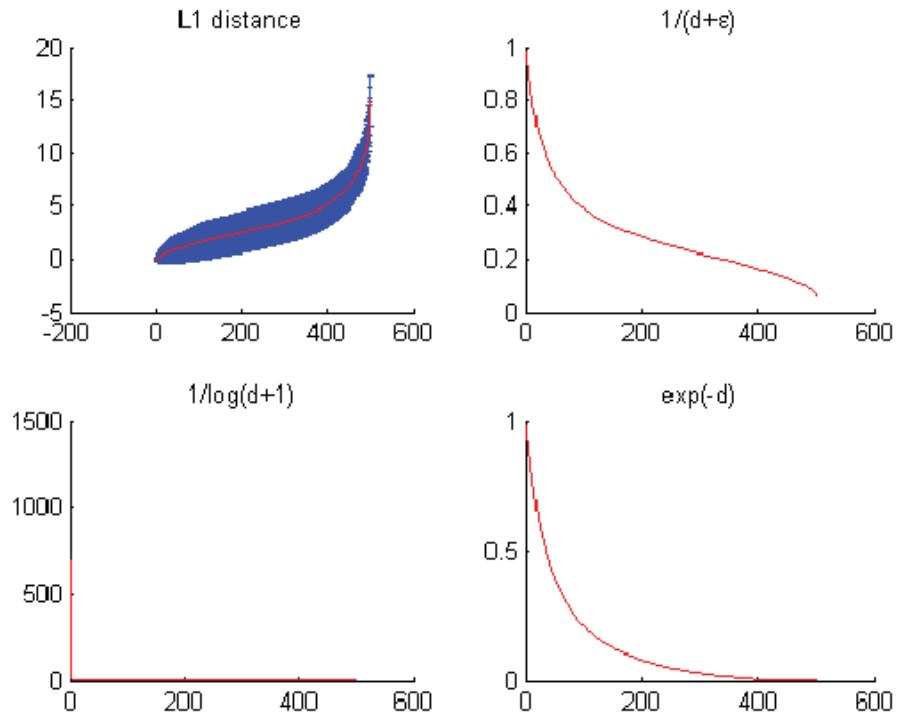

FIGURE 4.9 – Evolution de la distance entre individus du jeu de données Complex (d=L1)

Le jeu de données Breast Cancer Wisconsin présente la plus grande variation de distance entre les individus (tables 4.3 et 4.4). Les données du jeu Complex présentent également de fortes variations mais de beaucoup plus petites distances (tables 4.1 et 4.2). Avec les résultats associés, nous observons que $1/(d+\varepsilon)$ est plus discriminant que $\exp(-d)$ qui est limité par son domaine de définition ($[0,1]$). L'expression $1/\log(1+d)$ atténue également la répartition entre les voi-

d	0	0.010	0.014	0.015	0.018	0.029	0.030	0.031	0.034
$1/(d+\varepsilon)$	1.000	0.990	0.986	0.985	0.982	0.972	0.970	0.970	0.967
$1/\log(1+d)$	969	88	66	61	52	33	32	31	29
$\exp(-d)$	1.000	0.990	0.986	0.985	0.982	0.971	0.970	0.969	0.967

TABLE 4.1 – 9 plus proches voisins pour l'individu moyennement le plus proche - Données Complex

d	1.580	2.910	3.000	3.890	4.400	4.530	4.720	4.910	5.320
$1/(d+\varepsilon)$	0.388	0.256	0.250	0.204	0.185	0.181	0.175	0.169	0.158
$1/\log(1+d)$	1.050	0.733	0.721	0.630	0.593	0.585	0.573	0.563	0.542
$\exp(-d)$	0.206	0.544	0.049	0.020	0.012	0.010	0.009	0.007	0.005

TABLE 4.2 – 9 plus proches voisins pour l'individu moyennement le plus lointain - Données Complex

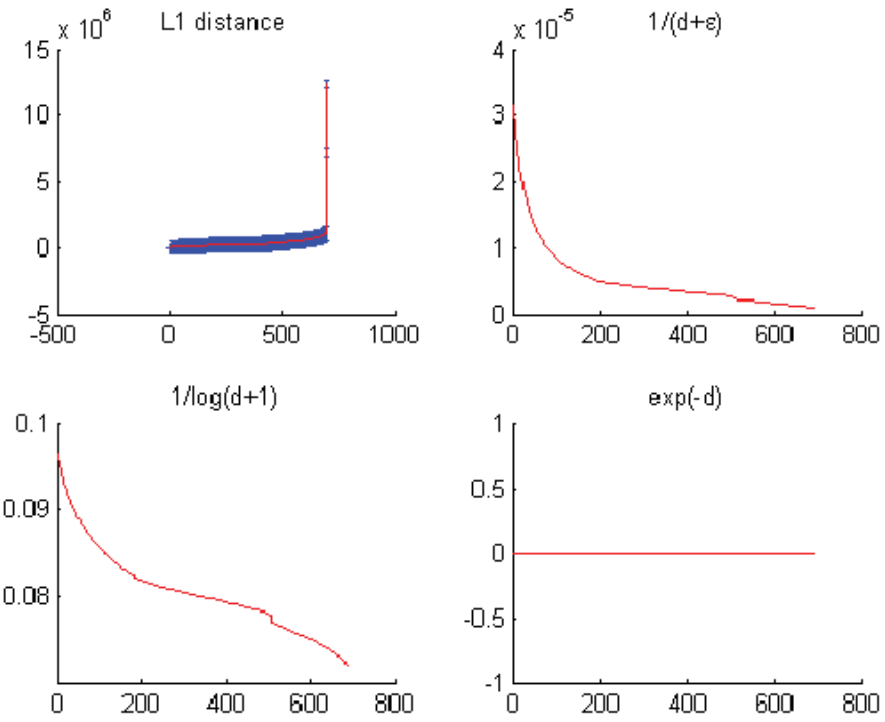

FIGURE 4.10 – Evolution de la distance pour les individus du jeu Breast Cancer Wisconsin dataset (d=L1)

d	0.360	9.30	133	162	480	766	1250	1250	1530
$1/(d+\varepsilon)$	0.027	0.010	0.007	0.006	0.002	0.001	0	0	0
$1/\log(1+d)$	0.277	0.220	0.204	0.196	0.162	0.151	0.140	0.140	0.136
$\exp(-d)$	2e-16	4e-41	1e-58	4e-71	3e-209	0	0	0	0

TABLE 4.3 – 9 plus proches voisins pour l'individu moyennement le plus proche - Données Breast Cancer Wisconsin

d	5e6	1e7	1e7	1e7	1e7	1e7	1e7	1e7	1e7
$1/(d+\varepsilon)$	1e-7	8e-8	8e-8	8e-8	8e-8	8e-8	8e-8	8e-8	8e-8
$1/\log(1+d)$	0.064	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061
$\exp(-d)$	0	0	0	0	0	0	0	0	0

TABLE 4.4 – 9 plsu proches voisins pour l'individu moyennement le plus lointain - Données Breast Cancer Wisconsin

sins. Comme il s'agit d'un problème de dynamique de distance, le choix de la méthode de pondération est donc basé sur le gradient Δ de $f(d)$. Par exemple, quand $\Delta(d) \to 20$, $\Delta(\frac{1}{d+\varepsilon}) \to 5.10^{-2}$ tandis que $\Delta(\exp(-d)) \to 0$, nous considérons alors tout le voisinage dans un cas non dans l'autre. C'est pour cela que nous proposons l'usage direct de la distance L1 comme pondération de la précision locale afin d'améliorer les résultats en utilisant une méthode de calcul nécessitant un faible coût de calcul.

$LAS_{j,k}$: mesure de la densité des individus correctement classés ($x \in S$).

$$LAS_{j,k} = \frac{1}{k} \sum_{x \in S} \frac{1}{d_n + \varepsilon} \tag{4.25}$$

Cette mesure estime le taux d'individus bien classés en prenant en compte la distance à l'individu requête. De la même manière que la précision locale *a priori*, elle ne prend en compte que la capacité du classifieur à ne pas faire d'erreur sur l'ensemble des classes du voisinage.

4.3.3 Résultats

Les jeux de tests

Afin d'étudier le comportement des sélections dynamiques, nous réalisons différents types de tests. Le premier est basé sur un jeu de données artificielles simulant différents types de nuages spécifiques dans un espace d'attributs à deux dimensions. Le second utilise des données réelles sur de plus grandes dimensions. Nous rappelons que les bases de données sont décrites au chapitre 2. Tous les tests se font sur des problèmes à deux classes, utilisant les données suivantes :

- données artificielles : 4 types de problèmes (Banana, Complex, Difficult et Simple) avec 250 individus par classe, sur 2 dimensions (voir la figure 4.11), produites avec un générateur aléatoire
- données naturelles : 5 types de problèmes (base *UCI Repository of machine learning databases* [3]) décrites dans la table 4.5

Nom	Taille classe 1	Taille classe 2	Dimension
Wisconsin breast cancer	458	241	10
Ionosphere	225	126	34
Pima Indians diabetes	500	268	8
Sonar	97	111	60
German	700	300	24

TABLE 4.5 – Caractéristiques du jeu de données naturelles

Les classifieurs combinés

La performance de la combinaison est évaluée à partir des 6 classifieurs suivants :

- QBNC : Classifieur quadratique Bayesien Normal à un noyau par classe(Bayes-Normal-2) ;
- FLSLC : Classifieur linéaire moindres carrées de Fisher ;

3. http ://www.ics.uci.edu/∼mlearn/MLRepository.html

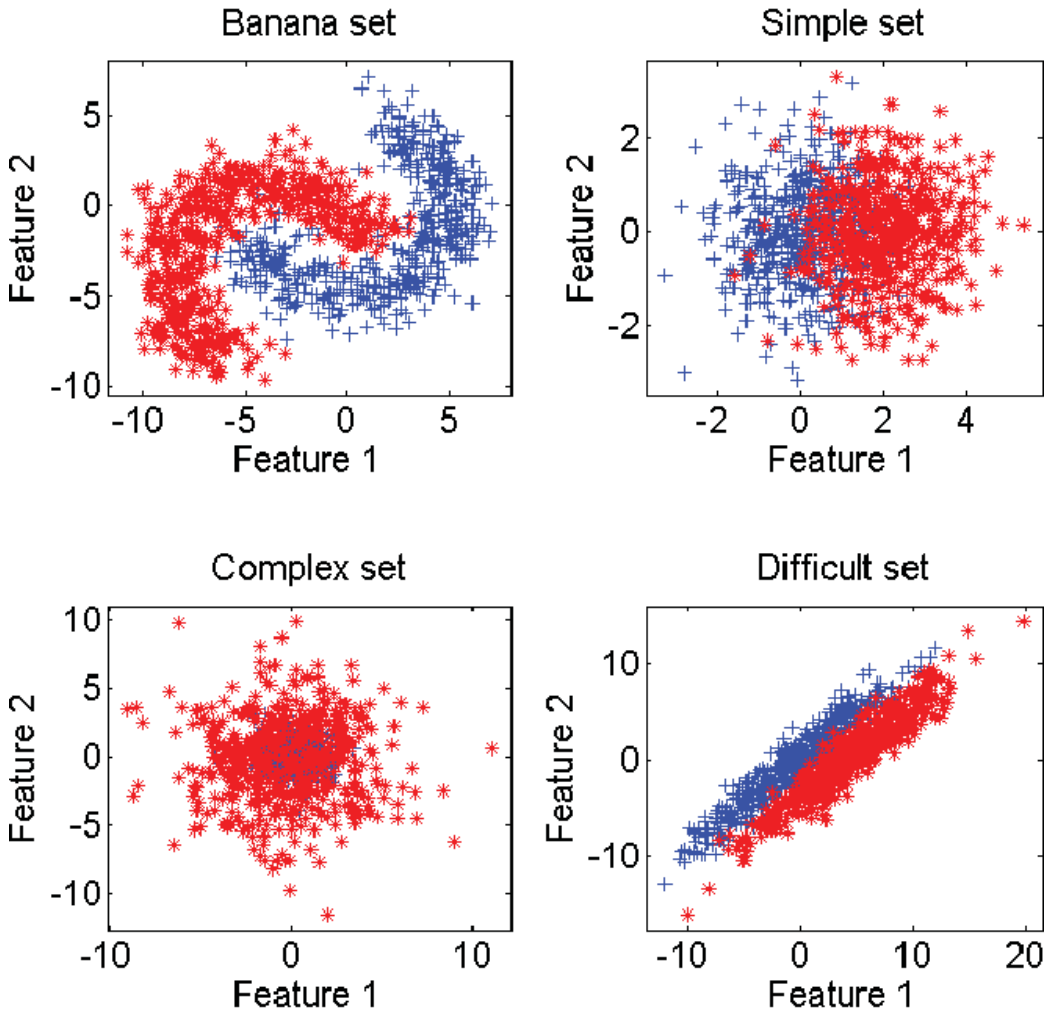

FIGURE 4.11 – Sorties des différents générateurs de données artificielles

– NMC : Classifieur aux plus proches moyennes ;
– KNNC : Classifieur aux k plus proches voisins ;
– ParzenC : Classifieur Parzen ;
– BayesN : Classifieur Bayesien naïf.

Chacun de ces classifieurs a d'abord été entraîné et testé sur tous les jeux de données afin d'en évaluer leurs performances globales. Les jeux d'entraînement et de test ont été construits par **bootstrapping**[47] à partir des données originales, avec intersection nulle entre les 2 jeux de données à chaque fois.

Le bootstrap est généralement utile à l'estimation de la distribution d'une statistique (moyenne, variance, etc.). Parmi les différentes méthodes citons l'algorithme Monte Carlo[57] permettant un ré-échantillonnage simple. L'idée est de faire une première mesure de la statistique recherchée sur le jeu de données disponible, puis d'échantillonner aléatoirement (avec remise) ce jeu de données et de recommencer la mesure. Cette méthode sera répétée jusqu'à atteindre une précision acceptable.

Les classifieurs KNNC, ParzenC et NMC sont **adaptatifs**, la valeur de leurs paramètres (i.e. k pour KNNC et ParzenC) est optimisée durant la phase d'apprentissage. Ensuite, une fois le paramètre fixé, il n'est pas remis en cause durant la reconnaissance.

Les méthodes de combinaison

Nous choisissons de tester les différentes combinaisons selon les différentes façons de calculer la précision locale :
– LAO : précision locale *a priori* selon [87] (équation 4.22) ;
– LCA : précision locale *a posteriori* toujours selon [87] (équation 4.22) ;
– LAP : La méthode de Didaci [21] (équation 4.23) ;
– LAS : la méthode proposée selon l'équation 4.25.

Les méthodes LAO, LCA et LAS ont également été comparées en utilisant la contrainte de similarité de Giacinto (approche MCB)[39, 31]. Pour ces tests, nous faisons évoluer la taille du voisinage (k) afin de tester son influence sur les résultats. Nous fixons une valeur maximale de k plus faible que le nombre total d'individus afin de rester dans le cas de l'analyse de la précision locale, et non globale. Pour les résultats présentés ci-dessous, nous utilisons la légende de la figure 4.17.

Afin de situer les résultats exprimés en temps de calcul, il faut savoir que tous les tests sont exécutés sur un ordinateur équipé d'un AMD Athlon XP2600+ avec 1Go RAM sous Ubuntu 7.x.

4.3.4 Discussion

Les approches directes

L'analyse des comportements sur les jeux de données Banana et Difficult (figures 4.12 et 4.13) donnent une première idée sur le comportement des mesures de la précision. Pour le jeu Banana, la mesure de précision *a posteriori* (LCA) est la seule qui augmente en fonction de la taille du voisinage. Dans ce cas de figure, l'accroissement de la taille du voisinage (k) permet l'accroissement du ratio entre le nombre d'individus bien classés et le nombre d'individus reconnus. Pour le jeu de données Difficult (figure 4.13), c'est la méthode LAP qui augmente légèrement avec k. Ici, l'accroissement de k est pondéré par la distance à l'individu à reconnaître pour améliorer la mesure de précision. Dans tous les autres cas, la précision est au mieux à peu près stable et au pire décroît légèrement avec l'accroissement de la taille du voisinage.

Les approches par limitation du voisinage

A propos des résultats produits avec les jeux de données artificielles (figures 4.12 et 4.13) notons la sensible variation de performance pour les méthodes LAO, LCA and LAP (avec ou sans MCB), en comparaisons des résultats obtenus avec les jeux de données naturelles (figures 4.14,4.15 et 4.16). Les méthodes LAO and LCA prennent en compte le nombre d'individus du voisinage bien classés, mais, lorsque le voisinage grandit, la performance tend vers celle du plus efficace des classifieurs pris seul, ce que nous notons en intégrant les performances des classifieurs seuls (tables 4.7 et 4.9) dans les figures ci-dessous. Prenons *a contrario* les données naturelles, dont la densité de distribution est plus faible (les individus sont mieux répartis dans l'espace des attributs), une approche statistique est alors plus représentative du comportement global (et non le meilleur classifieur). Comme les lois de probabilité sont inconnues, nous mesurons la distribution des individus selon 2 critères : la distance moyenne Euclidienne $\bar{d}$ entre individus, et son écart-type σ_d. Les données sont présentées dans le tableau 4.6. Plus la distance moyenne et l'écart-type sont grands, moins la taille du voisinage pour le DCS a d'influence sur la performance globale. Il est également intéressant de noter que la méthode LAS (avec ou sans MCB) semble la moins sensible à la taille du voisinage, excepté bien sûr, le fait que plus k est grand, moins bonne est la performance globale (car elle tend vers la performance globale du meilleur classifieur pris seul).

Notons dans la table 4.6 que pour les données artificielles (Banana, Complex, Difficult et Simple) les moyennes et écarts-type sont du même ordre, pour une dimension fixée à 2. Ce qui n'est pas le cas pour les données naturelles, pour lesquelles les rapports entre les distances moyennes et les écarts-type indiquent la très grande complexité de ces bases.

données	$\bar{d}$	σ_d	dimension
Wisconsin breast cancer	$7,62.10^{11}$	$7,53.10^{25}$	10
Pima Indians diabetes	$3,03.10^{4}$	4.10^{9}	8
German	$2,17.10^{3}$	$1,21.10^{7}$	24
Ionosphere	18,99	164,50	34
Sonar	3,99	3,71	60
Banana set	66,7	4.10^{3}	2
Complex set	71,13	$4,4.10^{3}$	2
Difficult set	68,14	$3,9.10^{3}$	2
Simple set	65,7	$3,8.10^{3}$	2

TABLE 4.6 – Distance moyenne et écart-type entre individus

L'apport du Multiple Classifieur Behavior (MCB)

Les k plus proches voisins d'un individu de test sont d'abord identifiés dans le jeux de données d'apprentissage. Ceux caractérisés par des MCB "similaires" (équation 4.19) à ceux des individus inconnus (issus du jeu de test) sont alors sélectionnés pour calculer les précisions locales et effectuer la sélection dynamique.

Les tables 4.7 et 4.9 représentent la meilleure performance pour chaque classifieur seul, puis pour chaque méthode de DCS (pour tous k). Notons en premier lieu (table 4.9), le gain de performance qu'apporte le MCB aux méthodes LAO et LCA. Ces résultats correspondent aux remarques de Huang formulées dans [39]. En seconde constatation, les résultats de la méthode LAS sont dégradés lors d'une combinaison avec l'approche MCB. En effet, la méthode LAS, basée sur un calcul moyen, nécessite un maximum d'informations à propos du voisinage, or le MCB ne prend pas en compte les individus les plus proches en terme de distance, mais seuls ceux en terme de réponse de classifieur. Ainsi, pour chaque méthode de calcul de la précision globale, le MCB améliore la probabilité de prendre les individus pour lesquels le système présente une meilleure **certitude** (qui peut s'apparenter à une mesure de crédibilité en théorie de l'évidence), sauf pour notre approche qui n'en a pas besoin. Ainsi, le MCB améliore les performances en sélectionnant l'information selon la similarité des sorties des classifieurs. Cependant, il y a une

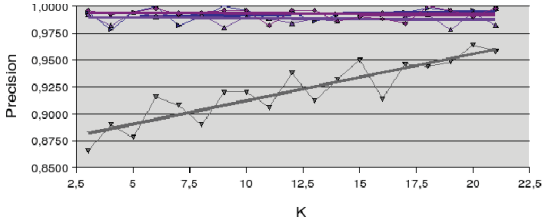

FIGURE 4.12 –

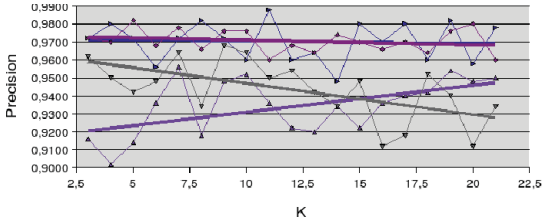

FIGURE 4.13 –

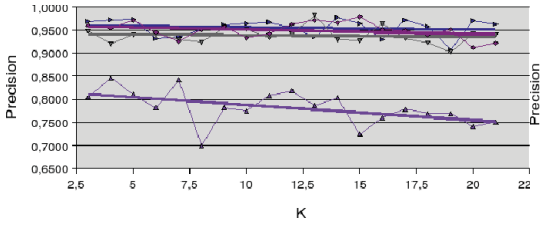

FIGURE 4.14 –

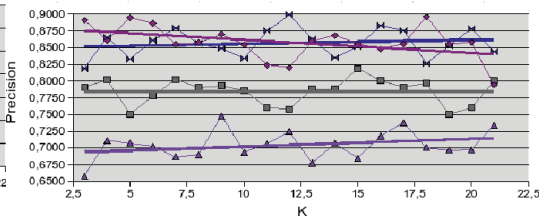

FIGURE 4.15 –

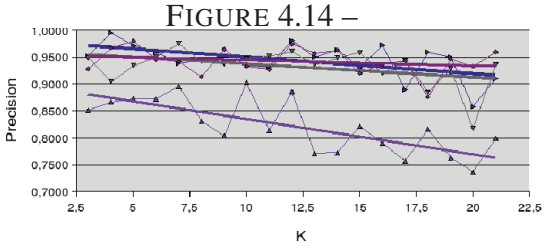

FIGURE 4.16 –

FIGURE 4.17 – légende

FIGURE 4.18 – Influence de la taille du voisinage pour les données ”Banana”(Fig. 4.12), ”Difficult”(Fig 4.13), ”Ionosphere”(Fig 4.14), ”diabete”(Fig 4.15) and ”Sonar”(Fig 4.16)

exception pour les méthodes prenant en compte les distances entre les individus connus et l'individu requête x^*.

Nous avons également étudié le comportement d'un groupe de classifieurs construit *a priori* par optimisation de la diversité. La mesure utilisée est celle de l'entropie, et nous avons sélectionné, pour chaque jeu de données, l'association de classifieurs présentant l'entropie la plus forte. Le taux de bonne classification obtenu est présenté en table 4.8.

Nous observons ici que la méthode de construction d'un jeu de classifieur ba-

sée sur la diversité n'améliore pas les résultats globaux obtenus par les classifieurs seuls. Ce qui nous laisse à penser que cette méthode n'est pas adaptée pour de la construction dynamique d'ensemble de classifieurs, comparé à une construction statique *a priori*.

	Banana	Complex	Difficult	Simple	Diabetes	Sonar	German	Cancer	Ionosphere
Quadratic	0.87	0.83	0.96	0.86	0.71	0.81	0.74	0.96	0.92
Pseudo Fisher	0.87	0.47	0.96	0.86	0.72	0.84	0.69	0.95	0.85
Nearest Mean	0.82	0.47	0.72	0.86	0.59	0.70	0.58	0.52	0.76
KNN	0.99	0.93	0.97	0.93	0.86	0.92	0.86	0.84	0.94
Parzen	0.99	0.88	0.95	0.89	0.74	0.92	0.63	0.51	0.92
Naive Bayesian	0.94	0.80	0.69	0.85	0.74	0.88	0.68	0.98	0.50
Best Performance	0.99	0.93	0.97	0.93	0.86	0.92	0.86	0.98	0.94

TABLE 4.7 – Meilleure performance pour chaque classifieur seul

	Banana	Complex	Difficult	Simple	Diabetes	Sonar	German	Cancer	Ionosphere
Mesure de l'entropie	0.99	0.58	0.91	0.87	0.76	0.80	0.74	0.67	0.90

TABLE 4.8 – Meilleure performance par combinaison construite sur la diversité

	Banana	Complex	Difficult	Simple	Diabetes	Sonar	German	Cancer	Ionosphere
LAO	1.00	0.93	0.98	0.95	0.90	0.98	0.89	0.98	0.98
LAO+MCB	1.00	0.94	0.98	0.95	0.90	0.99	0.90	0.99	0.99
LCA	0.96	0.91	0.97	0.91	0.82	0.97	0.85	0.97	0.98
LCA+MCB	1.00	0.95	0.98	0.96	0.89	1.00	0.89	0.99	0.99
LAP	0.99	0.86	0.97	0.87	0.75	0.90	0.65	0.65	0.85
LAS	1.00	0.95	0.99	0.95	0.90	0.99	0.92	0.99	0.98
LAS+MCB	0.94	0.90	0.98	0.93	0.83	0.96	0.85	0.99	0.99
Best Performance	1.00	0.95	0.99	0.96	0.90	1.00	0.92	0.99	0.99

TABLE 4.9 – Meilleure performance pour chaque méthode testée

L'intérêt de la combinaison dynamique

Les résultats obtenus par combinaison statique *a priori* de classifieurs par optimisation de la diversité sont très faibles, comparés aux classifieurs seuls ou combinés dynamiquement, pour les données "Complex" ou "Breast Cancer Wisconsin". Ce comportement peut s'expliquer par le fait que la diversité est un critère basé sur le comportement global qui compare un classifieur par rapport au groupe, tandis que la précision locale est un critère basé sur la répartition des individus dans

l'espace des attributs.

Notons finalement que pour les jeux de données testés, les résultats donnés par DCS sont toujours meilleurs que ceux obtenus par le meilleur classifieur pris seul (tables 4.7 et 4.9).

Les méthodes LCA+MCB et LAS donnent même le meilleur taux de reconnaissance global (table 4.9) pour un temps de calcul similaire (environ 80ms par individu pour les données artificielles sur un AMD Athlon XP2600+ avec 1Go RAM sous Ubuntu 7.x). Les bénéfices attendus de la sélection dynamique de classifieurs sont pour les très grosses bases de données, il est alors possible de n'implémenter que quelques classifieurs, certainement pas les meilleurs en terme de performance, mais au moins les meilleurs en terme de temps de calcul. Ainsi, le temps de calcul est grandement amélioré par rapport à un unique classifieur trop complexe. Ceci est particulièrement intéressant face à l'augmentation des techniques d'extraction de la connaissance, **data mining**[26], où une réponse de plus en plus rapide est attendue alors qu'il faut gérer des quantités de données de plus en plus grandes.

Néanmoins, pour être complet, il faut prendre en compte le coût le plus important de la méthode qui est liée à une recherche de voisinage sur l'ensemble d'apprentissage. Une implémentation pratique de la méthode verrait ici un très grand gain à utiliser des approches optimisées et une organisation de l'ensemble d'apprentissage sous forme d'arbre par exemple.

4.3.5 Conclusion

Deux idées ont été présentées ici. La première est la mesure de la précision locale pour la classification. La seconde est l'utilisation de cette précision locale

dans les systèmes de classifieurs dynamiques (DCS). Nous avons montré l'amélioration apportée par ces idées, comparée à l'utilisation *a priori* d'un seul classifieur. Nous avons également expliqué l'importance d'utiliser, pour le calcul de la précision locale d'un classifieur, un voisinage de faible taille (la précision locale tendant vers la précision globale lorsque la taille du voisinage augmente). Nous avons finalement proposé une façon différente de calculer la précision locale qui donne des résultats aussi bons que ceux donnés par la meilleure des méthodes existantes mais avec une plus faible complexité algorithmique.

Il est possible de développer d'autres méthodes de mesure de la précision locale en utilisant la connaissance sur les classifieurs, spécialement en utilisant la position des individus inconnus par rapport aux frontières et aux centres des classes. Cette adaptation à la configuration est proche de celle du principe des machines à support vectoriel (SVM) [68]. Cependant, sur ce genre d'approche, quelques précautions sont à prendre en compte avec les contraintes imposées par les données naturelles (i.e. des zéros sur les matrices de variance/covariance).

Nous montrons ici l'intérêt d'utiliser l'information sur la répartition des individus dans l'espace des attributs, spécialement dans un voisinage donné, en utilisant une distance L1 pour le calcul de la précision locale. La suite de ce travail pourra être d'exploiter la forme du voisinage, en utilisant l'information portée par la covariance entre individus (matrices de variance/covariance, valeurs propres), dans l'idée de créer une métrique adaptative comme montré dans [35]. Ce genre de métrique réduit le voisinage dans les directions le long desquelles les centres des classes diffèrent, avec l'intention de finir sur un voisinage pour lequel les centres des classes coïncident (et avoir ainsi le plus proche voisinage approprié pour la classification).

Néanmoins, la mesure de la précision est le point le plus intéressant pour une chaîne de traitement de l'information [27, 26]. Avec la capacité d'évaluer la qua-

lité de la décision, le système possède l'information à réinjecter pour l'exécution dynamique des opérateurs et la prise de décision [79].

Si nous revenons au principe de la décision avec intégration de la précision, en fin de chaîne de traitement tel qu'il est présenté au chapitre 3, nous venons de voir comment calculer cette précision pour les classifieurs, et comment l'utiliser pour choisir dynamiquement un classifieur, *a priori* le mieux adapté pour traiter l'individu inconnu en cours d'analyse. Le chapitre suivant détaille comment il est possible de calculer la précision en sortie d'un opérateur de calcul d'attributs.

"*Une description qui dépasse dix mots n'est plus visible.*"
Jules Renard

Chapitre 5

Attributs

Sommaire

5.1 Introduction

Au cours de notre approche, nous avons repris la notion de précision héritée de la métrologie pour l'exploiter dans un système décisionnaire à base de traitement d'images. Après avoir montré, dans le chapitre 3, comment utiliser cette mesure issue de l'extraction d'un attribut, nous avons expliqué comment elle pouvait être produite à l'issue d'un étage de classification soit directement, soit après sélection du meilleur classifieur pour produire dynamiquement une étiquette (décision finale).

Dans ce chapitre, nous allons nous intéresser à la production de cette mesure de précision dans la phase d'extraction d'un attribut pour associer à une mesure (le scalaire, le vecteur ou la matrice produit par l'attribut) une valeur de précision.

Le fait d'avoir organisé ce manuscrit en commençant par l'usage plutôt que la production nous a permis de comprendre que la mesure de précision devait être comprise entre 0 et 1 (précision optimale).

Les objectifs en arrière plan de ce travail concernent la gestion semi-supervisée de la chaîne de traitement d'images. Ce qui est visé, c'est la capacité du système à faire évoluer le paramétrage des différents opérateurs qui la composent (évolution linéaire) ou à changer les opérateurs (évolution non linéaire). Le critère utilisé pour le pilotage de la chaîne correspond alors au rapport du coût sur la précision finale.

Par chaîne de traitements, nous entendons l'enchaînement des opérateurs de traitement d'image (débruitage, segmentation, calcul d'attributs etc.) tel que nous le trouvons pour une application classique en imagerie numérique (voir la figure 5.1).

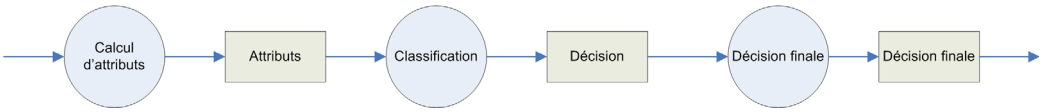

FIGURE 5.1 – Chaîne de traitements linéaire simple

5.2 Précision et attributs

L'application COROC (Cognitive Optical Recognition for Old Characters) qui a servi de toile de fond à ce travail portait essentiellement sur la reconnaissance de caractères manuscrits. Ce secteur technologique et scientifique a vu apparaître de nombreuses méthodes introduisant des notions d'invariances de disposition du caractère(comme les invariants de Hu[37][38], par exemple) basées sur des moments statistiques classiques, comme les moments cartésiens ou les moments centrés[69]. D'autres moments sont apparus ensuite comme les moments de Zernike[92] accompagnés d'une suite d'invariants calculés sur cette base[42][43]. Il existe aussi d'autres méthodes de descriptions des formes basées sur des décompositions en séries de Fourier[6][83] ou encore les coefficients d'ondelettes[8][10][59].

Puis, périodiquement avec l'apparition des nouvelles approches, le monde scientifique a produit des études comparatives pour confronter ces méthodes[3][4][84], principalement sur la base MNIST. Ces études montrent dans la majorité des cas la supériorité des invariants basés sur les moments de Zernike.

De façon étonnante, de notre point de vue, les études comparatives menées ne s'appuient que sur des résultats de classification et aucune sur la capacité de l'attribut à décrire correctement l'objet. Dombre, dans sa thèse[23] est un des premiers à discuter de cette notion et poser les limites d'approches "aveugles" face à l'individu à traiter. Ainsi, si les approches à partir de moments statistiques sont très utilisées en reconnaissance de formes et de caractères, nous savons que lesdits moments n'ont pas tous la même capacité discriminatoire (pour un même ordre

donné). Mais surtout, dans notre contexte, cette capacité varie en fonction de la forme de l'objet.

Lorsqu'au cours de ce travail nous avons dû choisir une définition pour la mesure de précision à associer à un opérateur d'extraction d'un attribut, nous avons décidé de prendre une formulation exploitant la préservation de l'information originale. Cette formulation est certes restrictive par rapport à d'autres formulations possibles mais elle présente surtout l'avantage d'être générique. Dés lors, comme nous le verrons, la mesure de précision estime la proportion d'information qui n'a pas été prise en compte lors du calcul de l'attribut. La question naturelle qui est posée à ce niveau correspond à la définition que nous donnons à l'information. Puisque ce traitement d'extraction de l'attribut est un traitement bas niveau (qui ne permet pas d'atteindre directement un niveau sémantique), l'information ici correspond aux données pixelliques.

Au niveau de l'opérateur lui même, les données recueillies ne permettent pas d'estimer une distance à l'objectif. Ici, clairement, la notion de chaîne de traitements est la seule qui permette d'établir une telle mesure en fin de la chaîne. Et comme nous l'avons vu au chapitre 3, la mesure alors fournie est un indice de confiance dans la décision.

Une critique supplémentaire à ce choix pourrait être apportée pour la généricité vers les opérateurs de filtrage, ou de débruitage. La proposition effectuée est là encore adaptée pour ces opérateurs, mais peut être pas la plus optimale. La plupart des opérateurs de restauration (par opposition aux méthodes de débruitage linéaires basées sur des filtres passe-bas, ou non linéaires) s'appuient sur des modèles de bruit qui peuvent donner lieu à des métriques prenant en compte cette spécificité. Mais ici aussi, selon nous, la qualité du prétraitement ne peut s'évaluer efficacement et qu'en fin de chaîne, dans la capacité apportée par ce prétraitement à améliorer l'indice de confiance dans la décision finale. Loin de nous

l'idée de rejeter toute la complexité dans l'étage de gestion de la chaîne de traitement d'images, nous considérons simplement que la mesure de précision associée à un opérateur ne peut travailler qu'au niveau des connaissances auquel elle intervient : pixellique pour le bas niveau, étiquette pour le niveau intermédiaire et décision finale (ou valeur) pour le haut niveau. Vu sous cet angle, il est également évident que l'étage de décision final induisant ou non le bouclage des traitements en vue de l'amélioration des résultats est un étage d'optimisation.

Dans ce chapitre, nous allons nous concentrer sur des descripteurs de formes simples et bien connus : les moments cartésiens, centrés, et ceux de Zernike. Nous les rappellerons brièvement ainsi que les invariants qui leur sont associés. Comme la mesure de précision que nous souhaitons établir est basée sur la quantité d'information prise en compte par l'attribut, nous avons choisi d'établir la différence entre l'image originale et l'image reconstruite à partir de l'information conservée. C'est pourquoi, lors de l'étude de chaque moment, nous nous intéresserons au calcul inverse. Au final, nous analyserons quelques résultats sur nos bases d'intérêt.

5.3 Moments et descripteurs

Dans le domaine de la reconnaissance de caractères, les moments statistiques sont largement utilisés ([3][4][84]). Les moments cartésiens, centrés et de Zernike peuvent décrire une forme en un vecteur multidimensionnel utilisable par un classifieur. Ils décrivent le contenu d'une image par rapport à leurs axes, et sont conçus pour capturer à la fois l'information géométrique globale et détaillée à propos des formes. Nous choisissons de travailler avec de tels moments car ils présentent la possibilité d'en calculer la transformée inverse.

5.3.1 Moments cartésiens

Le calcul des moments

En considérant une image comme une fonction de distribution de densité cartésienne $f(x,y)$ (fonction bidimensionnelle continue), Hu explique dans [37] et [38] que le $(p+q)^{\text{ième}}$ moment cartésien de dimension 2 est défini en tant qu'intégrale de Riemann comme :

$$m_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^p y^q f(x,y) dxdy \tag{5.1}$$

Sous les conditions décrites par Shutler[69], nous savons que :

Théorème 5.3.1. *Théorème d'unicité : la séquence de moment m_{pq} est définie de manière unique par $f(x,y)$ et inversement, $f(x,y)$ est définie de manière unique par m_{pq}.*

Ceci implique qu'une image peut être décrite et reconstruite si l'on utilise des moments d'un rang suffisamment élevé.

La version discrète de l'équation 5.1, pour une image constituée de pixels P_{xy}, donne :

$$m_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} x^p y^q P_{xy} \tag{5.2}$$

où m_{pq} est le moment bidimensionnel Cartésien, M et N les dimensions de l'image et $x^p y^q$ la fonction de base.
A partir du moment d'ordre m_{00} (la masse totale) défini par :

$$m_{00} = \sum_{x=1}^{M} \sum_{y=1}^{N} P_{xy} \tag{5.3}$$

on retrouve le centre de gravité de l'image, de coordonnées $(\hat{x},\hat{y})$ définies par :

$$\hat{x} = \frac{m_{10}}{m_{00}} \quad \hat{y} = \frac{m_{01}}{m_{00}} \tag{5.4}$$

Dans [38], Hu propose des moments centrés invariants à la translation par :

$$\mu_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} (x-\hat{x})^p (y-\hat{y})^q P_{xy} \tag{5.5}$$

et au changement d'échelle par :

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}{}^{\gamma}} \tag{5.6}$$

où

$$\gamma = \tfrac{p+q}{2} + 1 \quad \forall (p+q) \geq 2 \tag{5.7}$$

La reconstruction

En ce qui concerne la reconstruction, Shutler explique que si tous les moments M_{pq} (de l'ordre 0 à l'ordre N_{max}) d'une fonction $f(x,y)$ et d'ordre $N = (p+q)$ sont connus, il est possible d'obtenir la fonction continue $g(x,y)$ dont les moments correspondent à ceux de la fonction originale $f(x,y)$ jusqu'à l'ordre N_{max}.
En utilisant la décomposition en série de Taylor :

$$g(x,y) = g_{00} + g_{10}x + g_{01}y + g_{20}x^2 + g_{11}xy + \ldots + g_{pq}x^p y^q \tag{5.8}$$

réduite en :

$$g(x,y) = \textstyle\sum_{p=0} \sum_{q=0} g_{pq} x^p y^q \quad N_{max} = p+q \tag{5.9}$$

il faut calculer les coefficients constants g_{pq} de façon à ce que les moments de $g(x,y)$ correspondent à ceux de $f(x,y)$ en supposant que l'image soit une fonction continue bornée par $\{x,y\} \in [-1,1]$. Pour obtenir ces limites, il est possible de normaliser les valeurs des pixels sur celles des moments cartésiens :

$$\int_{-1}^{1} \int_{-1}^{1} g(x,y) x^p y^q dx dy \equiv M_{pq} \tag{5.10}$$

Ainsi, en substituant l'équation 5.8 dans 5.10 et en résolvant l'équation on obtient un ensemble d'équations linéaires dont le nombre est déterminé par l'ordre $(p+q)$

de la reconstruction. Il faut alors les résoudre pour les coefficients g_{pq} en inversant les matrices :

$$M_{pq} = \sum_i \sum_j g_{ij} \frac{1}{(i+p+1)(j+q+1)} (1-(-1)^{i+p+1})(1-(-1)^{j+q+1}) \quad \forall (i+j) \leq N \tag{5.11}$$

Il suffit enfin de réintégrer les coefficients g_{pq} dans 5.8 pour reconstruire une approximation de l'image originale.

La figure 5.2 illustre des exemples de reconstruction d'une image après décomposition en moments cartésiens à différents ordres.

Nous pouvons voir sur les images reconstruites la présence d'ondulations autour du caractère à reconnaître du fait des dépassements des intervalles sur les discontinuités. Ceci est dû à l'incapacité pour une fonction continue de recréer une fonction discrète, qu'importe combien de termes finis de grand ordre, un dépassement de la fonction se produit ; c'est le phénomène de Gibbs [71]. Le phénomène de Gibbs est un dépassement (ou **ringing**) des séries de Fourier et d'autres séries de fonction sur des simples discontinuités, qui augmente jusqu'à diluer totalement l'information utile (voir la reconstruction après décomposition à l'ordre 21).

Ceci peut être expliqué en termes de séries de Fourier :
Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction continue intégrable définie par parties, périodique de période $L > 0$. Supposons qu'à certains points x_0 les limites gauche $f(x_{0^+})$ et droite $f(x_{0^-})$ de la fonction f diffèrent d'une valeur non nulle de coupure a :$f(x_{0^+}) - f(x_{0^-}) = a/neq0$. Pour chaque entier $N \leq 1$, soit $S_N f$ la Nth série de Fourier partielle :

$$S_N f(x) := \sum_{-N \leq n \leq N} \hat{f}(n) e^{2\pi \sin x/L} = \frac{1}{2} a_0 + \sum_{n=1}^{N} a_n \cos(\frac{N\pi nx}{N}) + b_n \sin(\frac{2\pi nx}{N}) \tag{5.12}$$

où $\hat{f}$,a_n,b_n sont les coefficients de Fourier. Si x_n est une séquence de nombres réels

convergeant vers x_0 comme $N \to \infty$, et si la valeur de coupure a est positive alors :

$$\overset{+}{\lim_{N\to\infty}} S_n f(x_N) \leq f(x_0^+) + a.(0.089490...) \tag{5.13}$$

$$\overset{-}{\lim_{N\to\infty}} S_n f(x_N) \geq f(x_0^-) - a.(0.089490...) \tag{5.14}$$

Notons que si la valeur de coupure est négative, il faut échanger les limites supérieures et inférieures, ainsi que les termes d'inégalité.

La constatation principale que nous pouvons faire est que les méthodes de reconstruction d'images pour mesurer la quantité d'information conservée seront toujours biaisées par ce phénomène. Cependant, ceci peut être compensé, par exemple, grâce à l'application de méthodes de seuillage sur l'image reconstruite.

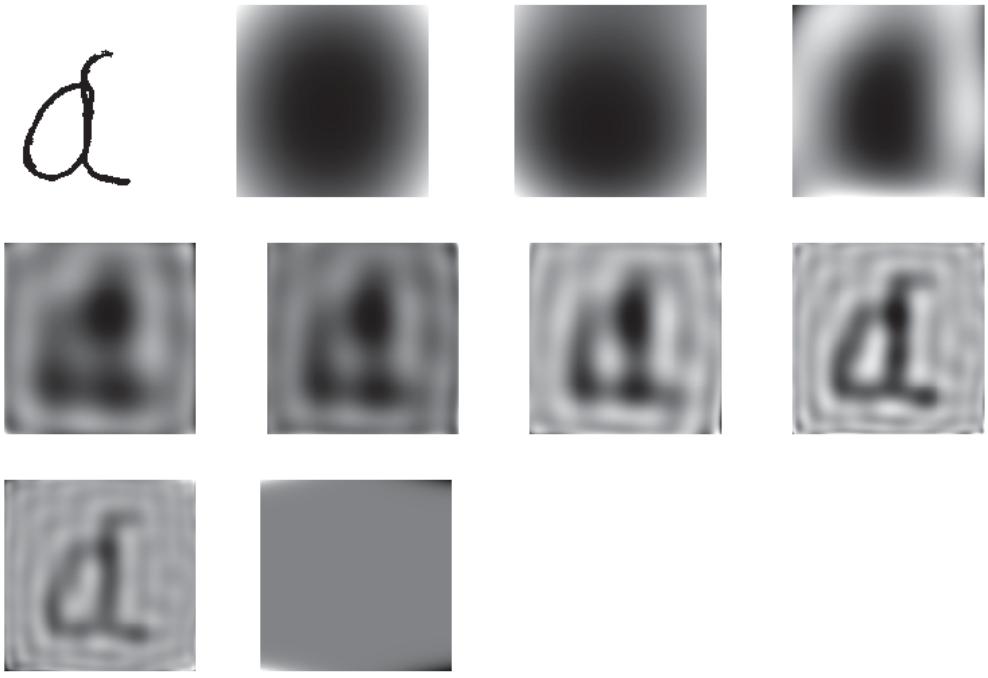

FIGURE 5.2 – Image originale et sa reconstruction avec les moments cartésiens à l'ordre 2, 3, 5, 9, 11, 13, 19 et 21

5.3.2 Les moments centrés

En gardant les même contraintes que pour les moments cartésiens, Shutler propose une extension aux moments centrés, la fonction $g(x,y)$ étant définie par :

$$g(x,y) = \sum_{p=0} \sum_{q=0} g_{pq}(x-\bar{x})^p (y-\bar{y})^q \quad N_{max} = p+q \tag{5.15}$$

ainsi 5.10 devient :

$$\int_{-1}^{1} \int_{-1}^{1} g(x,y)(x-\bar{x})^p (y-\bar{y})^q dxdy \equiv M_{pq} \tag{5.16}$$

De la même manière que nous avons calculé l'expression inverse des moments cartésiens, les moments centrés s'obtiennent à partir de :

$$M_{pq} = \sum_i \sum_j g_{ij} \frac{[(1-\bar{x})^{p+i+1} + (-1)^{p+i}(1+\bar{x})^{p+i+1}][(1-\bar{y})^{q+j+1} + (-1)^{q+i}(1+\bar{y})^{q+j+1}]}{(p+i+1)(q+j+1)} \tag{5.17}$$

Les moments centrés n'étant finalement qu'une extension des moments cartésiens permettant l'invariance à l'homothétie, les résultats obtenus après décomposition et reconstruction seront les mêmes, et il est possible de formuler les mêmes remarques (notamment le phénomène de Gibbs).

5.3.3 Moments de Zernike

Formulés à partir des travaux de Zernike ([92]) sous la forme complexe ([83]), les moments de Zernike A_{mn} peuvent être exprimés comme :

$$A_{mn} = \frac{m+1}{\pi} \sum_x \sum_y P_{xy}[V_{mn}(x,y)]^* \tag{5.18}$$

où $x^2 + y^2 \leq 1$, * dénote le complexe conjugué et $V(m,n)$ est le polynôme de Zernike exprimé en coordonnées polaires comme :

$$V_{mn}(r,\theta) = R_{mn}(r)\exp(jn\theta) \tag{5.19}$$

avec (r,θ) définis sur le disque unité, $j=\sqrt{-1}$ et

$$R_{mn}(r)=\sum_{s=0}^{\frac{m-|n|}{2}}(-1)^s F(m,n,s,r) \tag{5.20}$$

où

$$F(m,n,s,r)=\frac{(m-s)!}{s!(\frac{m+|n|}{2}-s)!(\frac{m-|n|}{2}-s)}r^{m-2s} \tag{5.21}$$

La reconstruction

De la même manière que nous l'avons vu précédemment, nous pouvons reconstruire la fonction originale avec les moments de Zernike. Shutler rapporte même une méthode plus rapide, basée sur la condition d'orthogonalité ([83]) qui considère que si tous les moments cartésiens d'une fonction $f(x,y)$ sont connus jusqu'à un ordre N_{max}, il est alors possible de reconstruire une fonction discrète $\hat{f}(x,y))$ dont les moments correspondent. Ce qui donne dans le cadre des moments de Zernike (exprimé par Khotanzad dans [42]) :

$$\hat{f}(r,\theta)=\sum_{m=0}^{N_{max}}\sum_{n}A_{mn}Vmn(r,\theta) \tag{5.22}$$

$m-|n|$ pair et $|n|\leq m$. Après développement nous obtenons :

$$\hat{f}(r,\theta)=\sum_{m=0}^{N_{max}}\sum_{n>0}(C_{mn}\cos n\theta+s_{mn}\sin n\theta)R_{mn}(r)+\frac{C_{m0}}{2}R_{m0}(r) \tag{5.23}$$

C_{mn} composant ainsi la partie réelle de A_{mn} comme :

$$C_{mn}=\frac{2m+2}{\pi}\sum_{x}\sum_{y}f(r,\theta)R_{mn}(r)\cos n\theta \tag{5.24}$$

et S_{mn} la partie imaginaire :

$$S_{mn}=\frac{-2m-2}{\pi}\sum_{x}\sum_{y}f(r,\theta)R_{mn}(r)\sin n\theta \tag{5.25}$$

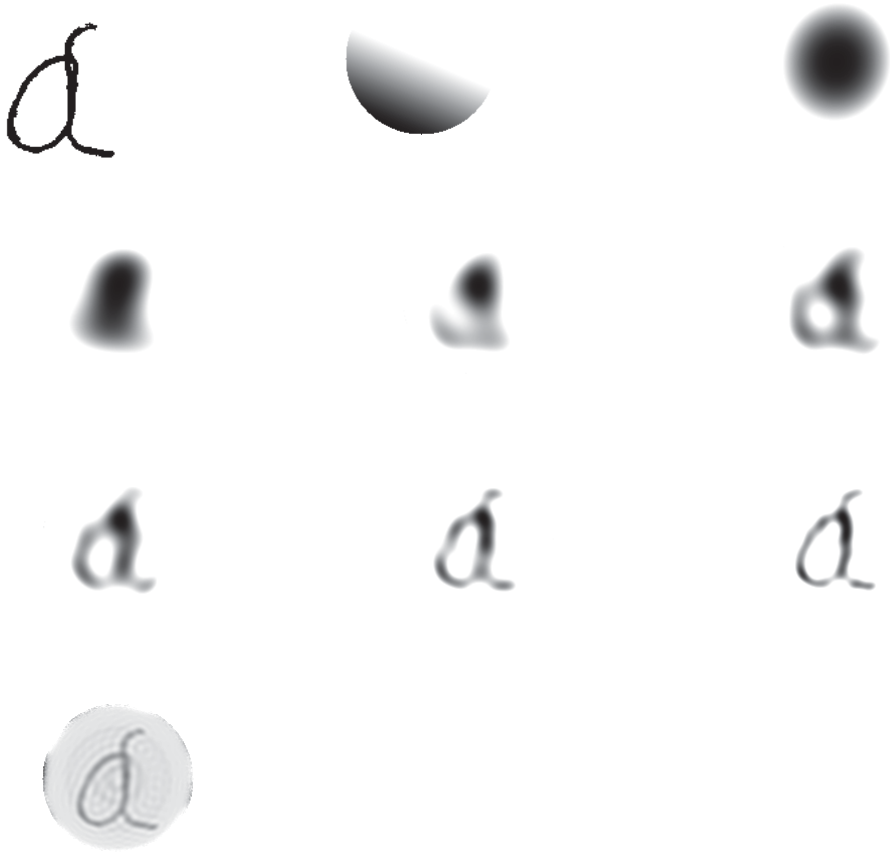

FIGURE 5.3 – Image originale et sa reconstruction avec les moments de Zernike à l'ordre 2, 3, 7, 9, 11, 13, 19, 25 et 49

Un exemple de reconstruction après décomposition par moments de Zernike à plusieurs ordre est illustré par la figure 5.3. Le demi disque obtenu à l'ordre 2 ainsi que le disque obtenu à l'ordre 3 sont liés au fait que les moments de Zernike s'expriment dans une base circulaire.

D'autre part, l'aspect "grisé" de la forme reconstruite à l'ordre 49 ainsi que les ondulations que nous voyons apparaître sont là encore liés au phénomène de Gibbs : l'information utile est noyée par le bruit généré aux discontinuité lors de la reconstruction.

Remarques sur les moments de Zernike

Certains auteurs utilisent les moments de Zernike pour de la reconnaissance de caractères binaires ([42], [43] et [3]) mais Trier[84] souligne la nécessité de développer des invariants à l'illumination car ces moments semblent adaptés aux images en niveaux de gris.

5.4 Invariants

Certaines familles de moments présentent l'intérêt d'être invariantes aux transformations géométriques. Les moments centrés sont invariants à la translation et au changement d'échelle, tandis que les moments de Zernike et de Fourier sont invariants à la translation, au changement d'échelle ainsi qu'à la rotation. Ces invariances sont fortement appréciées en reconnaissance de forme, facilitant l'apprentissage et la reconnaissance de caractères et symboles manuscrits. En effet, l'écriture humaine à ceci de particulier (ainsi que tous les objets naturels) qu'elle ne peut produire deux individus complètement identiques. Chaque individu possède une caractéristique différente des autres (inclinaison, taille, rotation etc.).

Or, statistiquement, la construction d'une classe basée sur un jeu d'apprentissage 100% naturel et sans aucune invariance produira de très forts écarts-types, les nuages de classes se chevauchant. Même si l'invariance aux transformations géométriques permet de corriger ce problème, il parait important pour nous de perdre des informations quant à ces transformations géométriques. Si notre hypothèse d'utilisation initiale est de considérer comme possible l'évolution de la définition ou de la résolution des images dans le temps, notre système doit pouvoir s'adapter au changement de contexte, sans remettre en cause la base d'apprentissage.

5.4.1 Les invariants dérivés des moments centrés

Les moments centrés μ_{pq} sont invariants à la translation car l'origine des moments est replacée au centre de gravité. Ils deviennent invariants au changement d'échelle en les normalisant :

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{5.26}$$

avec :

$$\gamma = \frac{p+q}{2} + 1 \quad \forall (p+q) \geq 2 \tag{5.27}$$

Ils nécessitent cependant d'être reformulés pour devenir invariants à la rotation. Hu[37][38] a dérivé les expressions des moments à partir des invariants algébriques appliqués à la fonction de génération de moments soumise à une rotation, ce qui donne un ensemble de moments non linéaires centrés. Le résultat, présenté

par Shutler[69], est un ensemble de moments orthogonaux invariants :

$$\iota_1 = \eta_{20} + \eta_{02} \tag{5.28}$$

$$\iota_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{5.29}$$

$$\iota_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{5.30}$$

$$\iota_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{5.31}$$

$$\iota_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \tag{5.32}$$

$$\iota_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})] \tag{5.33}$$

$$\iota_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \tag{5.34}$$

Ces moments sont d'ordre fini mais, contrairement aux moments centrés, ils ne proposent pas un jeu complet de descripteurs d'image. Il est cependant possible de dériver des moments d'ordre supérieur comme expliqué dans [38] et dans [3]. Li[53] propose au total 52 invariants d'ordres de 2 à 9, et Belkasim[3] en propose 32 (dont quelques uns de plus que Li) d'ordres de 2 à 7.

Reiss[61] a prouvé récemment que les moments proposé par Hu comme invariants à toute transformation linéaire étaient incorrects et proposé une nouvelle formulation. Par exemple, les invariants I_1 et I_2 sont calculés à partir des *invariants relatifs*[61][38] ; ils satisfont la condition :

$$I'_j = |A^T|^{\omega_j} |J|^{k_j} I_{j'} \tag{5.35}$$

où I_j est la fonction des moments dans l'espace (x,y) original, I'_j la même fonction calculée à partir des moments dans l'espace transformé (x',y'), ω_j est le poids de l'invariant relatif, |J| la valeur absolue la matrice Jacobienne de transformation transposée A^T et k_j est l'ordre de I_j. Pour générer des invariants absolus (i.e. satisfaisant $\psi'_j = \psi_j$), Reiss utilise $|A^T| = J$ et $\mu'_{00} = |J|\mu_{00}$ pour les transformations

linéaires :

$$I'_J = |J|^{\omega_j + k_j} I_J \qquad \text{pour } \omega_j \text{ pair} \tag{5.36}$$

$$I'_J = J|J|^{\omega_j + k_j - 1} I_J \qquad \text{pour } \omega_j \text{ impair} \tag{5.37}$$

pour prouver que $\psi_j = \frac{I_j}{\mu^{\omega_j + k_j}}$ est un invariant si ω_j est pair et $|\psi_j|$ est un invariant si ω_j est impair.

Pour les transformations linéaires générales, Hu et Reiss donnent les invariants relatifs suivants, fonctions de moments centrés d'ordre 2 et 3 :

$$I_1 = \mu_{20}\mu_{02} - \mu_{11}^2 \tag{5.38}$$

$$I_2 = (\mu_{30}\mu 03 - \mu_{21}\mu_{12})^2 - 4(\mu_{30}\mu_{12} - \mu_{21}^2)(\mu_{21}\mu_{03} - \mu_{12}^2) \tag{5.39}$$

$$I_3 = \mu_{20}(\mu_{21}\mu_{03} - \mu_{12}^2) - \mu_{11}(\mu_{30}\mu_{03} - \mu_{21}\mu_{12}) + \mu_{02}(\mu_{30}\mu_{12} - \mu_{21}^2) \tag{5.40}$$

$$\begin{aligned} I_4 = {} & \mu_{30}^2\mu_{02}^2 - 6\mu_{30}\mu_{21}\mu_{11}\mu_{02}^2 + 6\mu_{30}\mu_{12}\mu_{02}(2\mu_{11}^2 - \mu_{20}\mu_{02}) + \\ & \mu_{30}\mu_{03}(6\mu_{20}\mu_{11}\mu_{02} - 8\mu_{11}^3) + 9\mu_{21}^2\mu_{20}\mu_{02}^2 - 18\mu_{21}\mu_{12}\mu_{20}\mu_{11}\mu_{02} + \\ & 6\mu_{21}\mu_{03}\mu_{20}(2\mu_{11}^2 - \mu_{20}\mu_{02}) + 9\mu_{12}^2\mu_{20}^2\mu_{02} - 6\mu_{12}\mu_{03}\mu_{11}\mu_{20}^2 + \mu_{03}^2\mu_{20}^3 \end{aligned} \tag{5.41}$$

Reiss ayant trouvé les poids ω_j et les ordres k_j suivants :

$$\begin{array}{cccc} \omega_1 = 2 & \omega_2 = 6 & \omega_3 = 4 & \omega_4 = 6 \\ k_1 = 2 & k_2 = 4 & k_3 = 3 & k_4 = 5 \end{array}$$

les coefficients $\psi_1 = \frac{I_1}{\mu_{00}^4}$, $\psi_2 = \frac{I_2}{\mu_{00}^1 0}$, $\psi_3 = \frac{I_3}{\mu_{00}^7}$, et $\psi_4 = \frac{I_4}{\mu_{00}^1 1}$, sont invariants en translation ainsi qu'à toute transformation linéaire. Reiss apporte aussi des invariants aux changements de contraste (en plus d'être déjà invariants en translation ainsi qu'aux transformations linéaires), les trois premiers étant $\theta_1 = \frac{I_4}{\mu_{00} I_2}$, $\theta_2 = \frac{I_1^2}{\mu_{00} I_3}$, $\theta_3 = \frac{I_1 I_3}{I_4}$.

5.4.2 Les invariants dérivés des moments de Zernike

La grandeur $|A_{mn}|$ est invariante à la rotation. Trier montre que les ordres 1 et 2 représentent l'orientation, la hauteur et la largeur, et qu'il faut monter jus-

qu'aux ordres 8 à 11 pour retrouver une apparence raisonnable à la reconstruction. L'invariance à la translation et au changement d'échelle peut-être obtenue en décalant et en normalisant la taille de l'image avant le calcul des moments. Le moment d'ordre 1 peut-être utilisé pour retrouver le centre de l'image, et le moment d'ordre 0 donne une estimation de la taille. Belkasim propose aussi d'utiliser :

$$B_{n,n+1} = |A_{n-2,1}||A_{n1}|\cos(\phi_{n-2,1} - \phi_{n1}) \tag{5.42}$$

$$B_{n,n+L} = |A_{n1}||A_{nL}|^p \cos(p\phi_{nL} - \phi_{n1}) \tag{5.43}$$

où $L = 3,5,\ldots,n$, $p = 1/L$ et ϕ_{nm} est composant de la phase de A_{mn} tel que :

$$A_{mn} = |A_{mn}|\cos\phi_{mn} + j|A_{mn}|\sin\phi_{mn} \tag{5.44}$$

Une analyse de l'écriture formelle de ces invariants nous permet de conclure qu'ils sont finalement des attributs issus des moments statistiques : à partir des moments cartésiens, centrés ou de Zernike, il est possible d'établir une nouvelle classe d'attributs qui résume ou extrait l'information utile des moments. Cette information nous ne pouvons malheureusement pas l'utiliser directement pour mesurer la quantité d'information conservée par rapport à l'objet original. En effet, il n'existe pas d'écriture inverse de ces invariants, et quand bien même, cela nous ramènerait au moments originaux que nous avons déjà, et dont nous connaissons la transformée inverse.

Leur popularité et leur efficacité, compte tenu des résultats que nous trouvons dans la littérature, en font toutefois des candidats idéaux comme support d'information quant aux transformations géométriques qu'a subies l'objet à reconnaître par rapport à l'objet de référence.

5.5 Mesure de la précision

La mesure que nous cherchons à estimer correspond à la quantité d'information qui est prise en compte par l'attribut. Cette mesure correspond au pourcentage

d'information conservé par l'attribut par rapport à l'information complète. Dans notre cas, cette mesure s'obtient par le biais d'une mesure de distance entre l'objet original et l'objet reconstruit. Toute la difficulté provient de l'application de la mesure de distance à utiliser qui diffère selon la nature de l'information extraite par l'opérateur de calcul de l'attribut. Dans le cas d'un opérateur de quantification couleur, la distance s'appuiera forcément sur la perte de complexité obtenue, estimée par une somme de différences couleur établies dans un espace couleur adapté (ΔE ou ΔE_{2000} dans Lab par exemple). Les différences étant calculées entre la couleur du pixel de l'image originale $P_1(x,y)$ et le pixel de l'image quantifiée $P_2(x,y)$ et la somme conduite sur toutes les positions (x,y) de l'image.

Pour l'exemple qui nous intéresse, nous nous attachons à des objets binaires de couleur noire (valeur nulle) sur fond blanc. L'objet reconstruit à partir des moments que nous utilisons n'est en revanche pas binaire, mais en niveaux de gris à cause de la formulation héritée du domaine continu. La binarisation de l'objet est donc un étage supplémentaire à prévoir avant le calcul de la distance entre objet reconstruit et original.

5.5.1 Quelle fonction de distance ?

Soient deux images binaires P_1 et P_2 définies sur un même support spatial de taille $N \times M$ pixels avec $P_1(x,y)$ et $P2(x,y)$ nuls pour tous les pixels de l'objet représenté sur l'image originale (par voie de conséquence, $P_1(x,y) = 1$ pour tous les pixels du fond).

Les premières fonctions de distance auxquelles nous pourrions penser sont les métriques de type Minkowski :

- $L_1 = \sum_x \sum_y |P_1(x,y) - P_2(x,y)|$
- $L_2 = \sqrt{\sum_x \sum_y (P_1(x,y) - P_2(x,y))^2}$

– $L_\infty = \max_{x,y}(|P_1(x,y) - P_2(x,y)|)$

les métriques en L_2 étant notamment bien adaptées dans notre cas où $P_1(x,y)$ et $P_1(x,y)$ sont binaires.

La formulation à retenir pour une estimation de la précision doit être normalisée entre 0 et 1, comme nous l'a imposé l'exploitation en tant que coefficient d'affaiblissement dans le modèle des croyances transférables (chapitre 3).

La première mesure de précision proposée est alors :

$$d_P(P_1,P_2) = 1 - \frac{1}{N.M}\sum_{x=1}^{N}\sum_{y=1}^{M}|P_1(x,y) - P_2(x,y)| \tag{5.45}$$

Cette métrique apparemment simple peut masquer les difficultés posées par le choix d'une bonne formulation. Bien évidement, nous en avons formulé plusieurs et recherché différentes propositions dans la littérature. Nous pourrions, par exemple, dans le cadre d'objets binaires, proposer une expression du type :

$$\begin{cases} d_B(P_1,P_2) = \frac{1}{k_0}\sum_{x=1}^{N}\sum_{y=1}^{M}\overline{P_1(x,y)}.\overline{P_2(x,y)} \\ k_0 = \arg S_i, S_i = \{(x,y), P(x,y) = 0\} \end{cases} \tag{5.46}$$

où $\overline{X}$ correspond à l'inverse booléen de X et k_0 le nombre de pixels de P_1 qui sont nuls donc appartenant à l'objet représenté.

Cette formulation est séduisante car uniquement centrée sur l'objet binaire original et reconstruit. La normalisation par le nombre de pixels utiles de l'objet représenté dans l'image originale lui confère en plus une meilleure sensibilité aux évolutions des opérateurs en amont. Cependant, si nous posons complètement le problème en considérant que l'image originale P_1 est composée de k_0 pixels de valeur nulle (donc $(N.M) - k_0$ pixels de valeurs non nulle) et que l'image reconstruite est composée de l_0 pixels de valeur nulle et l_1 pixels de valeur non nulle ($l_1 = (N.M - k_1)$). Les différents cas de figure possibles sont représentés dans le tableau 5.5.1.

	l_0	l_1
k_0	l_{00}	l_{10}
$k_1 = (N.M) - k_0$	l_{01}	l_{11}

TABLE 5.1 –

Ce tableau permet de comparer les quantités l_{00} et l_{11}, nombre de pixels de l'image reconstruite présentant la bonne valeur avec les erreurs de reconstruction l_{10} et l_{01}.

Dans le cas de la métrique issue de la distance de Minkowski d'ordre 1 :

$$d_P(P_1,P_2) = 1 - \frac{1}{NM}\sum\sum|P_1(x,y) - P_2(x,y)| = 1 - \frac{l_{10}+l_{01}}{NM} \tag{5.47}$$

Si nous analysons différents cas de figure :

– Une reconstruction parfaite ($l_{00} = k_0$ et $l_{11} = k_1$) :
$l_{10} = l_{01} = 0 \Rightarrow d_P(P_1,P_2) = 1$,

– Une reconstruction inversée ($l_{00} = l_{11} = 0$) :
$l_{01} + l_{11} = NM \Rightarrow d_P(P_1,P_2) = 0$,

– Le pire cas d'une reconstruction d'une image blanche ($l_{00} = l_{01} = 0$) :
$d_P(P_1,P_2) = \frac{l_{10}}{NM}$,

– Le pire cas d'une reconstruction d'une image noire ($l_{10} = l_{11} = 0$) :
$d_P(P_1,P_2) = \frac{l_{01}}{NM}$,

la dynamique de la mesure est comme évoquée précédemment fonction des ratios $1/NM$ dépendant de la taille des images considérées.

Dans le cas de la seconde métrique, l'expression de la formule pour le cas proposé devient :

$$d_B(P_1,P_2) = \frac{l_{00}}{k}, \tag{5.48}$$

ce qui donne, en reprenant les mêmes cas de figure que précédemment :

– Une reconstruction parfaite ($l_{00} = k$ et $l_{11} = k$) :

$d_B(P_1,P_2) = \frac{l_{00}}{k} = \frac{k}{k} = 1$,

– Une reconstruction inversée ($l_{00} = l_{11} = 0$) :
$d_B(P_1,P_2) = \frac{0}{k} = 0$,

– Le pire cas d'une reconstruction d'une image blanche ($l_0 = 0$) :
$d_B(P_1,P_2) = \frac{0}{k} = 0$,

– Le pire cas d'une reconstruction d'une image noire ($l_1 = 0$) :
$d_B(P_1,P_2) = \frac{l_{00}}{k} = \frac{k}{k} = 1$.

La dynamique de la mesure est ici supérieure à la première puisqu'égale à $\frac{1}{k}$ avec $k < NM$ (par définition l'objet n'est pas une image noire).

Cependant, comme le montre la séquence de cas, la mesure d_B va avoir tendance à surévaluer la mesure lorsque l'image reconstruite va tendre vers un objet uniformément noir.

Au final, malgré une apparente simplicité de la formulation, l'expression $d_P(P_1,P_2)$ prend en compte tous les paramètres nécessaires : la partie de l'objet similaire et la partie du fond similaire entre les deux objets, et au final, la dynamique plus faible n'est due qu'à une prise en considération de plus d'informations dans la double somme.

5.5.2 La binarisation de l'image reconstruite

Les images reconstruites à partir des différents moments ne sont pas binaires mais continues discrétisées. Pour appliquer les mesures de distance il convient auparavant de binariser les images reconstruites. Comme nous avons pu le constater dans la première partie de ce chapitre, le niveau d'intensité moyen de l'image reconstruite peut varier de façon importante (figure 5.2 et 5.3), l'utilisation d'un seuil fixé *a priori* n'est donc pas envisageable.

Cette question est pour nous une des premières applications de la mesure de la précision locale pour déterminer le seuil de binarisation idéal. Nous avons choisi de boucler selon ce critère de précision un système cherchant le seuil idéal de binarisation (voir figure 5.4).

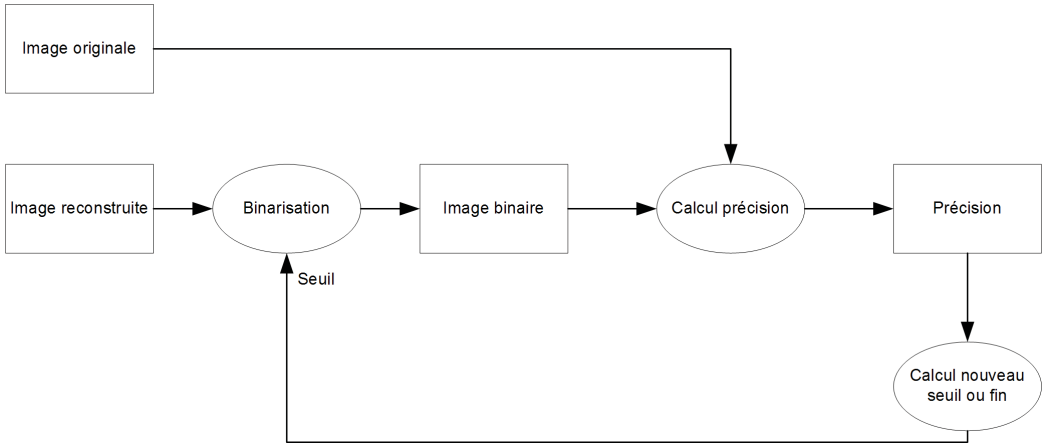

FIGURE 5.4 – Processus de calcul de seuil adaptatif

Ce schéma classique de système en boucle fermée laisse libre court au choix de l'algorithme d'établissement du nouveau seuil (tracking, dichotomie etc.). Les résultats qui sont présentés par la suite correspondent à un algorithme de poursuite simple. Ils montrent selon toute vraisemblance un accroissement de la précision jusqu'à un seuil idéal, puis une décroissance une fois ce seuil dépassé.

Les figures 5.5, 5.6 et 5.7 représentent chacune trois exemples de l'application de cette technique de seuillage. On retrouve en haut l'individu original et sa reconstruction, et en bas la reconstruction après seuillage accompagnée de l'évolution de la précision en fonction de la valeur du seuil. Nous avons fait le choix de prendre à chaque fois deux individus très différents au niveau de leur forme et de leur complexité pour illustrer la différence de comportement de la précision. Chaque figure est associée à une famille de moments, respectivement les moments cartésiens, les moments centrés et les moments de Zernike.

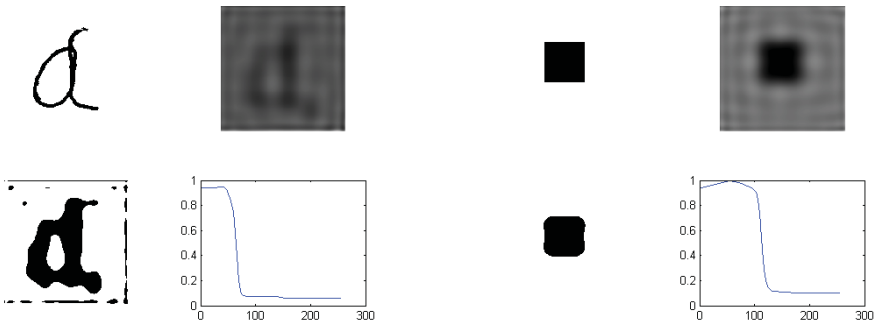

FIGURE 5.5 – Original, reconstruction, binarisation et évolution de la précision en fonction du seuil - Moments cartésiens à l'ordre 15

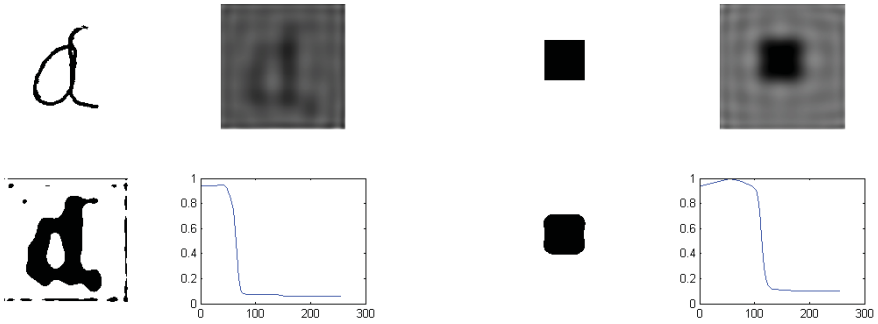

FIGURE 5.6 – Original, reconstruction, binarisation et évolution de la précision en fonction du seuil - Moments centrés à l'ordre 15

Nous constatons en premier lieu sur ces figures, et ce résultat était attendu selon la théorie, que les images obtenues après reconstruction sont strictement identiques que l'on utilise les moments cartésiens ou les moments centrés. Ce phénomène est dû au fait que les bases de décomposition sont strictement identiques, la seule différence étant que les moments centrés sont une réécriture des moments Cartésiens prenant en compte les transformations géométriques de type translation. Il est donc en effet logique, pour une même image donnée, d'obtenir strictement la même reconstruction.

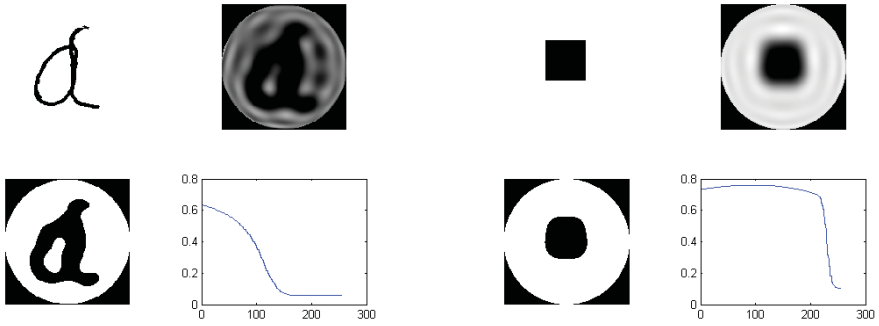

FIGURE 5.7 – Original, reconstruction, binarisation et évolution de la précision en fonction du seuil - Moments de Zernike à l'ordre 15

Notons également que, comme attendu, la précision évolue effectivement en même temps que la valeur du seuil, jusqu'à un optimal au delà duquel elle décroît. Il existe cependant des différences dans l'évolution de la précision. D'une part, avec les moments cartésiens (et centrés), l'optimum est atteint plus rapidement pour le caractère alpha que pour le rectangle noir. L'information utile (i.e. les pixels décrivant la forme) étant mieux répartie pour un rectangle plein que pour un caractère au trait fin et de forme complexe, elle subit de fait moins fortement l'influence du phénomène de Gibbs. Nous constatons d'ailleurs très bien sur l'image reconstituée que le rectangle semble plus compact sur un fond plus clair.

Notons d'autre part l'évolution de la précision au delà de l'optimum. Celle-ci chute brutalement avec les moments cartésiens (et centrés), tandis que la décroissance est plus douce avec les moments de Zernike. Ceci semble indiquer une meilleure conservation de l'information pour cette famille de moments, ainsi qu'une plus faible sensibilité au bruit induit par le phénomène de Gibbs.

Cependant, en s'attardant sur la valeur des optimums, en fonction de la forme ou de la famille de moments utilisés, nous nous apercevons que ceux obtenus à l'aide des moments cartésiens (ou centrés) sont meilleurs que ceux obtenus par les moments de Zernike. Cela peut être dû à une trop forte sensibilité de la méthode de

calcul de la précision (et du seuillage) au bruit généré par la reconstruction. Une meilleure conservation de l'information originale par les moments Cartésiens (ou centrés) peut également en être à l'origine. Trier[84] et Shutler[69] indiquent bien que chaque famille de moments induits des déformations géométriques : les moments de Zernike ayant tendance à favoriser les arrondis (du fait de leur écriture complexe) à l'inverse des moments Cartésiens qui eux favorisent les formes angulaires.

5.5.3 Protocole

Dans l'optique de vérifier si la quantité d'information utile en fonction de la taille du symbole est importante, un changement d'échelle est appliqué sur chaque image de 1 à 3 (avec un pas de 0.1). A cause du besoin d'une taille normalisée pour cette opération, nous ne pouvons utiliser que la base MNIST pour cette partie. Tous les moments statistiques sont calculés de l'ordre 1 à 35, afin de prendre en compte la dilution de l'information utile dans l'information générée par la reconstruction. Nous utilisons également un algorithme de dilatation pour obtenir quatre nouvelles images (1x, 2x, 3x and 4x) afin de procéder à un changement d'épaisseur de trait en compensation de la différence d'épaisseur entre les symboles manuscrits et ceux générés artificiellement.

Notons enfin la particularité des moments de Zernike : l'analyse et la reconstruction, ces moments s'inscrivant dans une base complexe, se font dans un cercle ; il existe alors une zone qui n'est jamais traitée, et visible en noir sur les figures vues précédemment. Afin de compenser ce phénomène, toute image analysée est d'abord recopiée au centre d'une image plus grande, de façon à ce que l'image originale soit inscrite dans le cercle d'analyse. Ensuite, pour la comparaison des données, la mesure de la précision ne se fera que dans une fenêtre centrée, de la taille de l'image d'origine.

5.6 Bilan et synthèse des différents résultats

Toutes les figures représentent l'évolution de la précision pour les 3 familles de moments, en fonction de l'ordre du moment. La figure 5.8 concerne les caractères manuscrits de notre propre base et les reproductions manuscrites des symboles de la base GREC'05 à gauche. Sur la droite, nous trouvons la précision obtenue en utilisant les symboles de la base de test GREC'05 et les caractères grecs générés en utilisant un logiciel de traitement de texte. La figure 5.9 a été générée en utilisant tous les caractères et symboles manuscrits ainsi que les images obtenues en utilisant les différentes dilatations, et la figure 5.14 a été générée à partir de la base MNIST et de différents niveaux de zoom, montrant l'impact du changement d'échelle.

La première observation est que les moments cartésiens et centrés donnent toujours sensiblement les mêmes niveaux de précision, ce qui est un comportement logique étant donné que la différence entre les deux écritures, comme nous l'avons vu plus haut, ne réside que dans une translation qui présente une meilleure invariance mais qui n'influe pas sur l'information conservée. De plus, les objets analysés ici sont à peu près centrés dans l'imagette qui les représente, ce qui rend caduc l'intérêt des moments centrés.

5.6.1 Influence de l'aspect manuscrit

Entre les parties gauche et droite de la figure 5.8, la différence est la précision maximale pour les moments de Zernike, cartésiens et centrés. Pour les caractères générés et les symboles, les moments non orthogonaux préservent plus d'information que les moments de Zernike, et donnent la possibilité de recréer quasi

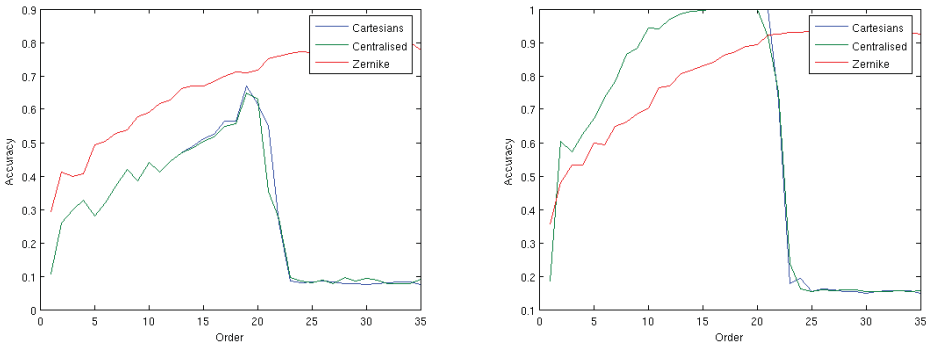

FIGURE 5.8 – Évolution de la précision en fonction de l'ordre pour les caractères et symboles manuscrits (à gauche) et générés (à droite)

parfaitement l'image originale en utilisant un ordre entre 15 et 20. Pour les documents manuscrits, les moments de Zernike donnent les meilleurs résultats. Ceci peut être expliqué d'une part par l'épaisseur du trait (les caractères générés sont plus épais que les manuscrits), et d'autre part par le fait que les symboles générés sont plus "angulaires" que les manuscrits. En effet, et comme l'expliquent Trier[84] et Shutler[69] du fait de leur base orthogonale et de leur forme complexe, les moments de Zernike semblent plus adaptés aux formes arrondies, telles que les caractères et les symboles manuscrits.

Afin de valider notre hypothèse sur l'épaisseur du trait, nous avons observé l'évolution de la précision en fonction de l'ordre des moments, pour les images d'origines dilatées plusieurs fois.

5.6.2 Influence de l'épaisseur de trait initial

Observons en premier lieu que pour la figure 5.9, la précision maximale obtenue avec les moments de Zernike n'augmente pas avec l'épaisseur du trait.

En ce qui concerne les moments orthogonaux (moments cartésiens et moments centrés), nous pouvons voir que plus le trait est épais, meilleure est la précision

(inférieure à 0.9 sans dilatation, supérieure à 0.9 après dilatation x4).

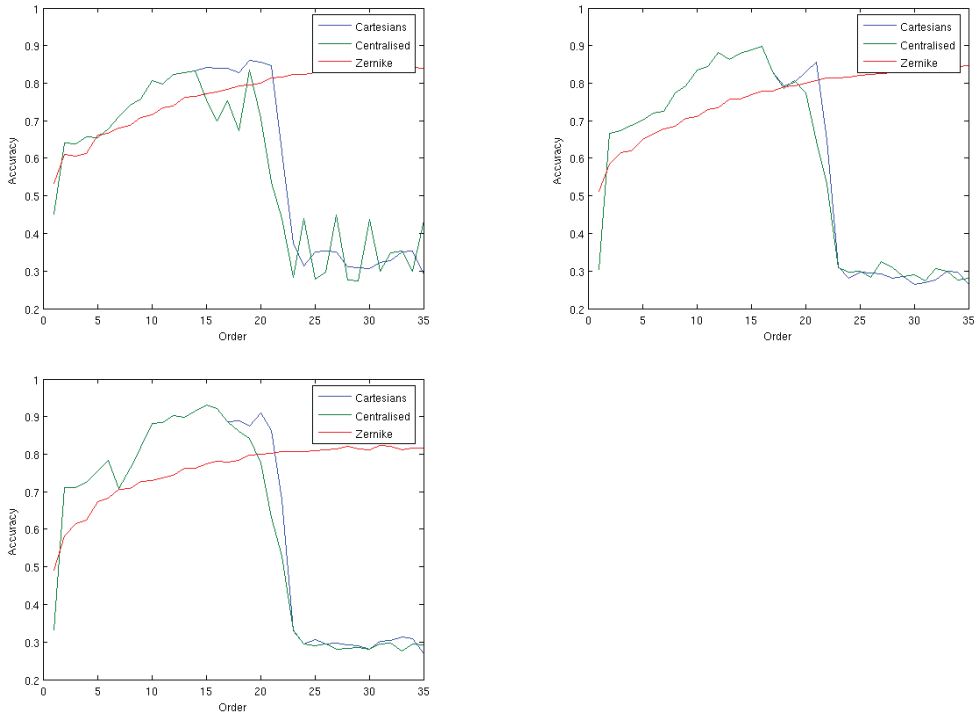

FIGURE 5.9 – Évolution de la précision pour différents niveaux de dilatation (1x, 3x, 4x)

Notons sur les figures 5.10, et 5.13 que les caractères et symboles manuscrits sont réalisés avec de feutres ou stylos à pointe relativement fine (trait de contour fin). Notons aussi, sur les figures 5.11, 5.12, que les traits des caractères et symboles générés artificiellement sont eux plus épais.

En commentaire de la figure 5.8 nous émettions l'hypothèse que les différences de précisions obtenues pour les caractères et symboles manuscrits, par rapport aux caractères dits "artificiels", pouvaient s'expliquer par l'influence de

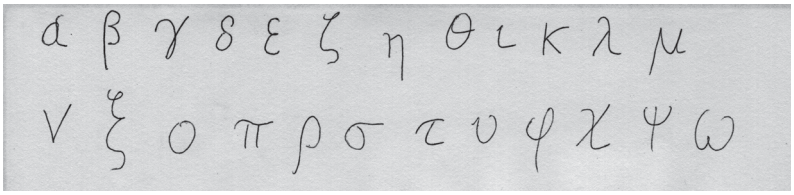

FIGURE 5.10 – Extrait de la base RCSOFT - caractères manuscrits

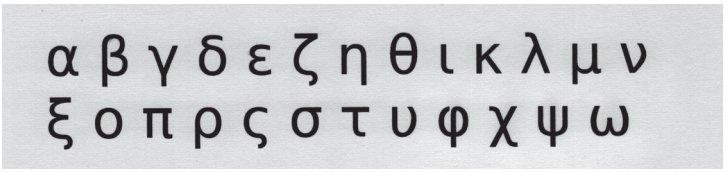

FIGURE 5.11 – Extrait de la base RCSOFT - caractères artificiels

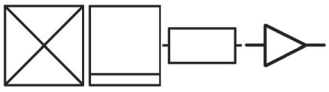

FIGURE 5.12 – Extrait de la base GREC'05

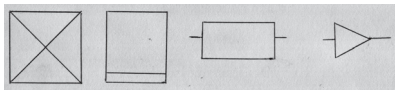

FIGURE 5.13 – Extrait de la base GREC/RCSOFT - symboles manuscrits

l'épaisseur du trait. Les résultats que nous venons d'exposer renforcent cette hypothèse : l'information utile sera d'autant mieux conservée par les moments cartésiens (et centrés) que le trait du caractère sera épais, alors que les moments de Zernike ne semblent pas sensibles à ce paramètre.

5.6.3 Influence du changement d'échelle

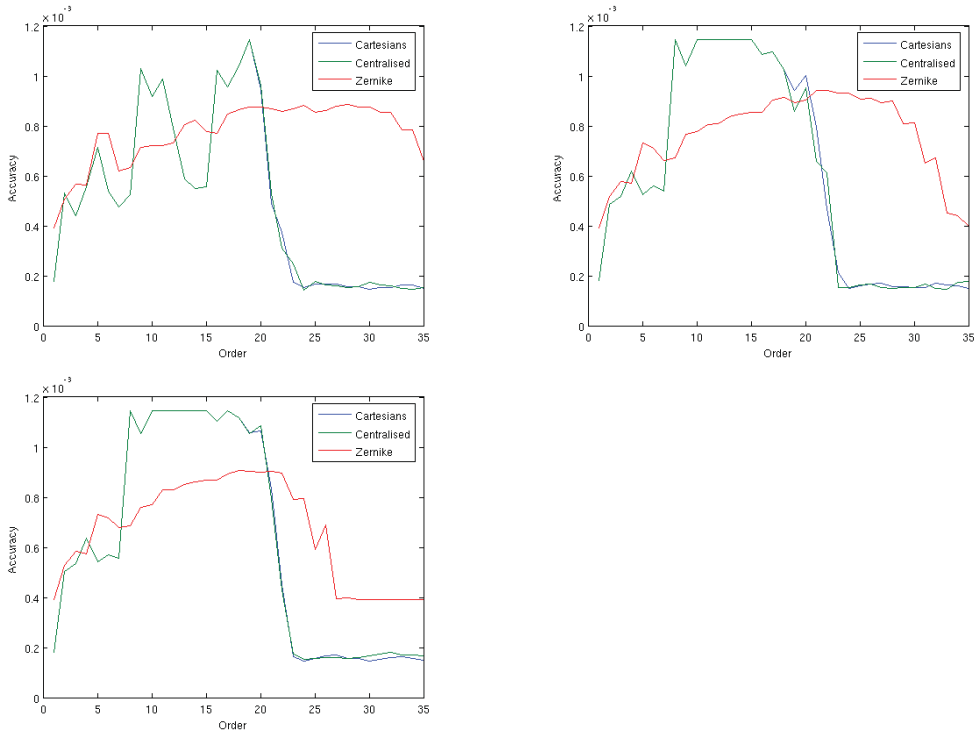

FIGURE 5.14 – Évolution de la précision pour différents facteurs d'échelle (x1, x2, x3)

Sur la figure 5.14, constatons d'abord que pour les résultats obtenus à l'échelle 1, des différences de comportement notable apparaissent entre les moments cartésiens et les moments centrés. Ce phénomène s'explique par le fait que les images

de la base MNIST sont petites (20x20 pixels) ; l'information utile (les pixels appartenant à la forme d'origine) se retrouve rapidement diluée dans le bruit généré par la reconstruction (interpolations, phénomène de Gibbs, seuillage etc.). Même si en terme de coût de calcul, il peut sembler intéressant de travailler sur des images de la plus petite taille possible, nous voyons très bien qu'une certaine stabilité de l'évolution de la précision en fonction de l'ordre s'installe à partir d'un changement d'échelle de 2.

Observons aussi la courbe en dents de scie de la précision pour les petites échelles, qui se stabilise ensuite pour des échelles plus élevées. Notons également une variation de l'ordre maximal avant la chute de la précision pour les moments de Zernike. Ceci est attendu du fait que l'algorithme d'échelle génère de l'information inutile. Du fait de la nature corrélée des moments non orthogonaux, chaque moment ne porte pas que sa propre contribution à l'information individuelle. Ceci dissout l'information utile pour les ordres de moment les plus grands. Même si accroître le nombre de pixels utiles n'améliore pas significativement la précision, cela contribue tout de même à en stabiliser l'évolution.

Finalement, nous pouvons voir sur toutes les figures que la précision augmente jusqu'à chuter fortement, pour toutes les familles de moments. Ceci est dû, comme nous l'avons vu précédemment, au phénomène de Gibbs et au fait que les transformées inverses et la reconstruction se fassent avec des fonctions continues pour estimer des fonctions discrètes : un dépassement de la fonction fini par se produire.

5.7 Conclusion

La sensibilité à l'ordre, épaisseur et taille peut être expliquée par la dilution de l'information utile dans l'information totale (partiellement expliquée par la malédiction de la dimension), et par le phénomène de Gibbs. La théorie le prévoie et

nous le voyons clairement dans nos résultats. Nous ne conseillons d'ailleurs pas de changer l'échelle de l'image à cause du bruit induit par ce genre de méthode.

Les moments centrés et cartésiens semblent être les méthodes les plus précises mais elles sont également les plus sensibles aux transformations géométriques, et c'est un réel problème concernant l'écriture manuscrite et les documents techniques complexes. Ces moments demandent cependant un très faible coût de calcul comparé aux moments de Zernike. D'une part, sachant l'ordre optimal (entre 10 et 22), nous supposons qu'il est possible de proposer un traitement de reconnaissance utilisant à la fois les moments statistiques et de l'information exogène (par exemple, de l'information concernant les transformations géométriques obtenues grâce aux différents invariants). D'autre part, nous retrouvons ici tout l'intérêt de la chaîne de traitement et de ses bouclages, permettant de combiner les différents types de vecteurs de moments d'ordre le plus faible possible pour une précision maximale et une sensibilité minimale aux transformations géométriques. Dans tous les cas, le choix se faisant selon la connaissance de l'évolution de la précision, impliquant un processus d'apprentissage en amont.

A propos des différences de comportement de la précision entre les caractères manuscrits et ceux générés, il ne doit pas exister d'influence de la géométrie des symboles (comme l'angularité), seule l'épaisseur compte. Dans ce cas, il doit être possible d'ajouter de l'information sur l'individu à reconnaître (pour une faible coût de calcul) et de décider d'utiliser une dilatation si c'est un symbole manuscrit.

Il pourrait être finalement intéressant d'analyser les effets de la précision si nous utilisons ensuite une étape de classification après le calcul des moments et de trouver une relation entre la quantité d'information préservée et les performances de reconnaissance. D'autres méthodes de similarité doivent également être testées, comme d'autres métriques de distances, de comparaison entre les formes

d'origine et celle reconstituée (c'est à dire une analyse au niveau de la forme et non pas une au niveau des pixels) ou alors en se basant sur des notions comme l'entropie de l'image.

Il est ainsi possible de choisir le descripteur de forme le plus approprié, en prenant en compte la précision, selon une connaissance *a priori* du document en cours d'analyse.

"*C'est le commencement qui est le pire, puis le milieu puis la fin ; à la fin, c'est la fin qui est le pire.*"
Samuel Beckett

Chapitre 6

Conclusion

6.1 Synthèse

Nous avons vu en introduction à ce document que les travaux présentés ici reposent sur l'hypothèse selon laquelle toute application de traitement d'images peut être représentée comme une chaîne de traitements simples, comme l'illustre la figure 6.1.

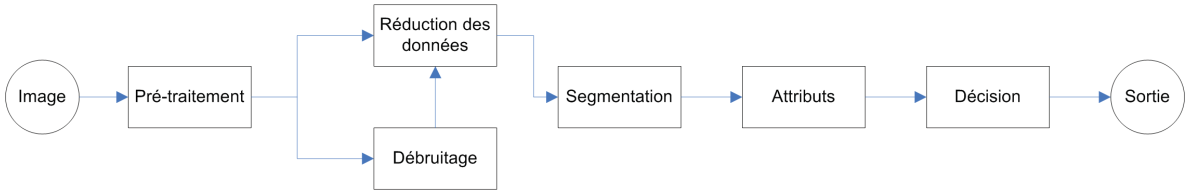

FIGURE 6.1 – Une application de traitement d'images vue sous forme de chaîne de traitements

Afin de gérer à la fois la combinaison de plusieurs opérateurs et les "essais" du système avec différents opérateurs et/ou paramètres, nous proposons la construction d'une chaîne non linéaire, avec bouclage, renvoyant aux travaux de Fayyad[27][26] sur la fouille de données et l'extraction de la connaissance dans les bases de données.

Les décisions du système quant aux opérateurs à exécuter ou au réglage des paramètres se basent sur les outils statistiques apportées par la théorie de l'évidence et particulièrement le modèle des croyances transférables. Nous intégrons dans ces outils une mesure de la qualité (connue *a priori*, ou estimée *a posteriori*) des traitements que nous appelons "précision". Cette précision est aussi utilisée pour choisir dynamiquement les outils de classification en se basant sur les travaux de Giacinto[32][31] sur la sélection dynamique des classifieurs. Tout ce processus est représenté par la figure 6.2.

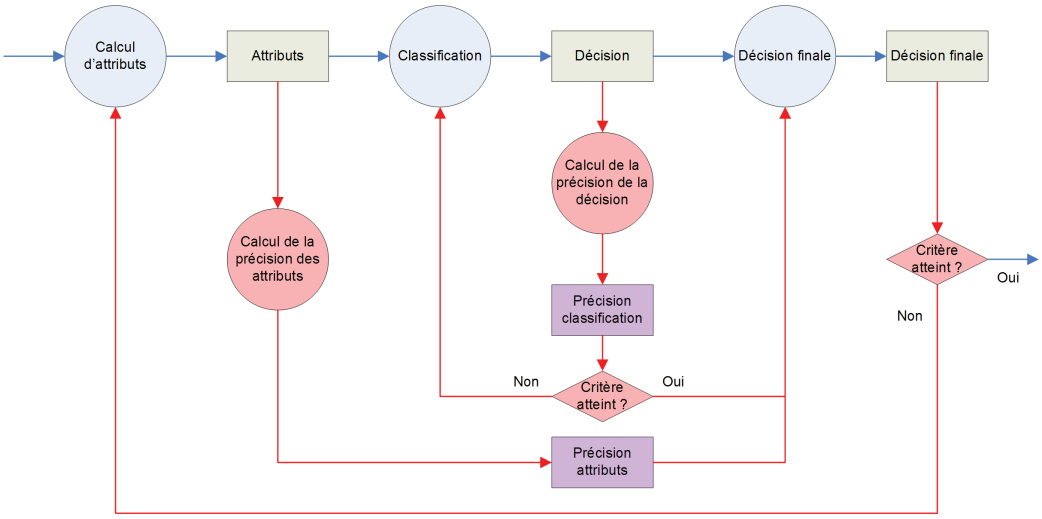

FIGURE 6.2 – Chaîne de traitement intégrant la précision des attributs et de la classification

6.2 Chaîne de traitement d'image : Théorie

Nous avons présenté dans le chapitre 3, les résultats obtenus avec une chaîne de traitements complète, exécutée avec bouclage sur la base de caractères MNIST. Le mode de fonctionnement de cette chaîne, a été réglé de manière à favoriser le temps de calcul au détriment de la précision. Pour cela, notre chaîne exécute une première passe avec une première famille de moments les moins coûteux en temps de calcul, puis réinjecte les individus rejetés pour une deuxième passe avec cette fois une famille de moments plus complexe, et ainsi de suite, jusqu'à ce qu'il n'y ait plus de rejets, ou plus de méthodes de calcul d'attributs disponibles. Le taux d'erreur global obtenu ainsi est présenté en figure 6.3 et le taux de rejet en figure 6.4.

Comme le réglage choisi impose en premier lieu le choix d'une méthode de calcul d'attributs peu coûteuse (mais moins précise), le taux d'erreur global ne peut pas être supérieur à celui obtenu avec la méthode choisie pour la première passe de la chaîne, les individus mal classés n'étant pas rejetés, ils ne sont pas ana-

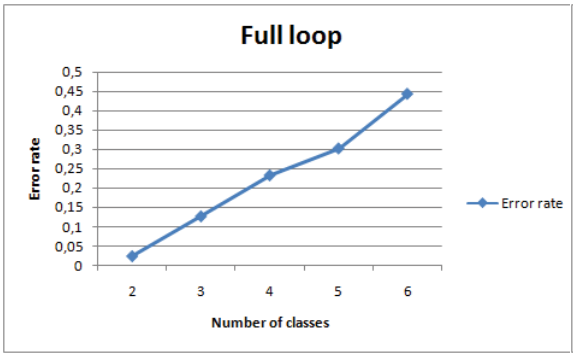

FIGURE 6.3 – Taux d'erreur pour une boucle à 3 étapes

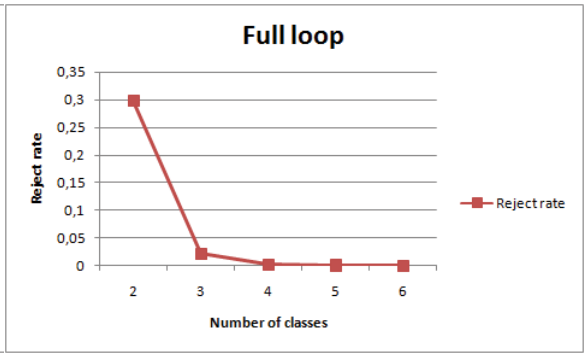

FIGURE 6.4 – Taux de rejet pour une boucle à 3 étapes

lysés avec une autre méthode ; le taux d'erreur obtenu avec la première méthode est la limite vers laquelle tend le taux d'erreur obtenu avec notre chaîne complète du fait de la recherche du temps de calcul minimum. Les résultats seraient nettement meilleurs en choisissant un réglage privilégiant la précision au détriment du temps de calcul.

D'autre part, la décision finale (bouclage ou rejet), se fait en utilisant le modèle des croyances transférables. Or, l'augmentation du nombre de classes implique une répartition équitable de la croyance en toutes les classes, provoquant une baisse du nombre de rejets. Notons que le système tend à ne plus rejeter aucun individu au delà de problèmes à quatre classes. Ce fait tend à privilégier des développements autour d'approches hiérarchiques de décision pour optimiser le critère coût/qualité.

Enfin, du fait que l'annulation des rejets coïncide avec la stabilisation des erreurs, nous supposons qu'à partir de problèmes à quatre classes, les erreurs entre les méthodes ne se compensent plus et que le système atteint un point d'équilibre. A l'instar de la méthode de sélection dynamique des classifieurs présentée au chapitre 4, il faudrait une décision non plus basée sur la répartition des croyances entre les hypothèses, mais sur le choix d'une méthode de décision localement la

plus précise dans l'espace des croyances.

Nous pouvons dire au final que nous ne proposons pas encore "LA" chaîne complète dynamique de traitement d'image, exploitant à tous les étages des mesures sur l'information en sortie des opérateurs pour choisir dynamiquement l'opérateur prochain. Nous proposons en revanche un formalisme permettant d'élaborer la structure d'une telle chaîne.

6.3 Critique/Perspectives

6.3.1 Critique

Le travail présenté ici n'est évidemment pas parfait, et souffre de plusieurs points critiquables.

Dans un premier temps, la méthode présentée au chapitre 3 ne permet de décider d'un bouclage éventuel qu'une fois arrivé en fin de chaîne. Afin d'optimiser les résultats et le temps de calcul, l'idéal voudrait que cette décision se prenne à chaque étape, c'est à dire après analyse de chaque sortie de chacun des opérateurs. Pour le moment, le système ne peut que constater un enchaînement non optimal une fois ce dernier réalisé.

Toujours en ce qui concerne le chapitre 3, l'écriture théorique montre une double influence de la précision de la sortie de l'opérateur de descripteur de formes ; elle est prise en compte à la fois par le classifieur (ce qui est logique puisque l'opérateur de classification sera moins performant avec des attributs peu précis) et sur la décision finale (intégration de la précision de l'attribut via le modèle des croyances transférable).

Nous voyons enfin, au cours du chapitre 4, que le classifieur n'est pas choisi

directement en fonction de l'opérateur de calcul des attributs (cf. idée de Singh & Singh [70]). Étant donné que le classifieur est choisi en fonction de sa précision locale, et que celle-ci est forcément liée à sa capacité à discriminer les données fournies par l'opérateur de descripteur de formes, nous ne savons pas jusqu'à quel niveau joue cette influence.

6.3.2 Perspectives

Au delà des critiques présentées ci dessus, quelques amélioration et optimisations pour compléter notre travail. Ces améliorations portent d'abord sur la partie "décision" (chapitre 3), et particulièrement la gestion du conflit. Le conflit, en théorie de l'évidence, a été intégré par Smets[73], quantifie la discordance entre les sources de croyance et s'écrit :

$$K = \sum_{B \cap C = 0} m_1(B) m_2(C) \tag{6.1}$$

On retrouve d'ailleurs cette valeur au dénominateur de la formule de combinaison de Dempster[15][16] :

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C \neq 0} m_1(B) m_2(C)} \quad \forall A \subseteq \Omega \tag{6.2}$$

Nous avons vu que le système mesure d'abord le conflit entre les sources (en fin de chaîne), et, si sa valeur est acceptable (faible conflit), calcule les probabilités pignistiques de chaque hypothèse. Pour optimiser le traitement et éviter le test d'une condition (i.e. si... alors... sinon...), il est tout à fait imaginable d'intégrer la mesure du conflit comme pondération des probabilités pignistiques favorisant l'hypothèse du rejet.

D'autre part, toujours en ce qui concerne le système de bouclage (chapitre 3), nous présentons une méthode de décision. Comme notre travail se porte sur de la combinaison d'opérateurs, nous pourrions considérer l'étage de décision finale

comme un opérateur, et intégrer dans le système différentes méthodes de décision, dont la combinaison serait la décision du système en fin de chaîne.

Pour finir avec les améliorations portant sur l'étape de décision, nous avons imaginé intégrer deux nouvelles notions :

- La précision d'un opérateur : en se basant sur les notions utilisées en métrologie, la précision d'un opérateur (à ne pas confondre avec les mesures de précision *a priori* et *a posteriori* de l'information manipulée par les opérateurs) serait la finesse de l'information en sortie de cet opérateur, ou une marge d'erreur. Cette notion correspond à un indice de confiance en l'opérateur, en fonction des données sur lesquelles il travaille.
- La pertinence d'un opérateur : cette information est, elle, plus sémantique. Elle décrit l'utilité d'un opérateur en fonction des données à manipuler. La pertinence d'un opérateur de traitement d'image couleur serait par exemple très faible pour une image binaire.

Ces informations contextuelles peuvent aisément être mises en place grâce notamment à une description sémantique de chaque opérateur qui serait intégrée à la base de connaissance du système.

La suite du travail réalisé sur les classifieurs (chapitre 4) pourra par exemple concerner l'exploitation de la forme du voisinage, en utilisant l'information portée par la covariance entre individus (matrices de variance/covariance, valeurs propres), dans l'idée de créer une métrique adaptative comme montré dans [35]. Ce genre de métrique réduit le voisinage dans les directions le long desquelles les centres des classes diffèrent, avec l'intention de finir sur un voisinage pour lequel les centres des classes coïncident (et avoir ainsi le plus proche voisinage approprié pour la classification).

Enfin, le travail présenté sur les opérateurs de calcul d'attributs (chapitre 5)

pourra se voir compléter par la génération des instances des opérateurs de calcul des attributs à la volée (i.e. gestion des paramètres). Il pourrait être finalement intéressant d'analyser les effets de la précision si nous utilisons ensuite une étape de classification après le calcul des moments et de trouver une relation entre la quantité d'information préservée et les performances de reconnaissance. D'autres méthodes de similarité doivent également être testées, comme d'autres métriques de distances ou en se basant sur des notions comme l'entropie. Il est ainsi possible de choisir le descripteur de forme le plus approprié, en tenant compte de la précision, selon une connaissance *a priori*, en fonction d'informations à propos du document en cours d'analyse.

C'est dans cette approche que s'inscrivent les travaux actuels de Sébastien Desbouchages[20][62], au sein du Laboratoire XLIM-SIC, basés sur un système à plusieurs familles d'attributs (histogrammes des couleurs, caractéristiques de texture, attributs de formes etc.). Un tel système choisit dynamiquement l'opérateur de calcul d'attributs, ou pondère sa décision, en fonction du critère de précision fourni par les attributs. À l'heure actuelle, ces travaux montrent qu'un tel système est réalisable, et améliore les taux de reconnaissance obtenus avec des classifieurs et des attributs simples.

6.4 Conclusion finale

Au delà des faiblesses et des amélioration de ce travail, nous avons posé en début de document un ensemble de questions mesurant si nos objectifs initiaux sont atteints ou pas. Revoyons à présent ces questions, ainsi que les réponses proposées, basées sur le contenu du manuscrit.

– Avons-nous une application de reconnaissance de caractères ? Clairement oui. Notre application présente évidement des faiblesses, elle tient plus du

prototype que de l'application grand public, mais nous avons bien réalisé un procédé capable de reconnaître des caractères manuscrits après apprentissage.

– Cette application se construit-elle dynamiquement par sélection dynamique des opérateurs disponibles parmi une bibliothèque d'opérateurs simples ? Nous montrons, dans le chapitre 5, comment choisir dynamiquement un opérateur de description de forme. Nous détaillons également, au cours du chapitre 4, comment il est possible de choisir dynamiquement un opérateur de classification en fonction de la précision locale de celui ci dans l'espace des attributs. Nous expliquons enfin, dans le chapitre 3, comment il est possible, d'une part, d'avoir une chaîne de traitements construite dynamiquement en fonction de l'individu en entrée, des critères d'estime et d'usage (i.e. privilégier le temps de calcul sur la précision du résultat ou vice versa). D'autre part, nous détaillons comment intégrer à la fois la précision des opérateurs de calcul d'attributs et la sélection dynamique de classifieurs, avec exploitation de la réponse de ceux ci comme mesure de précision *a posteriori*, par le biais du modèle des croyances transférable, pour obtenir une chaîne de traitement non linéaire et dynamique.

– Les performances en terme de taux de bonne reconnaissance sont elles meilleures que celles obtenues par le meilleur enchaînement statique des opérateurs utilisés ? Nos résultats sont encore très loin de ceux présentés dans la littérature, obtenus avec des méthodes bien plus complexes. Au regard par contre des familles d'attributs choisies, ainsi que les outils de classifications, nous proposons, dans les chapitre 4 et 5, des résultats meilleurs que ceux obtenus par choix statique de la méthode de calcul d'attributs et du classifieur.

– Ce travail apporte t'il une base de réflexion théorique au problème ? C'est

certainement l'objectif principal de notre travail et nous n'avons aucunement la prétention d'apporter la meilleure réponse. Nous expliquons toutefois le principe de sélection dynamique d'un classifieur et proposons un formalisme pour les opérateurs, leur sélection ainsi que la décision à prendre en fin de chaîne.

Au final, nous pouvons dire que, bien qu'il subsiste encore énormément de travail avant d'aboutir à une chaîne complète de traitement d'image générique, dynamique et auto régulée, nous avons atteint les objectifs visés au début de ce travail. Nous espérons fortement que ces travaux serviront de base pour des études et des développements ultérieurs sur le sujet.

"*Dans toute action, dans tout choix, le bien c'est la fin, car c'est en vue de cette fin qu'on accomplit toujours le reste.*"
Aristote

Bibliographie

[1] A. Appriou. Probabilités et incertitudes en fusion de données multi-senseurs. *Revue Scientifique et Technique de la Défense*, 11 :27–40, 1991.

[2] D. Arrivault. *Apport des graphes dans la reconnaissance non-contrainte de caractères manuscrits anciens*. PhD thesis, Université de Poitiers, 2002.

[3] S. O. Belkasim, M. Shridhar, and A. Ahmadi. Pattern recognition with moment invariants : A comparative study and new results. *Pattern Recognition*, 24 :1117–1138, December 1991.

[4] S. O. Belkasim, M. Shridhar, and A. Ahmadi. Corrigendum. *Pattern Recognition*, 26 :377, January 1993.

[5] I. Bloch. Fusion d'informations numériques : panorama méthodologique. In *Journées Nationales de la Recherche en Robotique*, 2005.

[6] H. Blum. A transformation for extracting new descriptions of shape. *Models for the Perception of Speech and Visual Form*, pages 362–380, 1967.

[7] M. Boll. *Que sais-je ? Les certitudes du hasard*. PUF, 1942.

[8] T. D. Bui and G. Chen. Invariant fourier-wavelet descriptor for pattern recognition. *Pattern Recognition*, 32 :1083–1088, 1999.

[9] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *ICML '06 : Proceedings of the 23rd international conference on Machine learning*, pages 161–168, New York, NY, USA, 2006. ACM.

[10] G. Y. Chen, T. D. Bui, and A. Krzyzak. Contour-based handwritten numeral recognition using multiwavelets and neural networks. *Pattern Recognition*, 36(7) :1597–1604, 2003.

[11] R. Clouard. *Raisonnement incrémental et opportuniste appliqué à la construction dynamique de plans de traitement d'images*. PhD thesis, Université de Caen, 1994.

[12] C. Conversano, R. Siciliano, and F. Mola. Supervised classifier combination through generalized additive multi-model. In *MCS '00 : Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 167–176, London, UK, 2000. Springer-Verlag.

[13] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento. A cascaded multiple expert system for verification. In *MCS '00 : Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 330–339, London, UK, 2000. Springer-Verlag.

[14] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Trans. on Information Theory*, 13(1) :21–27, 1967.

[15] A. P. Dempster. Upper and lower probabilities induced by multiple valued mappings. *Annals of Mathematical Statistics*, 38 :325–339, 1967.

[16] A. P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society*, 30 :205–247, 1968.

[17] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7) :1095–1107, 1997.

[18] Thierry Denœux. A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25 :804–813, 1995.

[19] Bureau International des Poids et Mesures. What is metrology. http ://www.bipm.org/en/convention/wmd/2004/, 2004.

[20] S. Desbouchages, N. Richard, and A.S. Capelle-Laizé. A new scheme of merger information based on accuracy for image classification. *Workshop on the Theory on Belief Functions*, pages 144–149, April 2010.

[21] L. Didaci, G. Giacinto, F. Roli, and G. L. Marcialis. A study on the performances of dynamic classifier selection based on local accuracy estimation. *Pattern Recognition*, 38 :2188–2191, 2005.

[22] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 7(10) :1895–1942, 1998.

[23] J. Dombre. *Systèmes de représentation multi-échelles pour l'indexation et la restauration d'archives médiévales couleurs*. PhD thesis, Université de Poitiers, 2003.

[24] D. Dubois and H. Prade. A set-theoretic view of belief functions : Logical operations and approximations by fuzzy sets. *Int. Jour. of General Systems*, 12 :193–226, 1986.

[25] R. P. W. Duin and D. M. J. Tax. Experiments with classifier combining rules. In *MCS '00 : Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 16–29, London, UK, 2000. Springer-Verlag.

[26] U. Fayyad, G. Pietetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3) :37–54, 1996.

[27] U. Fayyad, G. Pietetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining : towards a unifying framework. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 82–88, 1996.

[28] J. L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 1981.

[29] J. Ghosh. Multiclassifier systems : Back to the future. In *MCS '02 : Proceedings of the Third International Workshop on Multiple Classifier Systems*, pages 1–15, London, UK, 2002. Springer-Verlag.

[30] G. Giacinto and F. Roli. Adaptive selection of image classifiers. *ICIAP '97, Lecture Notes in Computer Science*, 1310 :38–45, 1997.

[31] G. Giacinto and F. Roli. Dynamic classifier selection based on multiple classifier behaviour. *Pattern Recognition*, 34 :1879–1881, 2001.

[32] G. Giacinto, F. Roli, and G. Fumera. Selection of image classifiers. *Electronics Letters*, 36(5) :420–422, 2000.

[33] D. J. Hand, N. M. Adams, and M. G. Kelly. Multiple classifier systems based on interpretable linear classifiers. In *MCS '01 : Proceedings of the Second International Workshop on Multiple Classifier Systems*, pages 136–147, London, UK, 2001. Springer-Verlag.

[34] D. Harmanec. Measure of uncertainty and information. In *Imprecise Probability Project,1999 (http ://ippserv.rug.ac.be/home/ipp.html)*, 1999.

[35] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(6) :607–616, 1996.

[36] T. K. Ho. Complexity of classification problems and comparative advantages of combined classifiers. In *MCS '00 : Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 97–106, London, UK, 2000. Springer-Verlag.

[37] M. K. Hu. Pattern recognition by moment invariants. *Proc. IRE (correspondence)*, 49 :1428, 1961.

[38] M. K. Hu. Visual pattern recognition by moment invariants. *IRE Trans. on Information Theory*, IT-8 :179–187, 1962.

[39] Y. S. Huang and C. Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17 :90–94, 1995.

[40] K. G. Ianakiev and V. Govindaraju. Architecture for classifier combination using entropy measures. In *MCS '00 : Proceedings of the First Internatio-*

nal Workshop on Multiple Classifier Systems, pages 340–350, London, UK, 2000. Springer-Verlag.

[41] J. Kholas and P. A. Monney. *Lecture Notes in Economics and Mathematical Systems*, volume 425, chapter A mathematical theory of hints : An approach to the Dempster-Shafer theory of evidence. Springer-Verlag, 1995.

[42] A. Khotanzad and Y. H. Hong. Invariant image recognition by zernike moments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(5) :489–497, 1990.

[43] A. Khotanzad and Y. H. Hong. Rotation invariant image recognition using features selected via a systematic method. *Pattern Recognition*, 23(10) :1089–1101, 1990.

[44] G. J. Klir and M. J. Wierman. *Uncertainty-based information. Elements of generalized information theory, 2nd edition. Studies in fuzzyness and soft computing*. Physica-Verlag, 1999.

[45] L. I. Kuncheva. Switching between selection and fusion in combining classifiers : an experiment. *IEEE Trans. Syst. Man. Cybernet. Part B 32*, 2002.

[46] L. I. Kuncheva. *Combining Pattern Classifiers*. Wiley, 2004.

[47] L. I. Kuncheva, M. Skurichina, and R.P.W. Duin. An experimental study on diversity for bagging and boosting with linear classifiers. *Information fusion*, pages 245–258, 2002.

[48] P. Latinne, O. Debeir, and C. Decaestecker. Different ways of weakening decision trees and their impact on classification accuracy of dt combination. In *MCS '00 : Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 200–209, London, UK, 2000. Springer-Verlag.

[49] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. DUNOD, 1995.

[50] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *IEEE*, 86(11) :2278–2324, November 1998.

[51] Y. LeCun and C. Cortes. The mnist database of handwritten digits. http ://yan.lecun.com/exdb/mnist, 1998.

[52] E. Lefevre, O. Colot, and P. Vannoorenberghe. Belief function combination and conflict management. *Information Fusion*, 3 :149–162, 2002.

[53] Y. Li. Reforming the theory of invariant moments for pattern recognition. *Pattern Recognition*, 25(7) :723–730, 1992.

[54] S. W. Looney. A statistical technique for comparing the accuracies of several classifiers. *Pattern Recognition Letters*, 8(1) :5–9, 19988.

[55] S. P. Luttrell. A self-organising approach to multiple classifier fusion. In *MCS '01 : Proceedings of the Second International Workshop on Multiple Classifier Systems*, pages 319–328, London, UK, 2001. Springer-Verlag.

[56] D. Mercier, B. Quost, and T. Denoœux. Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion*, 2007.

[57] N. Metropolis. The beginning of the monte carlo method. *Los Alamos Science*, 15 :125–130, 1987.

[58] J. Parker. Rank and response combination from confusion matrix data. *Information Fusion*, 2 :113–120, 2001.

[59] K. Rahbar, M. Rahbar, and F. M. Kazemi. Handwritten numeral recognition using multi-wavelets and neural networks. In *5th WSEAS Internation Conference on Signal Processing*, pages 56–58, May 2006.

[60] E. Ramasso. *Reconnaissance de séquences d'états par le Modèle des Croyances Transférables - Application à l'analyse de vidéos d'athlétisme*. PhD thesis, Université Joseph Fourier de Grenoble, 2007.

[61] T. H. Reiss. The revised fundamental theorem of moment invariants. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13 :830–834, August 1991.

[62] N. Richard, A.S. Capelle-Laizé, and S. Desbouchages. Combinaison dynamique d'attributs d'images selon un critère d'exactitude. *In COtation*

des Informations : Théorie et Applications, Ingénierie des Connaissances (IC2010), June 2010.

[63] I. Rish. An empirical study of the naive bayes classifier. In *IJCAI-01 workshop on "Empirical Methods in AI"*, 2001.

[64] F. Roli, G. Fumera, and J. Kittler. Fixed and trained combiners for fusion of imbalanced pattern classifiers. *5th International Conference on Information Fusion*, pages 278–284, 2002.

[65] F. Roli, G. Giacinto, and G. Vernazza. Methods for designing multiple classifier systems. In *MCS '01 : Proceedings of the Second International Workshop on Multiple Classifier Systems*, pages 78–87, London, UK, 2001. Springer-Verlag.

[66] J. W. Jr Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18 :401–409, 1969.

[67] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[68] J. Shawe-Taylor and N. Cristianini. *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[69] J. Shutler. Cvonline : 2d moments and their invariants. http ://homepages.inf.ed.ac.uk/cgi/rbf/CVONLINE/entries.pl ?TAG823, August 2002.

[70] S. Singh and M. Singh. A dynamic classifier selection and combination approach to image region labelling. *Signal Processing : Image Communication*, 20 :219–231, 2005.

[71] N. K. Sinha. *Linear Systems*, chapter 3. Wiley, 1991.

[72] M. Skurichina, L. I. Kuncheva, and R. P. W. Duin. Bagging and boosting for the nearest mean classifier : Effects of sample size on diversity and accuracy. In *MCS '02 : Proceedings of the Third International Workshop on Multiple Classifier Systems*, pages 62–71, London, UK, 2002. Springer-Verlag.

[73] P. Smets. The nature of the unnormalized beliefs encountered in the transferable belief model. In *Proceedings of the 8th Annual Conference on Uncertainty in Artificial Intelligence*, pages 292–297, 1992.

[74] P. Smets. Beliefs functions : The disjunctive rule of combination and the generalized bayesian theorem. *Int. Jour. of Approximate Reasoning*, 9 :1–35, 1993.

[75] P. Smets. *Advances in the Dempster-Shafer Theory of Evidence - What is Dempster-Shafer's model ?* Wiley, 1994.

[76] P. Smets. Decision making in the tbm : The necessity of the pignistic transformation. *Int. Jour. of Approximate Reasoning*, 38 :133–147, 2005.

[77] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66(2) :191–234, 1994.

[78] S. Sochacki. Planification des taches du traitement d'images. Master's thesis, Université de La Rochelle, 2003.

[79] S. Sochacki, N. Richard, and P. Bouyer. A comparative study of different statistical moments using an accuracy criterion. In *Seventh IAPR International Workshop on Graphics Recognition - GREC 2007*, September 2007.

[80] S. N. Srihari, J. J. Hull, and T. K. Ho. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1) :66–75, 1994.

[81] D. M. J. Tax and R. P. W. Duin. Combining one-class classifiers. In *MCS '01 : Proceedings of the Second International Workshop on Multiple Classifier Systems*, pages 299–308, London, UK, 2001. Springer-Verlag.

[82] J. R. Taylor. *An Introduction to Error Analysis : The Study of Uncertainties in Physical Measurements*. University Science Books, U.S., 2nd ed. edition, July 1997.

[83] M. R. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America*, 70(8) :920–930, 1979.

[84] Ø. D. Trier, A. K. Jain, and T. Taxt. Feature extraction methods for character recognition - a survey. *Pattern recognition*, 29(4) :641–662, 1996.

[85] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR01)*, volume 1, pages 511–518, Kaui(Hawaii), Décembre 2001.

[86] M. S. Wong and W. Y. Yan. Investigation of diversity and accuracy in ensemble of classifiers using bayesian decision rules. In *International Workshop on Earth Observation and Remote Sensing Applications (EORSA 2008)*, pages 1–6, 2008.

[87] K. Woods, W. P. Kegelmeyer, and K. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19 :405–410, 1997.

[88] L. Xu and A. Krzyzakand C. Y. Suen. Methods for combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3) :418–435, 1992.

[89] B. Yaghlane, P. Smets, and K. Mellouli. Belief function independence : I. the marginal case. *Int. Jour. of Approximate Reasoning*, 29 :47–70, 2001.

[90] B. Yaghlane, P. Smets, and K. Mellouli. Belief function independence : Ii. the conditionnal case. *Int. Jour. of Approximate Reasoning*, 31(1) :31–75, 2002.

[91] B. Yaghlane, P. Smets, and K. Mellouli. Independence concept for belief functions. In Physica-Verlag, editor, *Technologies for constructing intelligent systems : Tools*, 2002.

[92] F. Zernike. Beugungstheorie des schneidenverfahrens und seiner verbesserten form, der phasenkontrastmethode (diffraction theory of the cut procedure and its improved form, the phase contrast method). *Physica*, 1 :689–704, 1934.

[93] H. Zhang. The optimality of naive bayes. In *FLAIRS Conference*, 2004.

[94] H. Khoufi Zouari. *Contribution à l'évaluation des méthodes de combinaison parallèle de classifieurs par simulation*. PhD thesis, Université de Rouen, December 2004.