Développement d'un outil d'aide au diagnostic pour la production de maïs permettant la réduction de la consommation en eaux d'irrigation et en traitements phytosanitaires

Elisa Roux

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Institut National des Sciences Appliquées de Toulouse (INSA de Toulouse)*

Présentée et soutenue le *11/12/2015*

par : ELISA ROUX

Développement d'un outil d'aide au diagnostic pour la production de maïs permettant la réduction de la consommation en eaux d'irrigation et en traitements phytosanitaires

---

JURY

Louise TRAVE-MASSUYES	Professeur d'Université	Présidente du Jury
Guillaume HUBY	Ingénieur	Membre du Jury
Bouchra LAMRINI	Chargé de Recherche	Membre du Jury

École doctorale et spécialité :
EDSYS : Informatique 4200018

Unité de Recherche :
LAAS-CNRS

Directeur de Thèse :
Marie-Véronique LE LANN

Rapporteurs :
Nathalie PERROT et *Rafaël GOURIVEAU*

Table des matières

Introduction

La classification intervient dans de nombreux domaines scientifiques et industriels et apparaît bien souvent comme étant une étape cruciale de l'analyse d'un système. En chimie, on parle de classification périodique des éléments, en biologie de classification des espèces, en théorie des langages de la classification de Chomsky (ou hiérarchie de Chomsky), en astronomie de la classification des étoiles,... Les théorisations scientifiques se basent systématiquement sur des classifications.

La classification se définit comme étant une répartition en groupes, en catégories, d'objets concrets ou abstraits, de notions, de données, ayant des traits communs, généralement dans le but d'en faciliter l'étude. C'est d'abord une opération de l'esprit qui, pour la commodité des recherches ou de la nomenclature, pour le secours de la mémoire, pour les besoins l'enseignement, ou dans tout autre but relatif à l'homme, groupe artificiellement des objets auxquels il trouve quelques caractères communs, et donne au groupe artificiel ainsi formé une étiquette ou un nom générique [Cournot, 1851]. Pour Henri Poincarré, "la science n'est qu'une classification, et une classification ne peut être vraie, mais commode."

De fait, le passage de l'être à la pensée est déjà une classification : le mot "table" peut aussi bien désigner un objet particulier que l'image abstraite que notre esprit a formée de toutes les instances possédant les caractéristiques d'une table. C'est parce que ses parents désignent toutes les tables par ce terme que l'enfant définira conceptuellement ce qu'est une table. Cette opération systématique, transformant des signaux sensoriels en représentations mentales, permet de développer une compréhension de notre environnement et, de là, une interaction cohérentes avec les objets qui nous entourent ainsi que de communiquer avec nos semblables. Le langage manifeste cette classification en attribuant des étiquettes, des noms aux différents objets mentaux.

La classification établit donc un modèle compréhensible, interprétable de la réalité. La connaissance que l'on a de la réalité est donc étroitement liée à la représentation que nous fournit cette modélisation. Ainsi, bien que nécessaire pour fonder la pensée et la science, elle n'a pas systématiquement prétention à être une solution de modélisation unique ; plus le système à représenter est complexe, plus le nombre de classifications candidates est élevé. Néanmoins, l'enjeu d'une classification peut souvent aider à lever les indécisions et déterminer quels critères favoriser. Une classification est spécifique à une situation et un problème spécifiques et nécessite une analyse informée du contexte, de ses particularités et de ses contingences.

Notre thèse intervient dans le cadre d'un projet agronomique visant à apporter aux cultivateurs de maïs un conseil adapté à la situation particulière de sa parcelle, de manière à optimiser son rendement tout en limitant ses besoins en eau d'irrigation et en produits phytosanitaires. Ce conseil se base sur l'analyse des caractéristiques de sa parcelle, après qu'elle a été classée suivant un modèle préalablement établi.

Notre étude s'inscrit dans la continuité des travaux de l'équipe DISCO (DIagnostic, Supervision et COnduite) au LAAS (Laboratoire d'Analyse et d'Architecture des Systèmes), qui intervient dans le

domaine du diagnostic automatique pour proposer des solutions de manière à assurer une meilleure maîtrises des systèmes dynamiques complexes. Ses recherchent se basent sur des formalismes issus des domaines de l'Intelligence Artificielle et de l'Automatique.

Dans le présent document, nous allons présenter notre travail et sa démarche.

Notre travail s'inscrit dans un projet FUI (Fonds Unique Interministériel) global appelé MAISEO, piloté par VIVADOUR et regroupant six autres participants : GEOSYS, Pioneer Genetique, CACG (Compagnie d'Aménagement des Coteaux de Gascogne), Météo-France, CESBIO (Centre d'Etuds Spatiales et de la BIOsphère), et le LAAS. Nous allons tout d'abord présenter ce projet, ses motivations et ses objectifs. Nous en dégageons les enjeux et les caractéristiques particulières, de manière à en établir précisément les contraintes et les besoins. Cette étude préalable est déterminante pour le choix de la méthode de classification et des données à prendre en compte.

Nous décrivons ensuite les différentes méthodes de classification de données, en précisant l'intérêt que présente chacune d'elles mais également leurs limites dans le cadre de notre étude. Nous expliquons ainsi notre choix de la méthode LAMDA, qui nous a paru particulièrement adaptée à notre situation.

Le chapitre suivant est dédié à la présentation particulière de la méthode LAMDA. Nous expliquons son fonctionnement dans le détail et présentons un algorithme d'évaluation de la qualité des partitions qu'elle permet d'obtenir.

Nous poursuivons par la description des modifications que nous avons jugé intéressant de lui apporter pour traiter plus efficacement les types de données employées pour la classification.

Enfin, nous présentons les résultats obtenus par l'application de la méthode LAMDA dans notre contexte précis, ainsi que l'intérêt que présente notre contribution à cette méthode.

1 Le projet MAISEO

Cette thèse s'intègre dans le cadre d'un projet FUI agronomique afin de permettre une réduction globale de la consommation en eau d'irrigation et de l'usage de traitements phytosanitaires. Pour cela, le projet MAISEO intervient à différentes échelles et auprès de différents acteurs ; c'est dans deux lots du projet qu'apparaît notre contribution.

L'objectif de MAISEO est de développer de nouvelles approches pour maintenir la production de maïs grain et mieux gérer la ressource hydrique au niveau du territoire. Le territoire d'expérimentation du projet est centré sur le département très agricole du Gers et quelques départements limitrophes, en Midi-Pyrénées. MAISEO intervient :

- Du côté de l'exploitation agricole, puisqu'il s'agit de développer de nouvelles solutions agronomiques et assurer les mêmes rendements de production du maïs en réduisant l'irrigation de 20% afin de libérer de l'eau pour le milieu naturel, et de nouveaux systèmes d'informations pour définir et piloter la culture dont les données liées à l'irrigation pourront être transmises au gestionnaire du bassin,
- Du côté gestionnaire de la ressource hydrique : le but est de développer de nouveaux systèmes d'informations pour mieux gérer la relation ressource-besoin d'irrigation et anticiper les situations de crises.

Pour cela, il suit étapes de déroulement :

1. Cartographie et caractérisation des zones homogènes de culture en potentiel maïs sur les surfaces cultivées en irrigué et en sec du territoire d'expérimentation et sélection des parcelles expérimentales représentatives des différentes zones homogènes pour les tests des idéotypes maïs,
2. Développement d'un nouveau système d'information pour le conseil au semis du maïs en pré-campagne au niveau de la parcelle (en hiver au moment de l'achat des semences de décembre à février) - le système doit fournir les informations nécessaires au choix variétal, de date et de densité de semis répondant le mieux au potentiel et condition de la parcelle pour l'année,
3. Développement d'un nouvel Outil d'Aide à la Décision et à l'Action (OADA) basé sur le suivi journalier en temps réel et la simulation des rendemnets et du bilan hydrique à la parcelle,
4. Définition des idéotypes variétaux et culturaux de maïs grain pour les différentes zones de cultures, permettant de maintenir ou améliorer le rendement grain avec un itinéraire de culture avec une baisse d'au moins 20% de la consommation en eau et de 10% en engrais azoté,
5. Développement d'un nouveau système d'information pour déterminer les surfaces irriguées avant le début des irrigations en juin, de manière à fournir au gestionnaire du bassin versant une carte des assolements, des surfaces irriguées, et des dates de levées pour les cultures de printemps sur l'ensemble du bassin, afin d'estimer la demande en eau sur la période d'irrigation (juillet-août) et d'ajuster éventuellement les quotas en eau disponible entre les irrigants,
6. Développer et montrer l'intérêt d'un nouveau système d'information pour le suivi hydrique journalier en cours de campagne,

7. Evaluer les performances environnementales de la production de maïs grain issue des itinéraires techniques améliorants MAISEO en comparaison avec les performances environnementales des productions témoins sur les mêmes parcelles, la même année, et avec les impacts du maïs grain à l'échelle nationale.

En ce qui concerne notre implication dans le projet, un premier objectif est de fournir aux producteurs de maïs un outil automatisé, disponible en ligne, pour les guider dans les choix qui leur incombent. En effet, les décisions que doivent prendre les agriculteurs quant à la méthode employée pour gérer au mieux leurs cultures reposent sur un diagnostic préalable concernant l'état de leur parcelle, et peuvent avoir de nombreuses répercussions - tant sur le plan du rendement et du coût de revient que sur des aspects environnementaux. Ainsi, dans le but de les assister et de soumettre à leur attention un éventail de solutions optimales, le projet MAISEO a été mis en place pour proposer aux agriculteurs un outil d'aide à la décision intégrant un système de diagnostic basé sur un certain nombre d'informations disponibles concernant leurs parcelles. Il se propose d'établir le diagnostic de chaque exploitation en se basant sur ses particularités agronomiques et sur le profil météorologique de la zone à laquelle il appartient.

La seconde partie du projet au cours de laquelle nous intervenons s'intéresse à la prédiction des besoins en eau d'îlots d'exploitations de maïs, de manière à en gérer au mieux la distribution aux différentes zones. Le suivi hydrique au niveau du bassin versant vise à prévenir les situations de crise au maximum de la demande, et mieux satisfaire les DOE. Le DOE (Débit d'Objectif d'Etiage) est le débit de référence permettant le maintien d'un bon état des eaux et au-dessus duquel l'ensemble des usages est satisfait en moyenne 8 années sur 10.

La figure présentée dans l'annexe A.1 décrit le déroulement global du projet.

Dans ce chapitre, nous exposons tout d'abord le contexte dans lequel est né le projet MAISEO, afin d'en expliquer les objectifs en matière de consommation hydrique et pourquoi un soucis particulier est accordé à la culture du maïs. Ensuite, nous détaillerons les visées respectives des parties du projet dans lesquelles nous sommes impliqués, en dressant tout d'abord un état de l'art des pratiques culturales et gestionnaires avant MAISEO, et décrivant en suivant le principe de fonctionnement du système à mettre en place.

1.1 Contexte

La France est le premier producteur de maïs grain en Europe avec 15 Mt produites en moyenne par année. Le grand Sud-Ouest totalise 41% de la production maïs grain française, dont 20% en Aquitaine et 11% en Midi-Pyrénées. Pourtant, au cours des cinq dernières années, la consommation de maïs s'est montrée supérieure à sa production de sorte qu'en 2012 - date à laquelle a été conçue l'idée du projet - les stocks ont atteint leur seuil critique. De fait, outre son usage dans l'alimentation humaine, le maïs est surtout consommé par le bétail et la volaille, puisque 2/3 de sa production sont dévolus à l'alimentation animale. Enfin, le maïs est impliqué dans la fabrication de colle pour l'industrie textile, d'édulcorants, de produits de l'industrie pharmaceutique, et de plastiques biodégradables et biocarburants.

L'eau est indispensable aux plantes à tous les niveaux. A l'échelle moléculaire, l'eau agit comme matrice pour toutes les réactions enzymatiques au niveau de la phase photochimique de la photosynthèse, et apporte de l'hydrogène et de l'oxygène. A l'échelle de la cellule, l'eau a un impact direct sur l'architecture des organes et leur élongation. Enfin, à l'échelle de la plante, elle permet l'assimilation des solutés présents dans le sol et leur migration vers les parties aériennes de la plante, tout en assurant en parallèle une régulation thermique des tissus exposés aux rayons du soleil. Par conséquent, un déficit en eau prolongé modifie les composantes du rendement et de la fertilité. Le déficit hydrique s'installe dans la plante quand l'absorption ne peut pas satisfaire la demande de la transpiration. On parle de stress hydrique lorsque la culture a épuisé sa réserve facilement utilisable (c'est-à-dire sans réduction de sa croissance), évaluée aux deux tiers de la réserve en eau du sol en sol peu profond et entre la moitié et les deux tiers dans les sols profonds. L'irrigation est un facteur économique important qui génère un gain de rendement par rapport à la culture en sec d'environ 28 q/ha en moyenne pour le maïs en région Midi-Pyrénées.

La culture du maïs est celle qui représente le rendement le plus élevé à l'hectare, mais ses besoins en eau sont problématiques : si la disponibilité en eau constitue un facteur limitant pour la croissance et la productivité des plantes en général, le maïs est particulièrement soumis à ce problème puisqu'il termine son cycle en été - c'est-à-dire à une période de faible pluviométrie. De plus, bien que sa photosythèse soit dite en "C4", son empreinte eau reste très élevée. Les plantes en "C4", décrites ainsi parce que le premier glucide formé comporte quatre atomes de carbone, bénéficient d'une efficacité photosynthétique et d'une efficience de l'eau bien supérieure à celles des plantes en "C3" (le blé, l'orge,...). Mais le maïs reste une plante exotique et se trouve très gourmand en eau. (voir Annexe Tableau A.1) La culture du maïs grain représente 60% des surfaces irriguées et consomme 70 à 80% des volumes d'irrigation, soit environ $250 Mm^3/an$ Dans ce contexte, la production de maïs est en danger avec les perspectives du réchauffement climatique et de restriction de la ressource hydrique ; le Plan National d'Adaptation au Changement Climatique a fixé comme objectif d'économiser 20% de l'eau prélevée d'ici 2020.

1.2 Classification et apprentissage de parcelles de maïs

1.2.1 Pratiques culturales avant le projet MAISEO

L'initiative du projet MAISEO a été doublement motivée par des préoccupations d'ordre écologique et par les difficultés manifestées par les agriculteurs quant au choix de leurs pratiques culturales.

Le rendement du maïs est en particulier influencé par divers facteurs génétiques, climatiques et agronomiques. Pourtant, peu d'agriculteurs possèdent les connaissances scientifiques nécessaires à une approche visant l'optimisation. De fait, il n'existe actuellement pas d'outil permettant aux agriculteurs d'établir un diagnostic de leurs parcelles - ni d'outil d'aide à la décision. Ces évaluations ont, jusqu'à présent, toujours été réalisées de manière intuitive et empirique par les agriculteurs eux-mêmes et les techniciens en s'appuyant sur leur propres connaissances historiques de la parcelle ; ces références qui ne sont pas formalisées mais reposent sur leurs souvenirs des campagnes précédentes ou d'événements

exceptionnels marquants, comme par exemple un niveau de récolte inhabituel au cours d'une année climatique extrême.

Les seules méthodes de classification de parcelles existantes avaient été réalisées dans le domaine des vignobles, pour lequel la notion de pédoclimat est essentielle. En ce qui concerne les cultures céréalières, quelques études avaient été menées sur la classification parcellaire, sur des petites zones, sans vocation à être étendues ni généralisables. Ces travaux ne couvrent qu'une petite partie du territoire d'une coopérative et n'ont pas été diffusés dans le milieu agricole, ne permettant ainsi pas aux agriculteurs d'en exploiter les résultats pour l'amélioration de leurs pratiques culturales. Les agriculteurs gèrent l'assolement de leurs exploitations - et donc le choix de leurs cultures - en réalisant intuitivement des classifications relatives à l'ensemble de leurs parcelles. Les éléments pris en compte sont :

- Le potentiel de la parcelle, estimé essentiellement d'après le souvenir des rendements passés. C'est l'expression de la capacité de rendement du sol pour des itinéraires techniques identiques.
- Les critères organisationnels, c'est-à-dire la distance de l'exploitation, la taille de la parcelle, la topographie et son impact sur la mécanisation, le fait que la parcelle soit irriguée ou non.

Des méthodes de classification des sols ont été explorées, mais elles ne sont pas spécifiques à une culture précise et ne couvrent qu'une zone très restreinte. L'étude de S. Bruckert [Bruckert, 1989] notamment, dégage les critères impliqués dans l'aptitude à la mise en culture des sols. Il en a retenu trois principaux :

- La situation géographique, et ses incidences sur les conditions météorologiques et la pente,
- L'organisation pédologique, et en particulier l'aération du sol, sa porosité, et les obstacles à l'enracinement
- La constitution chimique des sols.

Les parcelles à faible potentiel (caillouteuses, peu profondes, avec mouillères non drainées, inondables...) sont dévolues à des cultures moins rémunératrices voire à des jachères. Pour gérer leurs parcelles et avoir une meilleure visibilité, les exploitants peuvent avoir accès à des SIG. Un SIG (Système d'Information Géographique) est un système d'information conçu pour recueillir, stocker, traiter, analyser, gérer et présenter tous les types de données spatiales et géographiques. L'intérêt principal de cette technique réside dans sa capacité à réunir des données pouvant être très hétérogènes dans un même environnement, quels que soient leur type et leurs origines (coordonnées, latitude et longitude, adresse, altitude, temps, médias sociaux, ...) Mais cette technique demeure très peu répandue, notamment du fait du coût important qu'elle représente [Renoult *et al.*, 2006]. Le soucis de représentation du potentiel agronomique de terres agricoles, entendu ici comme interaction entre les données physiques des terres (caractéristiques pédologiques, topographiques,...) et les données socio-économiques (conditions des marchés locaux, législation fiscales et administratives,...), a donné lieu à la réalisation de cartes à l'échelle régionales, mais leur conception se heurte à quelques difficultés : le problème de l'accessibilité et de la spatialisation des données, la complexité de la combinaison des différents critères,... L'intégration, dans ces cartes, des données socio-économiques posent en particulier la question de la pérennité des informations. En effet, si les seules données physiques bénéficient d'une stabilité suffisante pour une classification des sols durablement signifiante, l'intégration des données socio-économiques menacent les cartes ainsi réalisées d'obsolétisme prématuré. Quelques solutions ont été proposées mais demeurent insuffisantes pour assurer la viabilité des cartes [Guyot et Bornand, 1987].

Aucune étude conjuguant type de la parcelle et données météorologiques n'a été ramenée à l'échelle du territoire d'une coopérative - et encore moins à l'échelle parcellaire. En outre, les seules cartes disponibles ne sont accessibles qu'aux organismes institutionnels. La prise en compte des données météorologiques revêt une importance cruciale dans le choix des variétés. En effet, les différentes variétés de maïs montrent des résistances inégales aux stress qui peuvent les accabler : certaines sont plus vulnérables au froid, d'autres à la chaleur, d'autres encore au manque d'eau... Et ces écueils peuvent sensiblement affecter leur rendement et leur fertilité. Pour le choix des variétés, les agriculteurs et techniciens s'appuient sur des essais réalisés dans le même secteur géographique. Leurs critères de choix concernent le rendement, la sensibilité à des maladies ou ravageurs présents sur les parcelles de l'exploitation, la

vigueur de départ, la vitesse de dessiccation, et l'humidité à la récolte, et la précocité de la variété du maïs à semer.

Les agriculteurs peuvent demander un profil cultural. Il s'agit de creuser une fosse profonde de 1.80 à 2m et d'observer les couches de sol. Généralement cette technique est réalisée par un pédologue ou un technicien spécialement formé pour ce travail. Cette pratique reste néanmoins très peu répandue. Il s'agit d'une technique très laborieuse, nécessitant un matériel adapté et onéreux, qui ne peut être proposée que par certaines chambres d'agriculture et les CETA peuvent proposer une telle prestation. Très peu d'agriculteurs sont sensibilisés à ce genre d'observations : en pratique, les informations décrivant l'opération et les bénéfices qu'elle peut apporter ne se trouvent exposées qu'à quelques groupes d'une dizaine d'agriculteurs qui se réunissent pour assister à une présentation commentée du profil cultural.

En zone vulnérable néanmoins, l'analyse de sol est obligatoire ; elle concerne l'analyse chimique pour la fertilisation. La granulométrie, c'est-à-dire l'analyse physique, est une prestation supplémentaire coûteuse qui présente un intérêt limité pour les agriculteurs, et est, de ce fait, peu répandue [Desbourdes *et al.*, 2007].

1.2.2 Description du projet

L'outil proposé au terme du projet est un portail internet interactif disponible pour les techniciens des coopératives assurant le conseil aux agriculteurs, et les agriculteurs eux-mêmes. L'objectif est de proposer à chacun un diagnostic personnalisé de sa parcelle et de lui soumettre différents scénarios envisageables pour l'aider à prendre une décision. Pour générer ces scénarios, le système mis en place par le projet MAISEO doit disposer d'une base de données riche, basée sur un historique intégrant des parcelles aux profils variés et des informations concernant les différents acteurs de leurs productions. Pour cela, il a fallu déterminer quels pouvaient être ces acteurs, évaluer leur importance et leur pertinence dans le cadre de ce projet, dégager des profils de parcelles, s'assurer de la cohérence et de l'intérêt des résultats obtenus, et lier les différents outils entre eux de manière à fournir un outil complet.

Le projet s'est articulé autour de quatre phases :

- La sélection des informations pertinentes pour l'apprentissage et le diagnostic,
- La classification de parcelles test choisies pour le projet, à partir de la base de données historiques constituée de données issues de capteurs, des relevés météorologiques, des informations communiquées par les agriculteurs,... ,
- La validation des résultats obtenus par la classification,
- L'intégration de ces développements dans un logiciel d'aide au pilotage de la parcelle.

1.2.2.1 Fonctionnement global du système

Le système conçu au cours de ce lot a été réalisé en collaboration avec le groupe GEOSYS. Fondée par des agronomes, ce groupe est consultant en ingénierie et contribue à développer les performances de l'agriculture sur les cinq continents, en fournissant aux acteurs des filières agro-industrielles des solutions basées sur des techniques d'imagerie récentes, pour optimiser l'utilisation des ressources, améliorer les rendements et réduire l'impact des aléas.

Le projet MAISEO prévoit deux phases de traitements des données : l'apprentissage et la reconnaissance. L'apprentissage consiste en la séparation en classes d'un échantillon de parcelles, de sorte à établir des profils représentatifs et déterminer quelles sont, pour chacun d'eux, les pratiques culturales les plus adaptées et offrant le meilleur rendement. La partie reconnaissance permet en suivant de sélectionner l'ensemble des classes dont la définition est la plus proche de la parcelle à analyser, afin que soit soumise à l'agriculteur une vue des différents scénarios envisageables - associant description des pratiques et résultats pouvant être espérés. En fin de campagne, les informations relatives aux choix finaux des cultivateurs sont recueillies chaque année, de manière à alimenter la base de données, et bénéficier chaque année d'un espace de données plus riche.

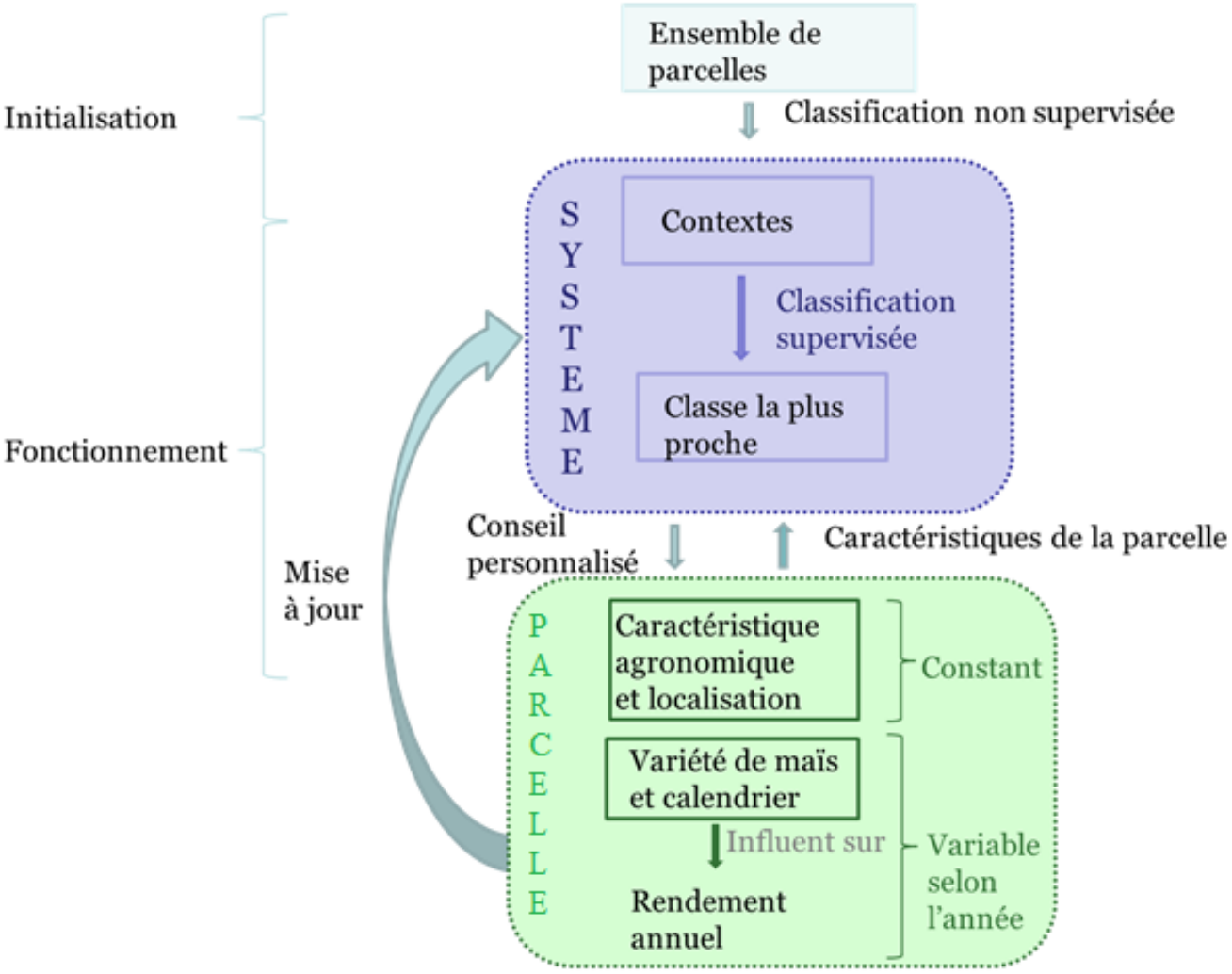


Figure 1. Description du fonctionnement du système

1.2.2.2 Sélection des informations pertinentes

La sélection des variables traitées par le système est une étape primordiale dans sa conception. Elles doivent être à la fois pertinentes au regard de l'objectif de l'outil, et systématiquement disponibles tant pour l'apprentissage que pour la classification.

En premier lieu, pour que l'apprentissage fournisse une base d'analyse fiable, les données d'entraînement doivent être complètes, précises, et représentatives. De ce fait, la sélection des descripteurs s'est opérée dans l'optique de fournir une partition claire et pertinente de l'échantillon d'apprentissage, de sorte à délimiter et définir des types précis de parcelles, déterminants pour l'estimation de leurs contraintes culturales et des besoins en eau. Le diagnostic ne peut ensuite être réalisé que dans la mesure où les critères utilisés pour l'apprentissage représentent des informations systématiquement accessibles par les agriculteurs. En d'autres termes, pour la sélection des critères, il est impératif de prendre en compte les limitations dans la connaissance que les cultivateurs ont de leurs parcelles ; les critères employés pour la définition des classes au cours de l'apprentissage doivent être les mêmes que ceux ensuite demandés aux agriculteurs pour estimer la proximité de leur parcelles aux différentes profils. Les seuls qui diffèrent sont ceux relatifs au conseil à apporter, puisqu'il s'agit là précisément des informations sollicitées par les cultivateurs. La sélection des autres caractéristiques nécessite, de ce fait, de ne considérer que les informations dont disposent habituellement les agriculteurs concernant leurs parcelles, et selon le même degré de précision - dans certains cas, lors de la reconnaissance, l'absence de

quelques informations relatives à la parcelle peuvent être admises sans en affecter la pertinence du résultat final mais cette situation doit être évitée au maximum pour assurer une reconnaissance optimale. Ainsi, si de nombreux critères avaient été d'abord envisagés, seuls quatre ont été retenus. A titre d'exemples, citons la surface de la parcelle et sa composition chimique. Le premier critère n'a pas été considéré comme représentatif de son rendement, et le second concerne une information que ne détiennent que peu d'agriculteurs. En ce qui concerne le soucis porté à la précision des informations détenues par les cultivateurs, la date de semis en est une illustration éloquente ; il est souvent très difficiles pour eux de prévoir avec exactitude la date à laquelle elle peut avoir lieu, du fait notamment du caractère imprévisible des conditions climatiques, aussi le conseil se porte-t-il sur une période de quinze jours.

Au vu de ces contraintes, les critères retenus sont :

- Le type de sol. Le maïs présente un rendement optimal lorsqu'il est cultivé dans des sols profonds et riches, mais il peut s'accommoder de conditions plus difficiles, comme des sols sableux ou plus argileux, voire calcaires, sous réserve de lui assurer les apports d'eau et d'éléments nutritifs nécessaires. Etant entendu que la majorité des agriculteurs ne possèdent pas la connaissance de la composition précise de leur sol, cette caractéristique est représentée par une valeur qualitative, comme par exemple "limon", "limon-argile", "boulbènes",... sans en indiquer la teneur. Il s'agit ainsi d'une modalité intégrant la liste des différents composés présents dans la terre.
- La possibilité d'irriguer. Bien qu'un des objectifs de MAISEO soit de limiter la quantité d'eau d'irrigation, une parcelle irriguée présente bien souvent un meilleur rendement qu'une parcelle non irriguée. Le but n'est donc pas de priver la parcelle d'eau d'irrigation, mais de l'irriguer judicieusement, en se référant aux conditions météorologiques et aux besoins de la plante, de manière à éviter aux plantes de se trouver en état de stress hydrique. La disponibilité de la parcelle en eau d'irrigation est représentée par les modalités "oui" et "non".
- La profondeur de la réserve utile. Il s'agit de la quantité d'eau que le sol peut absorber et restituer à la plante. Plus la réserve utile sera élevée et moins l'agriculteur se verra contraint d'irriguer par lui-même. Evidemment, le niveau de remplissage de la réserve utile sera fonction de sa capacité et de la quantité d'eau de pluie tombée au préalable. Elle est généralement maximale en début de campagne et minimale à la fin de l'été [Bouthiert, 2013]. Elle est ici exprimée en cm.
- La zone météorologique de la parcelle - non pas indiquée par l'agriculteur directement, mais obtenue par géolocalisation. La classification météorologique est importante puisque le maïs nécessite, pour une germination active, une température minimum de 10°C et au moins 18°C pour sa floraison (liée également à une certaine quantité de degrés jours de croissance dépendant de la variété). De plus, la prédiction des stress que pourraient subir le maïs permet de choisir la variété la plus adaptée. Selon 11 critères agro-météorologiques (Annexe Tableau A.2) calculés annuellement sur la période 1991-2010, Météo-France a réalisé un zonage en 7 classes (figure I.2). Le zonage s'appuie sur les données météorologiques Safran (226 mailles sur le domaine) et sur le Q20 et le Q80 (de la période 1991-2010) des 11 variables retenues soit un total de 22 variables. Pour l'ensemble des mailles d'une zone, la médiane de chacune des 22 variables météorologiques a été calculée, ce qui caractérise le comportement moyen de chaque zone (Annexe Tableaux A.3 et A.4). Une maille est un carré de 8km de côté.

En ce qui concerne les critères employés pour l'apprentissage seul, c'est-à-dire les informations destinées à aiguiller l'agriculteur dans son choix, leur nombre est de trois :

- Le rendement espéré. C'est le critère qui permet à l'utilisateur de faire un choix entre les différentes classes dont sa parcelle est diagnostiquée comme étant proche. Il n'est évidemment pas possible de garantir que le rendement indiqué sera atteint, il s'agit d'une estimation basée sur l'observation des rendements des parcelles test. Indiquer le rendement moyen obtenu est apparu comme étant trop risqué, en terme d'espoir donné à l'agriculteur ; c'est donc un intervalle qui lui est présenté, prenant comme bornes la valeur minimale et la valeur maximale du rendement des parcelles d'entraînement ayant formé la classe. Les deux valeurs sont exprimées en q/ha.
- Le degré de précocité du grain employé. Les différentes modalités pouvant être "très tardif", "tar-

dif", "semi-tardif", semi-précoce", "précoce", "très précoce", cette caractéristiques est également représentée par une valeur allant de 1 à 6, 1 symbolisant le "très tardif" et 6 le "très précoce". Contrairement à l'idée que l'on peut s'en faire intuitivement, plus un grain est tardif, plus il est préconisé de le planter tôt dans l'année. La précocité correspond à la durée du cycle de développement de la plante, entre le semis et la récolte. Une variété très précoce a un cycle court, par opposition à une variété tardive dont le cycle est long : plus la variété est précoce et moins elle a besoin d'unités de chaleur pour atteindre la maturité. À l'inverse, plus la variété est tardive et plus ses besoins en unités de chaleur - ou somme des températures - sont élevés. Les variétés tardives témoignent d'un rendement supérieur aux variétés précoces - de 15 à 20% supérieur ; mais les variétés précoces présentent un avantage au cours des années plus froides puisqu'elles permettent de produire du maïs à cycle court (1 600 degrés-jours), là où le maïs tardif n'aurait pas assez de temps pour pousser.

– La date de semis. Elle dépend de la précocité du grain semé, et aussi des conditions climatiques auxquelles la parcelle est soumise, en particulier la chaleur et la pluie, puisqu'elles influent respectivement sur la rapidité de floraison de la plante, et sur son bon développement. L'éventail des dates de semis a été découpé en 6 plages de 15 jours, s'étalant du 15 mars au 15 juin, aussi cette information est-elle représentée par une valeur allant de 1 à 6.

Pour une parcelle donnée, les informations relatives à la parcelle en elle-même (le type de sol, la profondeur de la réserve utile, et sa position géographique) restent inchangées au fil des années. Les pratiques culturales, qu'il s'agisse du type de grain choisi (et donc sa précocité) ou de la date de semis, peuvent varier d'une année sur l'autre en fonction des préférences de l'agriculteur et de sa volonté de s'adapter aux conditions particulières de l'année courante. La question de l'irrigation pourrait être considérée comme faisant partie des pratiques culturales, et être ainsi dépendre de paramètres changeants, s'il s'agissait de la traiter en terme de quantité d'eau - ce qui augmenterait la précision du système. Malheureusement, cette information étant apparue à nos partenaires comme trop difficile à obtenir avec suffisamment de précision pour être intéressante, nous ne retenons que de critère de la présence ou non d'un matériel d'irrigation, permettant d'y recourir si besoin. Cette information est donc considérée comme invariable par parcelle. Ce sont évidemment les données invariables uniquement qui sont utilisées lors de la reconnaissance puisqu'il s'agit de conseiller l'agriculteur quant à ses pratiques culturales, dont l'efficacité dépendra directement des caractéristiques inhérentes à sa parcelle, en lui fournissant les différents scenarii auxquels le mèneraient ses stratégies.

L'information du rendement n'est pas prise en compte pour l'apprentissage, il ne s'agit que d'une donnée "d'aide à la décision" permettant à l'utilisateur de choisir entre les différents scénarii qui lui seront proposés : elle décrit le profil et fait partie intégrante de sa définition, mais pas du processus de création du profil. Elle est mise à jour systématiquement - en même temps que les autres caractéristiques variables du profil. Il est donc primordial, pour aider au mieux les agriculteurs à prendre des décisions informées, de disposer de cette information au moment de l'apprentissage et, en fin de campagne, lors de la mise à jour.

1.2.2.3 La partie apprentissage

L'apprentissage a été réalisé sur un échantillon de parcelles situées en Midi-Pyrénées, où l'irrigation est utilisée par 27% des exploitants agricoles et la culture du maïs grain représente 60% des surfaces irriguées. Cette étape vise à définir les différents profils, à partir des informations rassemblées sur parcelles, chaque profil étant établi en fonction des sept caractéristiques sus-indiquées. Le but final consiste à proposer à l'agriculteur, au moment de la reconnaissance, les couples liant une date de semis et une précocité de grain préconisés pour son type de parcelle et le renseigner sur le rendement qu'il peut espérer s'il venait à calquer ses pratiques culturales sur le couple proposé ; aussi l'apprentissage doit-il associer, comme le montre la figure 2, un duo précis à chaque profil, ainsi qu'un intervalle de rendement.

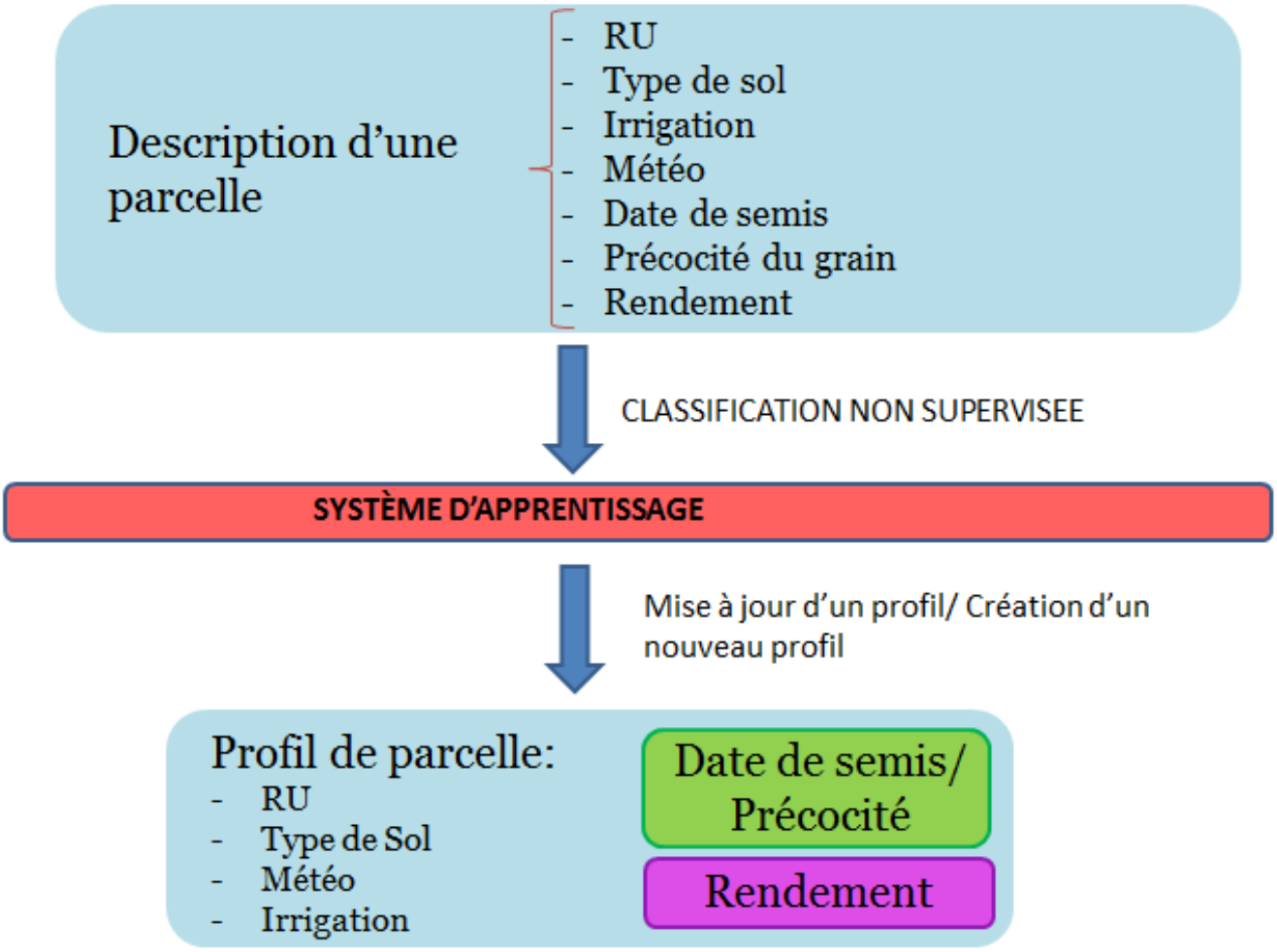


Figure 2. Description du fonctionnement du système d'apprentissage

Pour la définition des profils, chaque parcelle doit être décrite par les sept caractéristiques - qu'elles interviennent dans le processus d'apprentissage ou non, de manière à bénéficier de la plus grande précision possible pour la définition des profils.

L'apprentissage se fait au moyen d'une classification non supervisée, que nous décrirons plus précisément dans les prochains chapitres. Chaque parcelle de l'échantillon d'entraînement associe une date de semis à une précocité de grain. Au moment de l'apprentissage, s'il existe un profil décrit par le même couple exactement, et que ce profil est suffisamment proche de la parcelle, alors le profil est mis à jour. Dans tous les autres cas, c'est-à-dire si aucun profil ne présente la même association ou si aucun profil présentant la même association ne ressemble à la parcelle, alors un nouveau profil est créé.

1.2.2.4 La partie reconnaissance

La reconnaissance, dont le principe de fonctionnement est schématisé sur la figure 3, prend en entrée les informations relatives à une parcelle, décrites par les quatre caractéristiques décrites, et propose en sorties les différents couples associant une date de semis à une variété conseillés, assortis pour chacun du rendement espéré et d'un indice de confiance.

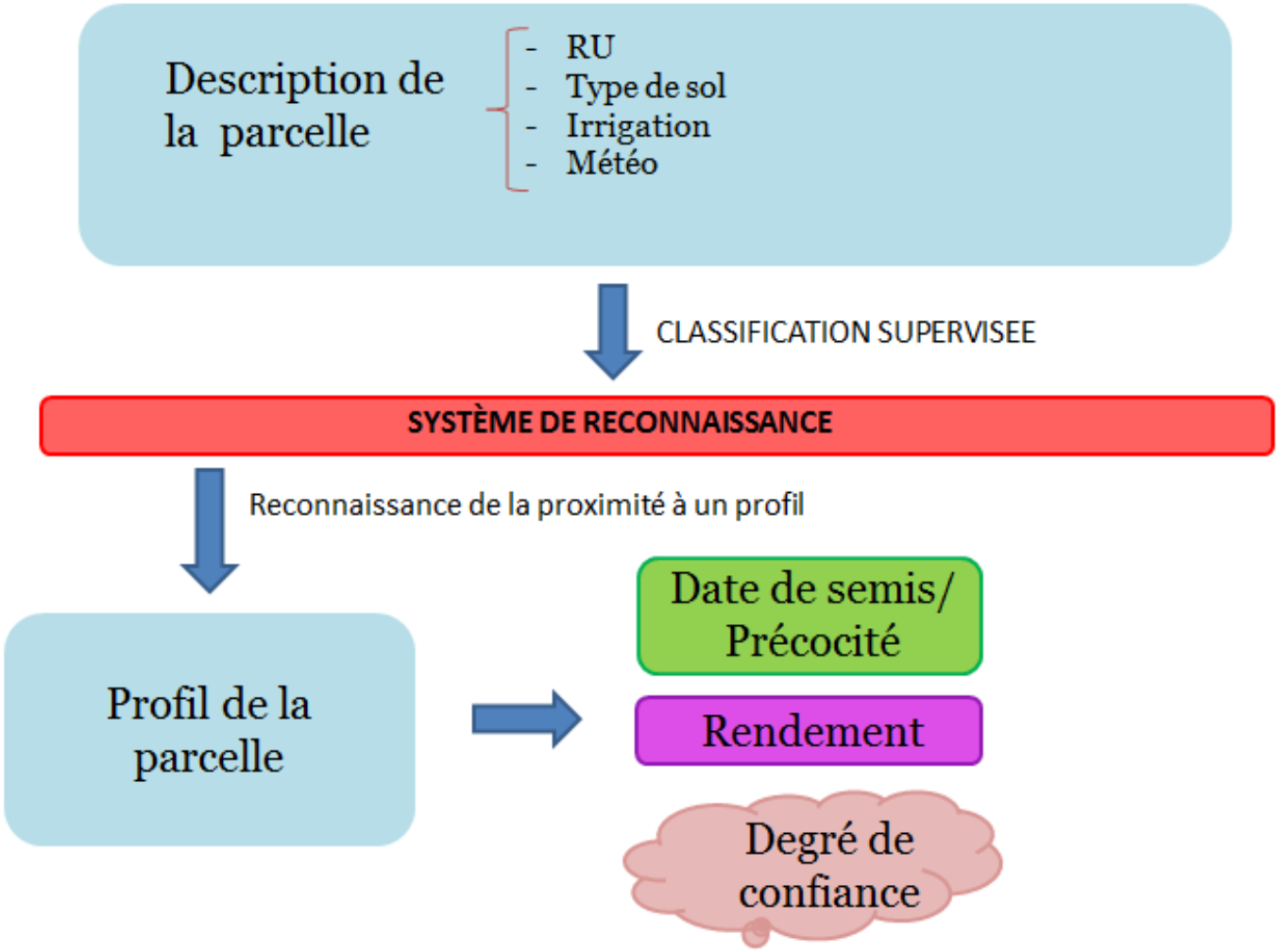


Figure 3. Description du fonctionnement du système de reconnaissance

1.3 Classification de zones agricoles pour une meilleure gestion de l'eau d'irrigation

1.3.1 Objectifs

A une époque où le soucis de préserver l'eau pour les générations futures préconise une politique d'économie, la question de la gestion du partage de l'eau s'avère constituer un problème crucial ; il s'agit d'optimiser les usages en répartissant équitablement la ressource en eau entre les différents besoins (naturels, agricoles, économiques, humains) en s'adaptant aux habitudes existantes et aux spécificités de chaque secteur.

Une grande partie de l'eau destinée à l'irrigation est prélevée dans les eaux superficielles, principalement pour une question d'accessibilité, mais aussi pour éviter des chocs thermiques aux plantes. La concurrence avec les autres utilisations de l'eau et le maintien d'un débit minimal résiduel pour assurer les diverses fonctions des cours d'eau (p.ex. habitat pour la faune et la flore aquatique, couloir à faune, structuration du paysage, ou dégradation de polluants) peuvent alors conduire à des situations critiques voire conflictuelles. Le risque de pénurie en eau durant les mois d'été à l'échelle des bassins versants augmente, en outre, considérablement avec le changement climatique. Par conséquent, préserver l'eau en assurant une répartition équilibrée et ajustée aux besoins de chaque zone apparaît maintenant primordial. Plus précisément, pour les agriculteurs de maïs, les besoins en eau peuvent sensiblement varier en fonction de différents critères pédologiques, géographiques, et météorologiques. Le but de ce volet

du projet MAISEO est donc d'assurer l'apport nécessaire et minimal d'eau à chaque exploitation de manière à ce que les plants atteignent leur potentiel de rendement, en tenant compte des spécificités de leur zone géographique [Hunger, 2010]. En d'autres termes, il s'agit d'apporter aux irrigants, aux gestionnaires des ouvrages, et à l'administration chargée du contrôle, les éléments d'appréciation de la meilleure stratégie de gestion de la demande en eau ayant lieu en cours de campagne, et d'anticiper des décisions concernant l'autorisation ou l'interdiction d'irriguer potentiellement lourdes de conséquences pour les rendements des cultures. A terme, le projet MAISEO doit permettre aux compagnies de gestion de l'eau de connaître a priori les zones d'exploitation nécessitant une alimentation en eau d'irrigation, les quantités respectives, de manière à établir une juste prédiction, en recoupant ces informations avec les données météorologiques prévisionnelles, des besoins à venir et d'anticiper les problèmes issus de la politique de restriction d'arrosage sur un bassin entier.

1.3.2 La problématique de gestion de l'eau avant le projet MAISEO

Les besoins en eau des différentes exploitations peuvent être estimés de différentes manières. La plus simple - mais fastidieuse et à la fiabilité inégale - est l'enquête auprès des agriculteurs, associée à une étude météorologique prévisionnelle. Les questionnaires soumis aux exploitants traitent des infrastructures et équipements disponibles actuellement ou attendus, la surfaces des différentes parcelles à irriguer, l'historique des quantités d'eau utilisées, les facteurs limitants, les stratégies d'adaptation,... Les résultats des enquêtes peuvent ainsi permettre de dresser une carte représentant les besoins en eau estimés par zone. Néanmoins, cette méthode est tributaire de la bonne foi des agriculteurs, de la précision de leurs réponses, et aussi du nombre de questionnaires renvoyés - car bon nombre d'entre eux n'y répondent pas, malgré les sollicitations [Fuhrer, 2010]. Pour davantage de précision et de certitude vis-à-vis des données accessibles, et aussi une répartition plus homogène et complète de ces informations, il est donc préférable d'associer à ces questionnaires des données obtenues à partir de modèles experts et d'observations systématiques sur l'ensemble de la zone à analyser.

Une estimation des besoins hydriques peut aussi être obtenue à partir de calculs décrits par Fuhrer [Fuhrer et Jasper, 2009] et mis en pratique par des modèles, cherchant à simuler la dynamique et le régime de l'eau du sol de la manière la plus proche possible de la réalité. Ces calculs s'appuient sur l'estimation de l'évapotranspiration. Plus précisément, il s'agit d'estimer l'humidité moyenne du sol à la profondeur d'enracinement en tenant compte de la différenciation par surface, ainsi que le rapport entre l'évapotranspiration actuelle et potentielle (ET/ETP). La quantité d'eau d'irrigation nécessaire est estimée relativement à la différence entre la valeur cible - correspondant à l'état d'humidité du sol à partir de laquelle la transpiration actuelle diminue par rapport à la transpiration potentielle - et la valeur actuelle de l'humidité du sol à la profondeur d'enracinement. Il est important de faire la différence entre l'ET qui décrit la quantité de vapeur d'eau transférée dans l'atmosphère par transpiration des plantes et par évaporation de l'eau contenue par le sol, par des retenues d'eau -naturelles ou non, ... et l'ETP définie comme étant la vapeur maximale d'évapotranspiration d'un couvert végétal continu lorsqu'il y a suffisamment d'eau disponible dans le sol pour satisfaire la demande évaporatrice de l'atmosphère [Bouchet].

D'autres modèles, orientés vers une approche basée sur le bilan hydrique du sol, prennent en compte le type de culture et son évolution du stade phénologique, les caractéristiques texturales et hydriques du sol (capacité au champ, mouvement capillaire, point de flétrissement), et la météorologie [et al., 2011].

Certains logiciels, basés sur ces modèles, permettent d'assurer un conseil à l'irrigation précis et efficace. [Pepin et Bourgeois, 1992]

Néanmoins, ces modèles demandent un niveau de détail vis-à-vis des données en entrée qu'il est souvent difficile d'obtenir, en particulier lorsqu'il s'agit de gérer la distribution de l'eau sur des zones étendues. Il demeure bien souvent des incertitudes relatives aux prévisions climatiques, aux propriété-

tés pédologiques et aux interventions anthropogènes (prélèvements d'eau, exploitation des lacs et des réservoirs,...) Ainsi une analyse suffisamment fine nécessiterait l'emploi de données spécifiques à la culture du maïs et précisément connues pour chaque parcelle, ce qui ne semble pas réalisable dans le cadre de ce projet - et une étude grossière ne s'avérerait pas véritablement discriminante ; c'est pourquoi dans le projet MAISEO, la question de la distribution de l'eau n'a pas vocation à être étudiée à l'échelle parcellaire mais en traitant des îlots constitués de plusieurs exploitations. Il s'agit donc d'analyser les besoins en eau de zones, à partir de caractéristiques objectives et facilement accessibles par enquêtes ou données satellites.

1.3.3 Fonctionnement global du système

Pour ce lot, nous avons collaboré avec la CACG. La Compagnie d'Aménagement des Coteaux de Gascogne est une société qui conçoit, construit, et met en œuvre des projets dans le but de concourir à l'aménagement du territoire. Dans ce cadre, ses principaux clients sont les collectivités territoriales, les agriculteurs et les entreprises privées. Elle gère notamment le canal de la Neste, qui alimente la grande majorité des rivières de Gascogne. Cette mission s'inscrit dans le cadre d'une concession d'état.

Notre implication dans cette partie du projet réside dans l'importance d'établir une classification des besoins en eau d'îlots d'exploitation pour une région donnée. L'outil permet de définir, comme représenté dans la figure 4, à partir d'un certain nombre de caractéristiques, des profils de zones afin d'évaluer la quantité d'eau d'irrigation nécessaire à chacune d'elles pour la culture du maïs.

La définition des profils s'est opérée - au même titre que le lot précédent - au moyen d'une classification non supervisée, basée un ensemble d'individus décrits par six caractéristisques. Chacune d'elles a été obtenue à partir de mesures précises basées sur des observations satellites ou des observations de terrain. Toutes les caractéristiques décrivant les données d'entraînement sont impliquées dans le processus de définition des profils et dans la reconnaissance. Ce sont les experts de la CACG qui, par la suite, analysent les différents profils obtenus afin d'associer à chacun d'eux une estimation des besoins en eau.

Les caractéristiques sont les suivantes :

- La distance au cours d'eau le plus proche et géré par la CACG. Elle est exprimée en mètres.
- L'altitude entre îlot et la retenue collinaire la plus proche. Elle permet d'estimer la difficulté rencontrée par les agriculteurs pour irriguer leur parcelle par eux-mêmes. Elle est exprimée en mètres.
- La distance entre l'îlot et le réseau en concession d'état. Elle est exprimée en mètres.
- Le type du point de prélèvement. Cet attribut a son importance car il implique des différences en terme de température et de qualité de l'eau. Ces écarts trouvent leur explication dans des caractéristiques physiques que sont la profondeur, la pente de la berge, le renouvellement de l'eau, la densité et la nature du fond. Tous ces éléments sont reliés à la genèse du plan d'eau. Ici, il peut s'agir d'un barrage, d'une gravière, ou d'un DCE (Directive Cadre Européenne sur l'Eau). Un plan d'eau DCE fait l'objet de suivis biologique, physico-chimique et chimique bien définis, et doit témoigner d'un respect de normes et de valeurs-seuil formellement établies.
- La pente. Elle joue un rôle important dans la capacité du sol à retenir l'eau pour les plantes. Plus le sol sur lequel se trouve l'exploitation est pentu, plus l'eau aura tendance à s'écouler en aval. Elle est exprimée en degrés.
- L'altitude de la parcelle. Elle est exprimée en mètres.
- L'orientation. Elle influe sur le temps et le degré d'exposition de la plante au soleil, et donc sur le rayonnement subi et l'évapotranspiration. Elle est exprimée en fonction des points cardinaux.

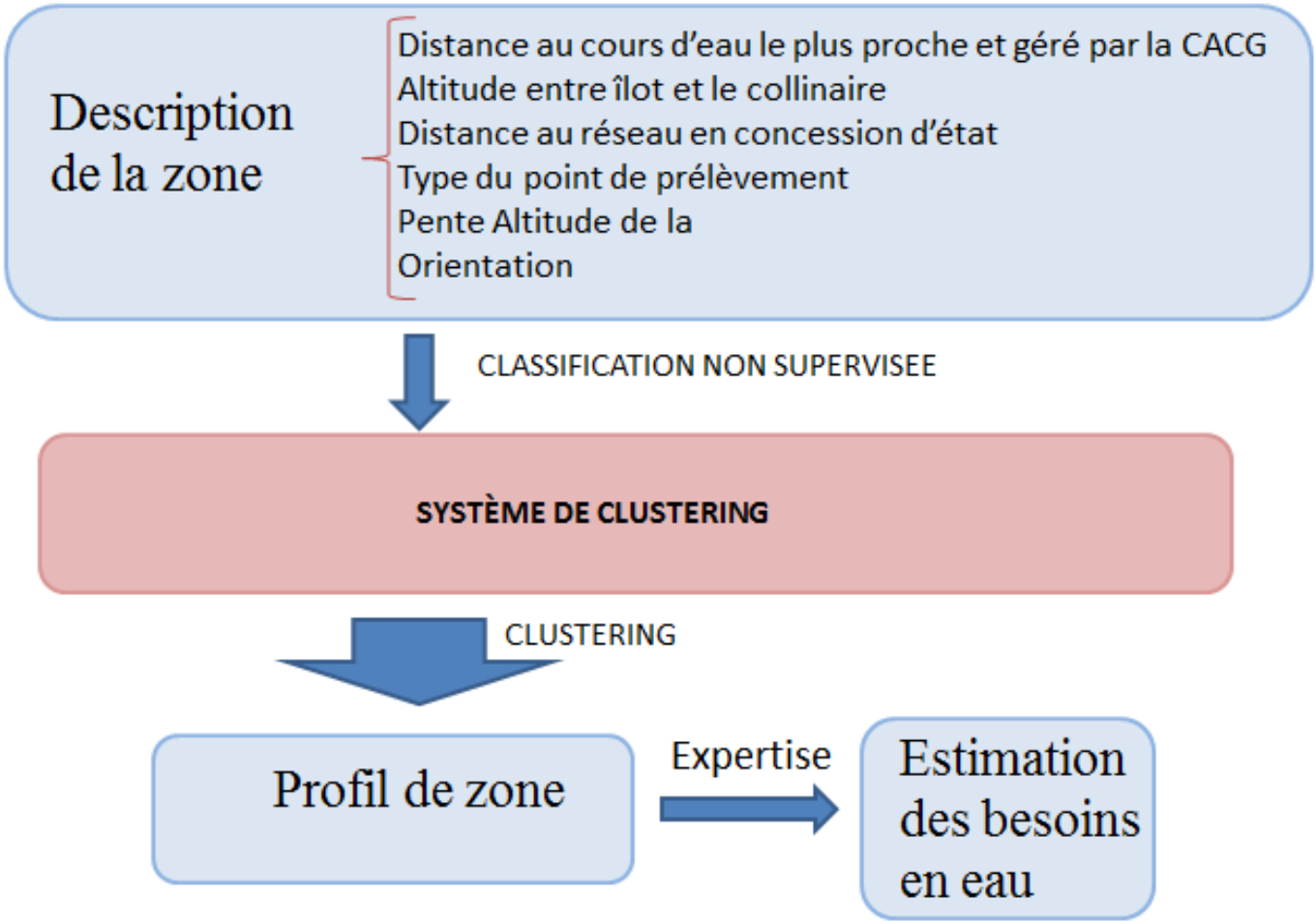


Figure 4. Description du fonctionnement du système

1.4 Conclusion

L'initiative du projet MAISEO a été motivée par la nécessité de produire davantage de maïs dans un contexte de ressource en eau limitée. Pour cela, il doit intervenir à différents niveaux : le conseil personnalisé aux agriculteurs et une gestion territoriale plus efficace de l'eau.

Le diagnostic précampagne vise à caractériser en temps réel le potentiel maïs d'une parcelle en irrigué ou en sec dans les conditions climatiques et hydriques de l'année à travers différents scénarii assortis d'un degré de probabilité de réalisation. Les résultats de ce diagnostic doivent permettre de définir les meilleures stratégies de semis, c'est-à-dire de déterminer quels sont les choix optimaux concernant les précocités, les variétés, des dates, et les densités de semis pour l'année et les objectifs de l'agriculteur. En particulier, notre implication réside dans le développement d'un module de diagnostic précampagne permettant la sélection, la fusion, et le traitement de données hétérogènes en provenance des différentes bases (météorologique, réserve utile, parcelle, zonage intra-parcellaire..). La caractérisation et l'évolution du niveau de fiabilité de l'information est un des paramètres clés de ce module car les données d'entrées ont des niveaux d'incertitudes différents en fonction des sources (obervations par télédétection, mesures de terrain, statistiques, modèles,...) - certaines données prépondérantes peuvent même venir à manquer au moment de la reconnaissance.

Concernant le deuxième aspect de MAISEO, l'objectif est de développer de nouveaux systèmes d'informations permettant d'établir des stratégies de gestion de l'eau basées sur l'analyse de données complètes et représentatives de la demande hydrique du territoire pour l'année. Avant ce projet, le gestionnaire du bassin versant, ne disposait principalement que de données statistiques sur la ressource

hydrique et d'une connaissance sommaire des besoins en irrigation et de données de terrain très partielles et éparpillées. Il s'agit donc de fusionner des informations auparavant dispersées et de les classer de manière à définir la localisation des surfaces irriguées avec une bonne probabilité d'avoir raison. Cette classification doit permettre le diagnostic hydrique du bassin versant à l'époque des semis des cultures d'été de mars à mai, de manière à anticiper les besoins hydriques du bassin en relation avec des prévisions de ressources et de permettre ainsi les prises de décisions pour prévenir les risques de défaillance - c'est-à dire la non satisfaction des DOE ou des besoins d'irrigation.

Dans le prochain chapitre, nous présenterons les différentes méthodes de classification existantes et expliquerons sur quels critères s'est basée notre sélection de celle qui allait nous assister dans le projet MAISEO.

2 Outil de classification pour le diagnostic de parcelles de maïs

Diagnostic provient du grec διαγνωοη : dia (à travers) - gnôsè (la connaissance, le jugement). Un diagnostic désigne une conclusion, généralement prospective, faisant suite à l'examen analytique d'une situation souvent jugée critique ou complexe : c'est par les signes d'une situation qu'on la juge, qu'on en acquiert la connaissance. L'enjeu est généralement d'identifier l'origine d'une défaillance, d'un dysfonctionnement, ou d'une complication. En intelligence artificielle, le diagnostic est une discipline visant au développement d'algorithmes permettant de déterminer si le comportement d'un système est conforme au comportement espéré. Dans le cas contraire, l'algorithme doit définir le problème avec précision et en déterminer les causes. Ainsi, un outil de diagnostic est un moyen employé pour obtenir cette connaissance ; son rôle est d'analyser le comportement d'un système et de le comparer à ce qui en est attendu.

Les deux principaux types d'algorithmes sur lesquels sont basés les diagnostics dans le champ de l'intelligence artificielle sont ceux basés sur le modèle, et ceux orientés données. Dans le premier cas, il s'agit de comparer les sorties d'un modèle théorique du système, décrivant le comportement attendu, aux observations fournies par le système opérant, de manière à détecter les potentiels dévoiements. Dans le deuxième cas - et c'est celui qui nous intéressera dans le cadre de ce projet - les mesures des systèmes sont regroupées dans des classes de manière à rassembler les données ayant des caractéristiques communes.

Dans le cas de la première partie de notre projet, le diagnostic ne vise pas qu'à analyser le comportement de la parcelle, mais aussi à prévoir ce qu'il devrait être. En effet, en fonction de la catégorie à laquelle appartient la parcelle, il est possible d'extrapoler sa productivité et son besoin en eau. L'objectif de notre outil de diagnostic est donc de classer cette parcelle dans la catégorie adéquate de manière à mieux la "connaître" dans notre contexte, c'est-à-dire à anticiper sa conduite pour être en mesure de fournir à l'agriculteur les conseils appropriés.
Notre tâche se découpe en plusieurs étapes :

1. L'apprentissage des différentes catégories auxquelles peut appartenir une parcelle à partir d'un échantillon de données,
2. La classification des parcelles à diagnostiquer dans les catégories apprises,
3. L'évolution des catégories en fonction des réponses des parcelles aux traitements appliqués.

En ce qui concerne la seconde partie, il nous faut être en mesure d'anticiper les besoins en eau d'irrigation d'îlots d'exploitations. Pour cela, nous devons les diagnostiquer et déterminer quel est leur profil hydrique. Il s'agit là encore d'anticiper la conduite de ces îlots de manière à fournir -non plus à l'exploitant - au gestionnaire du bassin versant une estimation des quantités d'eau d'irrigation à allouer à chaque îlot de sa zone.

Notre travail implique donc plusieurs aspects, de sorte que toutes les méthodes de classification de données ne sont pas appropriées à ce traitement. Il est donc important, dans un premier temps, d'établir

quelle méthode serait la plus adaptée.
Dans ce chapitre, après avoir défini les enjeux d'une classification, nous décrivons différentes méthodes et expliquons les raisons de notre choix.

2.1 Les enjeux de la classification

Un problème de classification doit être étudié sous trois angles particuliers :

1. La fréquence relative des classes de la population concernée, exprimée formellement par la probabilité de distribution,
2. Un critère implicite ou explicite de séparation des classes, c'est-à-dire une relation qui prend en compte les différents attributs pour déterminer l'appartenance de chacun des individus à sa classe,
3. Le coût associé à une mauvaise classification. Dans le cas du diagnostic, une erreur de classification peut être très grave : par exemple, il peut s'agir de la classification d'un individu déficient dans la classe des individus normaux (faux négatif) ou, inversement, de la classification d'un individu normal dans la classe des individus déficients (faux positif).

Théoriquement, chacune de ces trois questions doit être examinée individuellement et les résultats combinés formellement en une règle de classification. Pourtant, la majorité des techniques employées les prend simultanément en compte et fournit une règle en accord avec la distribution particulière, de sorte qu'elles ne peuvent pas s'adapter facilement à des changements dans la fréquence des classes.

Dans notre étude, le diagnostic ne concerne pas la détection d'individus déficients mais la description de parcelles pour anticiper au mieux leur comportement. La différence entre ces deux objectifs est importante car, dans le premier cas, il n'y a généralement que deux classes en jeu (la classe des individus normaux et la classe des individus déficients) tandis que dans le second, le nombre de classes est a priori inconnu.

2.2 Les méthodes de classification

La nomenclature des méthodes de classification distingue deux types précis de classification : supervisée ou non-supervisée. Dans le premier cas, la difficulté est d'établir une règle permettant de classer un nouvel individu dans des classes préalablement définies, tandis que dans le second il s'agit de déterminer ces classes. Dans notre projet, nous aurons besoin des deux à des stades différents du traitement des données.

2.2.1 La classification supervisée

La classification supervisée est employée dans un contexte pré-défini, c'est-à-dire lorsque les classes sont connues et précisément définies a priori. Il s'agit ainsi d'assigner chacun des individus d'un échantillon à la classe dont il est le plus proche [Duda *et al.*, 2012]. Pour cela, les caractéristiques des individus de l'échantillon doivent être similaires aux caractéristiques des classes. L'enjeu est donc la reconnaissance des individus traités : il s'agit de les décrire par la classe à laquelle chacun d'eux appartient et de leur attribuer les propriétés de cette classe. Ces classes peuvent être par exemple définies par un expert, et les outils de classification se basent sur des algorithmes variés pour parvenir à une décision.

2.2.1.1 Les Réseaux de Neurones Artificiels (RNA)

L'intelligence artificielle trouve sa source chez les philosophes classiques, dont Leibniz (1646-1716). Le but du calculus ratiocinator [Leibniz] était de décrire le processus de la pensée humaine et sa concrétisation nous donnera l'ordinateur programmable dans les années 1940. Dès le départ, deux approches se confrontent :

– L'approche logiciste ou symbolique, qui vise à recréer les « lois universelles » de la pensée et s'inspirent du concept de machine de Turing,
– L'approche neuronale qui essaie d'imiter les processus biologiques cérébraux..

Suivant cette nomenclature, la méthode des réseaux de neurones se situe clairement dans l'approche neuronale : ses principes trouvent leur source dans l'observation du système neurologique humain. Il s'agit d'une technique non-linéaire impliquant des noeuds inter-connectés et communiquant directement entre eux [McCulloch, 1943]. En biologie, un neurone est une cellule excitable assurant la transmission d'un signal bioélectrique par conductivité. Un neurone convertit les stimulations en impulsions nerveuses et les transmet aux neurones voisins. La variation locale que subit un neurone, correspondant à une phase de dépolarisation et une phase de repolarisation de son axone, s'appelle le potentiel d'action. Chaque neurone a un potentiel seuil, de sorte que le potentiel d'action suit la loi du "tout ou rien".

La méthode des réseaux de neurones réinvestit ce principe : basé sur une entité appelée perceptron, dont le principe de fonctionnement est représenté sur la figure 5, un RNA prend en entrée un vecteur de valeurs réelles qu'il traite par combinaison linéaire [Vapnik et Naumovich, 1998]. Les perceptrons représentent une surface de décision en hyperplan dans un espace d'exmples en n-dimensions. Un RNA se compose de plusieurs couches de ces entités [Rosenblatt, 1962], chaque couche transférant ses informations à la couche qui lui est immédiatement supérieure. Les perceptrons d'une couche sont reliés à ceux de la couche précédente et traitent les données qu'ils leur transmettent ; le résultat est ensuite transmis à un module intégrant une fonction d'activation f. Le résultat obtenu décrit l'état interne du perceptron. Le nombre de couches dépend du RNA et de ses besoins. La dernière couche du réseau fournit le résultat de classification. La représentation schématique d'un RNA est visible sur la figure 6.

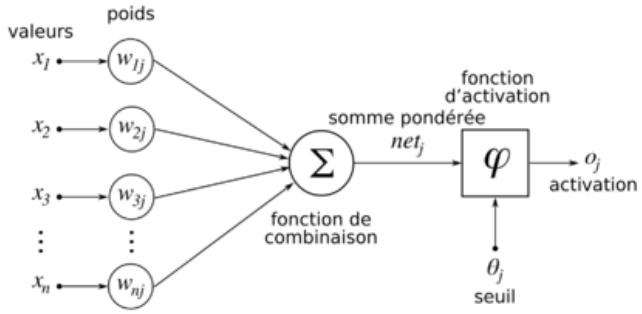


Figure 5. Schéma d'un neurone

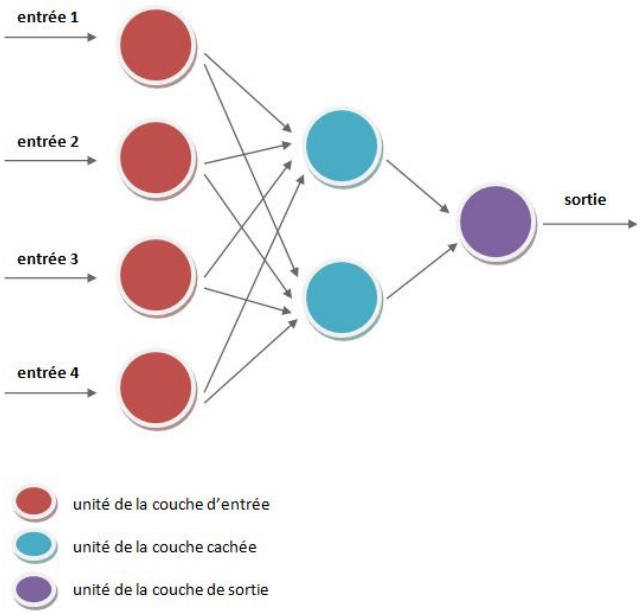


Figure 6. Vue simplifiée d'un réseau artificiel de neurones

Ainsi, pour un neurone, la première étape consiste en l'agrégation des différents signaux en entrée en une valeur v telle que $v = \zeta(x, w)$, imitant ainsi un neurone biologique puisqu'il permet ainsi de projeter des signaux en une réponse interne. Ensuite intervient une fonction d'activation θ, qui transforme la valeur obtenue en une valeur de sortie o telle que $o = \theta(v)$, simulant ainsi le comportement d'un neurone biologique lorsqu'il décide de renvoyer ou d'inhiber un signal, selon sa logique interne. La valeur de sortie est par la suite propagée à toutes les unités suivantes, telle qu'elles ont été déterminées par la topologie particulière. Le tableau 1 présente les différents types de fonctions d'activation, associés aux types de neurones correspondant. Les fonctions d'activation employées peuvent varier d'un RNA à un autre : fonction gaussienne, sigmoïde, fonction tangente hyperbolique,... en fonction du type de traitement désiré. Les plus courantes sont représentées sur la figure 7.

Type de Neurones	Fonction d'aggrégation	Fonction d'activation
Linéaire par palliers	Somme pondérée	Fonction par échelons ou signes
Linéaire	Somme pondérée	Fonction linéaire ou linéaire par morceaux
Sigmoïde	Somme pondérée	Fonction sigmoïde ou tanh
Distance	Distance	Fonction linéaire ou linéaire par morceaux
Gaussien	Distance	Noyau gaussien

Tableau 1. Types de neurones les plus utilisés dans les RNA

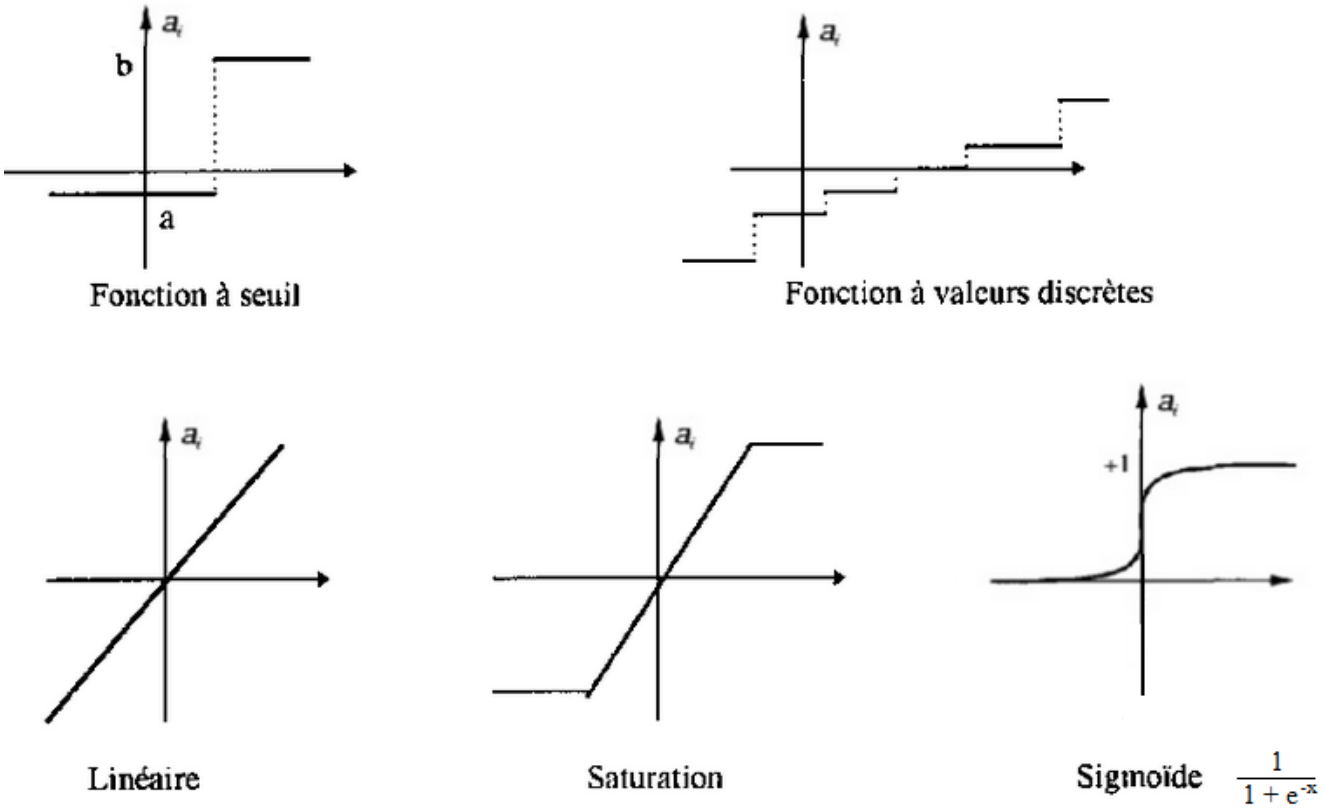


Figure 7. Fonctions d'activation les plus courantes

Les RNA sont donc des méthodes de classification multicouches puisque si, à l'échelle du neurone, les calculs sont effectués en deux étapes, à l'échelle du réseau, le traitement est découpé en autant de moments qu'il existe de couches de neurones. Cette approche multicouche est intéressante car elle permet des opérations complexes et un traitement en profondeur de l'information entrante, l'activation se

propageant à travers le réseau à partir de la couche d'entrée jusqu'à celle de sortie, de façon séquentielle. Selon la connectivité du réseau, le processus d'activation et de propagation peut être implémenté dans un système distribué parallèle ou un PDP. [Rumelhart *et al.*, 1986]

Dans les RNA sans "feedback" - c'est-à-dire sans retour de l'information de sortie vers l'entrée - le processus d'activation et de propagation se termine lors de la production des signaux de sortie. A l'inverse, dans les RNA recevant un feedback, tels que les réseaux récurrents, la propagation ne s'achève jamais et suit une trajectoire dynamique à travers l'espace d'état, de sorte que les neurones et les différentes couches soient continument mis à jour.

L'état du réseau est représenté par les valeurs de tous les poids pour tous les neurones. Les valeurs des poids, les connections entre les neurones, et le type des fonctions d'agrégation et d'activation correspondent au mécanisme employé par le réseau pour encoder la connaissance apprise à partir des données d'entrée. L'apprentissage par RNA équivaut à l'estimation de l'ensemble des poids qui concluent la tâche de classification de manière optimale, dans le but de minimiser le risque d'erreur de classification. Chaque type de RNA fournit un schéma d'apprentissage spécifique, adéquat à sa topologie particulière. La mise à jour du poids w_{ij} entre un neurone j recevant un signal x_i d'un neurone i et produisant un signal o_j doit refléter l'importance de l'interaction des deux neurones dans la production du signal en sortie du réseau, de sorte que :

$$\delta w_{ij} = F(x_i, o_j, w_{ij}, \nu)$$

avec

- ν : Paramètre de taux d'apprentissage préalablement défini,
- F() : Fonction multiplicative définie.

Les réseaux de neurones sont particulièrement adaptés aux données dont l'échantillon est bruité ou pour des problèmes dont la représentation est symbolique. Ils peuvent également fonctionner en mode non supervisé.

Le problème principal, pour notre situation, des RNA réside dans leur opacité : il est très difficile d'obtenir une représentation claire des connaissances, puisque la structure interne du réseau, les fonctions ainsi que les poids, sont cachés. Les résultats obtenus peuvent alors être difficiles à être interprétés a posteriori par un expert. En outre, il nécessite une base d'apprentissage conséquente pour déterminer les poids des connections entre les différents neurones des couches successives - ce sont les paramètres à identifier au même titre que les paramètres d'un modèle. Par conséquent il est nécessaire de disposer d'assez d'exemples sinon on se trouve devant une impossibilité de déterminer ces paramètres (phénomène de sur-paramétrisation)

2.2.1.2 Les arbres de décision

Les arbres de décision, comme représentés sur la figure 8, sont des outils très souvent utilisés pour la classification ou la prédiction. Leur fonctionnement est très intuitif, et leur comportement est décrit par des règles précises, de sorte que les résultats sont aisément analysables. Ils présentent également l'avantage de pouvoir traiter les données tant qualitatives que quantitatives. Il existe en outre des techniques pour traiter des données incomplètes.

Ce sont des outils de classification dont la structure est arborescente, c'est-à-dire constituée de branches, de noeuds, et de feuilles. A chaque noeud s'applique une règle de décision, de manière à ce que l'individu soit orienté, en fonction de ses caractéristiques, vers la branche qui lui correspond. Il sera ainsi amené au noeud suivant, qui le traitera et le dirigera à son tour sur une des branches qui émergent de lui, et ainsi de suite jusqu'à ce que l'individu ne rencontre plus un noeud mais une feuille. A chaque feuille correspond une classe précise. Ainsi, pour chaque noeud, la règle à appliquer doit être connue et définie au préalable.

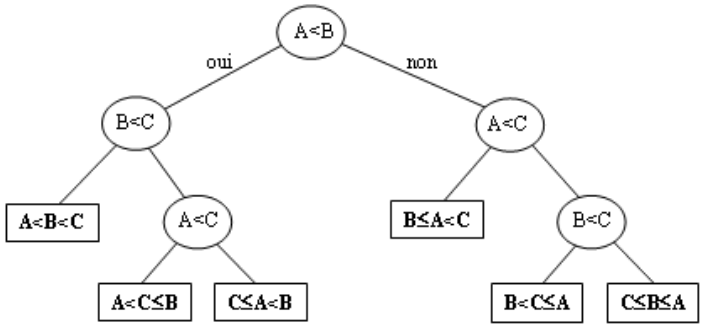


Figure 8. Exemple d'arbre de décision

La construction d'un arbre de décision n'est pas difficile, le problème repose le plus souvent sur la question de l'optimisation, qui réside dans deux aspects principaux :

- Rapidité de la construction de l'arbre,
- Rapidité dans la prédiction.

L'arbre doit également assurer des prédictions justes et robustes malgré le bruit ou les données manquantes. Le nombre d'arbres possibles augmente exponentiellement avec le nombre d'attributs et de valeurs distinctes pour chacun d'eux. Par exemple, si un arbre doit tester d attributs entre son tronc et ses feuilles, d ! chemins différents peuvent être établis. La construction d'un arbre de décision binaire qui permet une classification correcte d'un ensemble de N individus tel que le nombre d'étapes pour classer un individu soit minimal est NP-complet. [Buhrman et De Wolf, 2002] [hya, 1976]

Il existe de nombreux algorithmes employés pour la construction des arbres de décision, dont le plus connu est l'algorithme CART [Breiman *et al.*, 1984]. Il se base sur le principe de l'entropie de manière à assurer la meilleure homogénéité des classes ; plus un noeud est discriminant, plus il se situe près de la base de l'arbre.[Breiman, 1996] L'entropie de Shannon est une fonction mathématique qui, intuitivement, correspond à la quantité d'information contenue ou délivrée par une source d'information. Elle est caractérisée par le fait que :

- Le minimum de la fonction est atteint lorsque tous les nœuds sont purs,
- Le maximum de la fonction est atteint lorsque les individus sont équi-répartis entre les classes.

Elle est définie selon la formule :

$$Entropie(p) = -\sum_{c=1}^{C} P(c|p) ln P(c|p)$$

avec

- N(p)= Nombre d'individus associés au nœud p,
- N(c | p)= Nombre d'individus appartenant à la classe c sachant qu'ils appartiennent au nœud p,
- $P(k|p) = \frac{N(c|p)}{N(p)}$ = Proportion des individus appartenant à la classe c parmi ceux du nœud p.

Ainsi, un nœud est pur si tous les individus d'un même nœud appartiennent à la même classe, donc si P(c | p)=1 -auquel cas il s'agit d'une feuille. L'algorithme s'arrête lorsque tous les noeuds sont purs.

Les arbres de régression sont une généralisation des arbres de décision : leur sortie est représentée par une valeur réelle appartenant à un ensemble continu, au lieu d'une valeur qualitative.

Les classes produites par les arbres de décision sont homogènes, et c'est là que réside leurs limites par rapport à notre projet : ses classes s'excluent mutuellement, ce qui est requis dans certains cas mais ne représente pas la réalité du nôtre. En effet, les parcelles de maïs ont des données très variées, il n'existe

pas de règles définitives et systématiques permettant de les classer dans une catégorie avec certitude, et chaque parcelle peut appartenir à plusieurs classes dans une certaine mesure. Notre objectif n'est donc pas de déterminer l'appartenance absolue d'une parcelle à une classe mais de l'assigner à la classe à laquelle elle ressemble le plus, tout en conservant les informations relatives à sa similarité avec les autres classes.

2.2.1.3 Les réseaux bayésiens

La classification bayésienne est une approche probabiliste basée sur les probabilités conditionnelles - et la règle de Bayes. Elle suppose l'indépendance des attributs et nécessite des connaissances a priori, de manière à prévoir le futur à partir du passé. Il s'agit d'estimer la probabilité d'occurrence d'un événement sachant qu'une hypothèse préliminaire est vérifiée, pour ensuite se baser sur ces probabilités pour guider l'inférence. Ainsi, chaque hypothèse se voit associer une probabilité décrivant la chance d'être la solution, sachant que l'observation d'une ou plusieurs instances peut modifier cette estimation. Aussi, au vu des instances, parlera-t-on d'hypothèse la plus probable. Le but est de conserver cette hypothèse et de rejeter les autres, tout en les conservant pour un traitement ultérieur. Le théorème de Bayes est décrit par les formules :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(B|A)P(A)$$

Donc :

$$P(A|B) = fracP(B|A)P(A)P(B)$$

Ce théorème s'applique à la classification selon les termes :

$$P(k|x_1, ..., x_n) = P(x_1, ..., x_n|k)P(k)/P(x_1, ..., x_n)$$

où

- $P(x_1, ..., x_n|k)$, $P(x_1, ..., x_n)$ et P(k) peuvent être estimées sur les instances de l'ensemble d'apprentissage telles que
  - $P(x_i|k) = n_{ik}/n_k$,
  - n_{ik} = nombre d'instances de la classe c qui ont comme valeur x_i pour l'attribut considéré,
  - n_c = nombre d'instances de la classe.
- $P(k|x_1, ..., x_n) = \frac{P(x_1,...,x_n|k)P(k)}{P(x_1,...,x_n)}$ avec
  - P(k) : proportion d'instances de la classe k,
  - $P(x_1, ..., x_n)$: proportion d'instances des attributs $(x_1, ..., x_n)$,
  - $P(x_1, ..., x_n|k)$: nombre de fois où on rencontre $(x_1, ..., x_n)$ dans les instances de la classe k - c'est-à-dire vraisemblance de la classe.

Ainsi, pour un individu j à classer, l'algorithme suit les étapes suivantes :

1. Estimer P(j | k)·P(k) pour chaque classe k,
2. Affecter j à la classe k telle que la probabilité P(j | k)·P(k) est la plus grande

On observe que :

- $P(k|x_1, ..., x_n)$ croît quand P(k) croît : plus k est probable, plus il y a de chances qu'elle soit la bonne classe,
- $P(k|x_1, ..., x_n)$ croît quand $P(x_1, ..., x_n|k)$ croît : plus $(x_1, ..., x_n)$ arrive souvent quand k est la classe, plus il y a des chances que k soit la bonne classe,
- $P(k|x_1, ..., x_n)$ décroît quand $P(x_1, ..., x_n)$ croît : si $(x_1, ..., x_n)$ est trop courant, il nous apprend peu sur k.

La classification est dite optimale si les probabilités de chaque hypothèse sont connues. [Zhang, 2004] Ce n'est malheureusement que rarement le cas, car plus le nombre d'hypothèses considérées augmente, plus le temps de calcul s'allonge et le nombre d'estimations est élevé. Les classificateurs bayésiens présentent l'avantage d'être robustes au bruit. Néanmoins, ils ne permettent aucune lisibilité sur les résultats.

2.2.1.4 Machine à vecteurs de support (SVM : Support Vector Machine)

Les SVMs, dont un exemple de fonctionnement est schématisé sur la figure 9, sont utilisés pour résoudre des problèmes de discrimination, c'est-à-dire déterminer à quelle classe appartient un échantillon. Il s'agit de définir un hyperplan séparateur entre les différentes classes qui maximise la marge. En d'autres termes, le but est d'obtenir un hyperplan délimitant les classes tel que sa distance aux échantillons d'entraînement les plus proches est la plus grande possible, de manière à minimiser les risques d'erreur de classification. Cet hyperplan est le noyau d'une fonction linéaire non nulle ; c'est un sous-espace vectoriel de l'espace des données d'apprentissage.

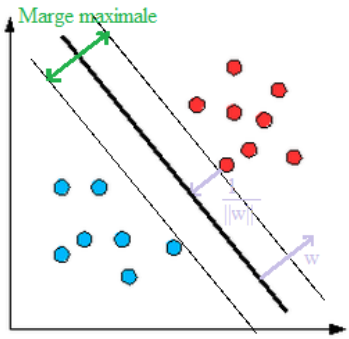


Figure 9. Exemple d'un SVM à dimension 2

La distance d'un point à l'hyperplan est donnée par la relation :

$$d(x) = \frac{\mid w.x + w_0 \mid}{\mid\mid w \mid\mid}$$

L'hyperplan optimal est celui pour lequel la distance aux points les plus proches est maximale. Cette distance vaut $\frac{2}{||w||}$. Le but étant de maximiser la marge, afin d'assurer la plus grande sécurité possible au moment où un nouvel exemple sera classé, on parle de "séparateurs à vaste marge". Étant donné que le couple (w, w_0) est défini à un coefficient multiplicatif près, on impose $y_i(w.x_i + w_0) \geq 1$. Maximiser la marge revient donc à minimiser $\mid\mid w \mid\mid$ sous les contraintes :

$$\begin{cases} min(\frac{1}{2} \mid\mid w \mid\mid^2) \\ \forall i, y_i(w.x_i + w_0) \geq 1 \end{cases}$$

Il s'agit d'un problème d'optimisation quadratique convexe sous contraintes linéaires (on a une seule contrainte) où la fonction objectif est le carré de l'inverse de la double marge. Il existe plusieurs méthodes pour résoudre un problème d'optimisation non linéaire tel que la méthode de Lagrange, la méthode de point intérieur, la méthode de gradient... En effet, un probleme convexe a un et un seul optimum et la fonction est derivable ; on peut suivre le gradient puisqu'il existe et est continu, et celui-ci nous mène nécessairement à l'optimum.

Les SVMs permettent de traiter des données de très grande dimension et sont très efficace. Leur principal inconvénient réside dans le fait que les meilleurs choix (paramètres du noyau) du paramètre de pénalisation de relâchement et du type de noyau posent énormément des problèmes et nécessitent

des lourdes tâches à l'utilisateur. De plus, cette méthode n'est pas adaptée au traitement des données qualitatives sans transformation préalable - ce qui n'est pas toujours possible sans perte d'information.

2.2.1.5 Les k-plus proches voisins

Le but des k-plus proches voisins est de trouver à quelle classe l'individu à classer a le plus de chances d'appartenir [Fix *et al.*, 1951]. Elle procède par analogie, c'est-à-dire la recherche de cas similaires ayant déjà été résolus. Cette méthode ne nécessite pas d'apprentissage mais simplement le stockage des données d'apprentissage. Pour cela, il s'agit de déterminer quelles sont, pour un nombre k défini, les k observations les plus proches parmi les exemples de l'échantillon d'apprentissage. L'individu est ainsi affecté à la classe majoritairement représentée par ses k-plus proches voisins. Un exemple de cette méthode de classification est présenté sur la figure 10.

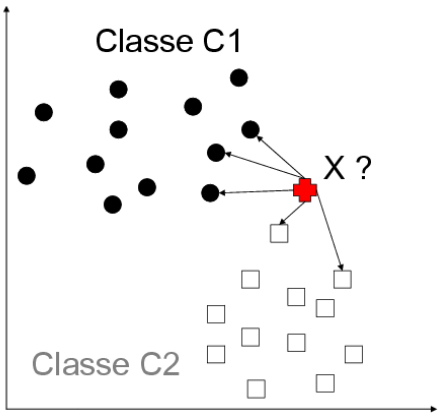


Figure 10. Exemple de l'algorithme des k-plus proches voisins pour k=5

Cette méthode procède ainsi de l'association entre une fonction de calcul de distance et une fonction d'élection de classe. Elle nécessite en entrée :

- Un paramètre k fixant le nombre de voisins à prendre en compte,
- Un échantillon de J individus, avec x caractéristiques, et leur classe respectivement associée,
- Un individu j, avec les mêmes x caractéristiques, à classer.

Elle nécessite des décisions préalables :

- Choix des attributs pertinents pour la tâche de classification considérée et des données à comparer,
- Choix de la distance par attribut, et du mode de combinaison des distances en fonction du type des attributs et des connaissances préalables du problème,
- Choix du nombre k de voisins déterminé par utilisation d'un ensemble test ou par validation croisée ; une heuristique fréquemment utilisée est de prendre k=x+1, où x est le nombre de caractéristiques.

Afin de trouver les k-plus proches voisins d'un individu à classer, la méthode la plus simple pour les données quantitatives consiste à utiliser la distance euclidienne. Elle est obtenue par la formule :

$$d(j_1, j_2) = \sqrt{\sum_{k=1}^{D} (j_{1k} - j_{2k})^2}$$

avec

- $d(j_1, j_2)$: distance entre deux individus j_1 et j_2,
- D : dimension du vecteur décrivant les différents individus telle que $x_i = (x_{i1}, x_{i2}, ...x_{iD})$.

Lorsque certains attributs manifestent une plus grande importance que d'autres, la distance euclidienne pondérée peut également être employée. Elle est définie par la formule :

$$d(j_1, j_2) = \sqrt{\sum_{k=1}^{D} w_k (j_{1k} - j_{2k})^2}$$

Concernant les données qualitatives, la fonction de distance dépend du type de modalités :
- Les modalités binaires ou les modalités énumératives : 1 dans le cas où les modalités sont strictement similaires et 0 sinon,
- Les modalités énumératives ordonnées.

Une fois les k-plus proches voisins détectés, le choix de la classe peut se faire à partir de la connaissance de la classe directement majoritaire, ou prendre en compte une pondération qui peut avoir été préalablement attribuée à chacune des classes ou calculée de sorte à être inversement proportionnelle à la distance entre l'individu à classer et le voisin courant.

La méthode peut s'appliquer dès qu'il est possible de définir une distance sur les attributs. La méthode permet de traiter des problèmes avec un grand nombre d'attributs. Néanmoins, plus le nombre d'attributs est important, plus le nombre d'exemples doit être grand.

Le problème principal de cette méthode vient de sa difficulté à gérer les grands ensembles de données, car toutes les données d'apprentissage doivent être stockées et examinées à chaque classification ; il devient alors très gourmand en temps CPU.

2.2.2 La classification non supervisée

La classification non supervisée est employée pour regrouper les individus d'un échantillon en différentes classes en fonction de leur ressemblance. Le but est d'obtenir des classes les plus homogènes possible et de séparer les individus considérés comme différents [Duda *et al.*, 2012]. Un algorithme de classification non supervisée ne connait pas les classes a priori et doit les définir à partir de l'échantillon en entrée. Les méthodes de coalescence visent donc à former une partition de l'espace par regroupements et séparation des individus en fonction de leurs affinités [Kaufman et Rousseeuw, 2009]. Le but est de maximiser la distance inter-classes et de minimiser la distance intra-classe de manière à obtenir la répartition optimale.

L'intérêt de la classification non supervisée est qu'elle offre la possibilité d'explorer l'espace des données sans aucune information sur la similarité des individus le composant et qu'elle peut ainsi en révéler des caractéristiques insoupçonnées. La liberté ainsi autorisée peut fournir des modèles très efficaces et rapides. Les classes ainsi générées peuvent par ailleurs être ensuite employées dans le cadre d'une classification supervisée comme classes prédéfinies où seront affectés les individus.

2.2.2.1 Les K-moyennes

K-means est un algorithme de quantification vectorielle, qui consiste en la minimisation alternée qui, étant donné un entier K, va chercher à séparer un ensemble de points en K clusters. Ainsi, l'algorithme des K-means permet de fournir un modèle de distribution des individus en différentes classes dont le nombre est défini au préalable par l'utilisateur [Kowalski *et al.*, 1982]. Les centres des classes sont initialement choisis à partir des données et sont décrits par un vecteur décrivant ses caractéristiques. Chaque individu est alors assigné à la classe dont les caractéristiques sont les plus proches des siennes. La définition initiale des centres peut ne pas constituer un bon modèle de la fonction de distribution de probabilité de l'espace en entrée, aussi les classes initialement déterminées évoluent-elles, par une séries d'itérations déplaçant le centre de chaque classe vers la moyenne des caractéristiques des individus la

constituant. La figure 11 montre un exemple de classification avec 3 classes.

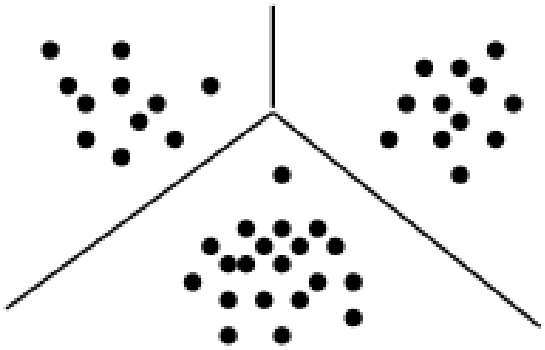


Figure 11. Classification avec 3 classes, utilisant l'algorithme des k-moyennes

La mesure de distorsion est définie selon la formule :

$$J(\nu, z) = \sum_{i=1}^{n} \sum_{k=1}^{n} z_i^k \mid\mid j_i - \nu_k \mid\mid^2$$

avec :
- ν : Vecteur des ν_k, où ν_k est le centre de la classe k,
- j_i : Individus à séparer,
- z_i^k : Variables indicatrices associées aux j_i telles que $z_k^i = 1$ si j_i appartient à la classe k, $z_k^i = 0$ sinon. z est la matrice des z_k^i.

Le but de l'algorithme est de minimiser J(μ, z) ; il est décrit par les étapes :

1. Choix du vecteur μ,
2. Minimisation de J par rapport à μ : $\mu_k^i = 1$ pour $k \in \arg\min \mid\mid j_i - \nu_k \mid\mid$. On associe à j_i le centre ν_k le plus proche,
3. Minimisation de J par rapport à ν : $\nu_k = \frac{\sum_i \mu_i^k j_i}{\mu_i^k}$,
4. Retour à l'étape 2 jusqu'à convergence.

L'étape de minimisation par rapport à μ revient à répartir les j_i selon les cellules de Voronoï dont les centres sont les ν_k. Dans l'étape de minimisation selon ν, ν_k est obtenu en annulant la kième coordonnée du gradient de J selon ν.

Le choix de K n'est pas universel, on remarque que si on augmente K, la distorsion diminue, et s'annule lorsque chaque point est centre de sa classe. Pour pallier ce phénomène, il est possible de rajouter un terme en fonction de K dans l'expression de J, mais son choix est arbitraire. On peut montrer que cet algorithme converge en un nombre fini d'opérations. Cependant la convergence est locale, ce qui pose le problème de l'initialisation. Une méthode classique consiste à lancer plusieurs fois l'algorithme en prenant les moyennes ν_k aléatoirement à chaque fois, puis on compare leur mesure de distorsion. On choisit la répartition qui possède la distorsion minimale. Dans le pire des cas, cet algorithme peut se révéler arbitrairement mauvais, mais dans la pratique, il réalise de très bons résultats. Néanmoins, et bien que l'algorithme finit toujours par converger vers une solution, celle-ci n'est pas toujours optimale.

2.2.2.2 Les K-moyennes floues

L'algorithme des k-moyennes floues est une variante de celui des k-moyennes, permettant d'obtenir des regroupements flous. Le critère de minimisation des distances intra-classe et de maximisation des

distances inter-classes prend en compte le degré d'appartenance. Concrètement, dans l'algorithme des k-moyennes, un individu appartient à une classe ou ne lui appartient pas. La notion de degré d'appartenance n'existe pas car il vaudrait 1 pour la classe à laquelle l'individu appartient et 0 pour les autres. Dans le cas des k-moyennes floues, un individu appartient à chaque classe selon un degré d'appartenance compris entre 0 et 1. Il est affecté à la classe qui maximise cet indicateur. Ces degrés évoluent donc nécessairement en même temps que les classes.

Ruspini fut le premier à proposer, en 1969, une approche de classification combinant le concept de sous-ensemble flou et les techniques basées sur la minimisation d'un critère [Ruspini, 1969], et qui a introduit la notion de partition floue. Les fonctions d'appartenance utilisées par Ruspini ont cependant une forte connotation probabiliste. Un peu plus tard, Dunn a largement généralisé l'approche de Ruspini. Dunn s'est intéressé à la définition de critères pour détecter la présence de classes compactes et séparables au sein d'un ensemble d'objets, et à la généralisation de l'algorithme des k-moyennes classiques [Dunn, 1973]. Mais c'est essentiellement Bezdek qui s'est intéressé aux aspects mathématiques des k-moyennes floues ; il a étudié leur convergence et généralisé le critère proposé par Dunn à toute une famille d'algorithmes qu'il a appelée k-moyennes floues [Bezdek, 1980] [Bezdek *et al.*, 1987] [Hathaway, 1987] [Hathaway, 1988]. Il a, pour cela, introduit un paramètre m, qui prend des valeurs strictement plus grandes que 1, et qui module le degré de flou de la partition obtenue à l'aide des k-moyennes floues. Lorsque m vaut 1 l'algorithme obtenu est quasiment équivalent, version floue, à la technique classique Isodata [Ball et Hall, 1965], qui permet de trouver, au moyen de fonctions caractéristiques, une solution approchée de la partition optimale sans passer par une recherche exhaustive de tous les cas possibles. La forme générale du critère générique, J_m, paramétré par m, est :

$$J_m(u(.), v) = \sum_{i=1}^{k} \sum_{x \in X} u_i^m(x) \; || \; x - v_i \; ||^2$$

Les centres et les degrés d'appartenance sont, pour une valeur de m donnée, calculés à l'aide des deux formules suivantes :

$$v_i = \frac{\sum_{x \in X} (\mu_i(x))^m x}{\sum_{x \in X} (\mu_i(x))^m}$$

$$\mu_i = \frac{\frac{1}{||x - v_i||^{2/(m-1)}}}{\sum_{j=1}^{k} \frac{1}{||x - v_j||^{2/(m-1)}}}$$

L'inconvénient principal de cet algorithme vient de l'étroite dépendance entre la partition initiale et le résultat final ; il n'est pas garanti qu'il soit optimal.

2.2.2.3 Les cartes de Kohonen

Comme pour les RNA, le principe des cartes de Kohonen s'inspire d'observations biologiques et, plus précisément, des mécanismes des cerveaux des vertébrés : chaque région du cerveau a des fonctions spécifiques, ainsi les neurones d'une même zone du cortex sont excités par le même type de stimuli. Les cartes de Konohen, dont la figure 12 propose une représentation, introduisent un type particulier de réseau neuronal : les cartes topologiques, fondées sur l'apprentissage compétitif. Une carte topologique est une grille composée d'unités appelées neurones.

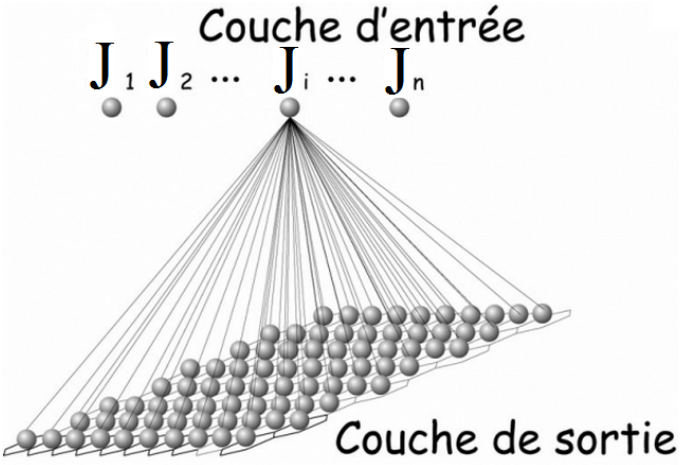


Figure 12. Carte de Kohonen

Inspirés de l'auto-organisation des régions du système nerveux, les neurones des cartes de Kohonen se spécialisent et regroupent un même type d'individus de l'échantillon en entrée [Kohonen, 2001] ; chaque neurone est lié aux autres neurones suivant une topologie et connecté à toutes les unités d'entrée dont le nombre correspond à la dimension des données d'entrée. Comme les connexions sont pondérées, chaque neurone peut être considéré comme un vecteur de poids dont les composants représentent la force des connexions synaptiques avec les données d'entrée. Le vecteur d'entrée et les vecteurs de poids de tous les neurones ont les mêmes dimensions.

La carte de Kohonen [Kohonen, 1982] a comme propriété de réaliser une quantification vectorielle de l'espace des entrées tout en respectant la distribution originale de ces entrées. En effet, l'application de l'algorithme de Kohonen crée un ensemble, de taille finie et fixée a priori, de prototypes, ou vecteurs codes, ayant les mêmes dimensions que les données en entrée. Après l'apprentissage, chacun de ces prototypes, reliés entre eux par une relation de voisinage sur la carte, représente un sous-ensemble de l'ensemble des entrées partageant certaines caractéristiques. A la convergence, les neurones de la carte représentent des zones formées autour de chacun des prototypes. Selon la terminologie de Voronoï, ces prototypes sont considérés comme des centroïdes des zones ou classes obtenues. Ces zones sont appelées »cellules de Voronoï». Les voisinages entre classes peuvent être choisis de manière variée, mais en général on suppose que les classes sont disposées sur une grille rectangulaire qui définit naturellement les voisins de chaque classe. Les figures 13, 14, et 15 montrent des topologies différentes, correspondant respectivement à des structures en grilles ou ficelles, en hexagone, et en cylindre.

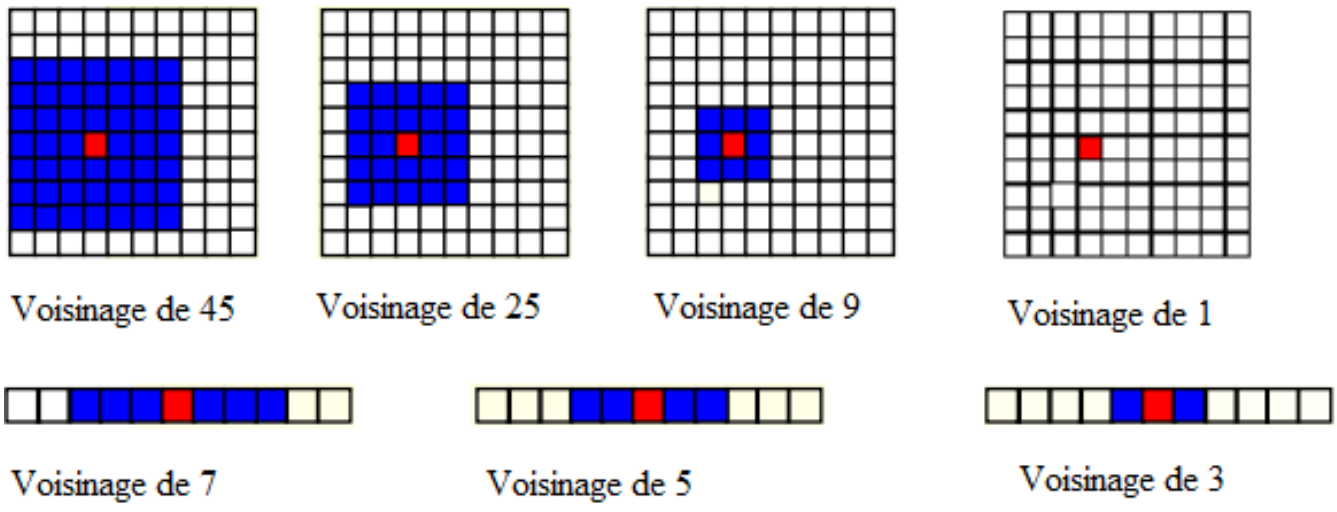


Figure 13. Structure en grille ou en ficelle

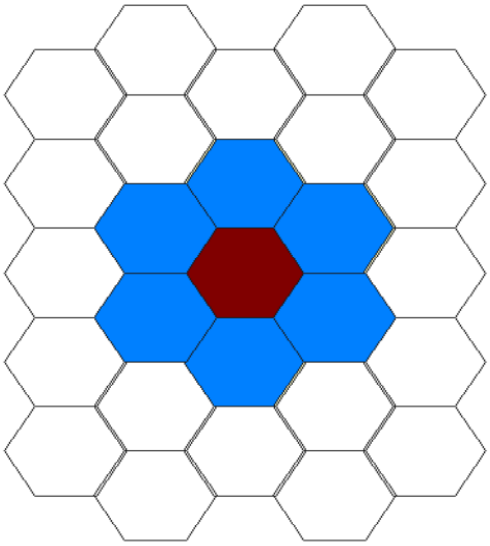


Figure 14. Structure en hexagone

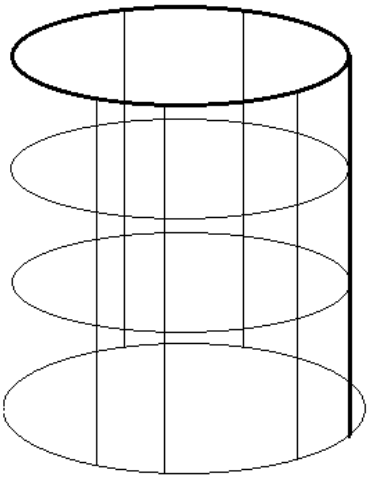


Figure 15. Structure en cylindre

La réduction de la dimension de l'espace d'entrée résulte de la combinaison de l'opération de regroupement et de quantification citée précédemment, avec la définition d'une structure de voisinage. Ainsi, en appliquant l'algorithme SOM, l'espace d'entrée constitué par n individus sera réduit en un espace de dimension inférieur (k<n) constitué par un ensemble de micro-classes. Chaque classe est représentée par la moyenne des caractéristiques des individus qui la constituent. Cet algorithme est itératif, et se déroule selon les étapes suivantes :

1. Initialisation : association à chaque classe d'un prototype dans l'espace des observations choisi de manière aléatoire,
2. Sélection aléatoire d'une observation, pour la comparer à tous les prototypes des classes, et déterminer la classe gagnante - c'est-à-dire celle dont le prototype est le plus proche au sens d'une distance donnée a priori,
3. Rapprochement de l'observation du prototype de la classe gagnante et des classes voisines.

L'association est exprimée par la fonction :

$$\Phi_w(j) = argmin_{k \in K} \mid\mid j - v_k \mid\mid$$

avec

- j : vecteur d'entrée, correspondant à l'individu à classer et appartenant à la liste L des individus à classer,
- v_k : prototype de k tel que $k \in K$; le vecteur correspondant aux caractéristiques de k est décrit par $\overrightarrow{k}$.

Le neurone vainqueur v, décrit par le $\overrightarrow{u}$, et ses voisins (définis par une fonction d'appartenance au voisinage) déplacent leurs vecteurs référents vers le vecteur d'entrée selon la formule :

$$v_k^{t+1} = v_k^t + \delta v_k^t$$

où

$$\delta v_k^t = \epsilon.h.(j - v_k^t)$$

avec

- $\epsilon = \epsilon(t)$: Coefficient d'apprentissage,
- h=h(k,v,t) : Fonction d'appartenance au voisinage, décrite le plus souvent par la formule $h = h(k, v, t) = \exp(-\frac{||\overrightarrow{k} - \overrightarrow{v}||}{2\sigma^2(t)})$, où σ est le coefficient de voisinage dont le rôle est de déterminer un rayon de voisinage autour du neurone vainqueur.

Cette méthode aboutit à une distribution souple de la topologie du réseau en un sous-espace non linéaire des données d'entrée. Elle est plus rapide que les méthodes des k-moyennes car elle considère les différents espaces comme reliés entre eux et non comme des zones isolées, ce qui réduit le coût des calculs.

L'inconvénient de ces cartes concerne la rigidité des liens entre les différents neurones.

2.2.3 Le deep-learning

Comme nous en avons donné un aperçu, les méthodes d'apprentissage peuvent être de différents types :

- L'apprentissage supervisé - très efficace lorsqu'on bénéficie d'un nombre suffisant d'exemples, c'est par exemple la méthode employée pour entraîner un détecteur de spams,
- L'apprentissage non supervisé - difficile à mettre en oeuvre mais intéressant dans le cas d'un manque conséquent de données labelisées, c'est ce qui permet le partitionnement de données,
- L'apprentissage par essai/erreurs - c'est ce qui est souvent utilisé pour apprendre un système à jouer aux échecs.

Les deux principaux enjeux pour la conception d'une méthode d'apprentissage concernent l'adaptabilité du système et sa robustesse : il s'agit d'établir des algorithmes capables de minimiser l'effort humain en garantissant l'autonomie et la fiabilité du système apprenant.

Le deep-learning est l'incarnation moderne des RNA et base son fonctionnement sur des unités mathématiques simples et entraînables, qui collaborent entre elles pour permettre la résolution d'une problème complexe [Hinton, 2006]. L'idée du deep-learning repose sur le fait que toute fonction continue peut être approximée, avec un seuil de précision arbitraire, par un RNA comprenant une couche cachée [Hornik *et al.*, 1989], conjugué à la difficulté d'élaborer un RNA face à des problèmes trop complexes - car alors, pour construire un RNA avec une couche cachée, un neurone différent est nécessaire pour chaque donnée d'entrée : les couches du réseau classique sont beaucoup plus spécifiques à un problème précis que celles obtenues par deep-learning. Le deep-learning permet dès le départ une hiérarchisation qu'il est impossible d'obtenir avec un réseau de neurones classique entrainé directement sur les données brutes. Pour cette raison, le temps de traitement devient alors extrêmement long - bien plus que

ce qui peut être généralement accordé à un système de cette nature pour pouvoir prendre une décision. C'est en cela que le deep-learning permet une généricité plus importante ainsi que des applications plus nombreuses et plus performantes.

En effet, s'inspirant des connaissances neurologiques, en particulier des modèles de communication du système nerveux, le deep-learning regroupe un ensemble de méthodes d'apprentissage, basées sur la classification supervisée ou non supervisée ; son objectif est de construire des représentations les plus immédiates possibles du réel et de créer des modèles capables d'apprendre ces représentations à partir de données non labellisées à grande échelle [Bengio *et al.*, 2015].

Le cortex supérieur du cerveau humain met en oeuvre des millions de structures neuronales très semblables et relativement simples jouant le rôle d'"identificateurs de permanences structurales", permettant l'émergence de concepts. Il s'agit d'acquis de l'évolution indispensables pour interpréter les messages sensoriels, endogènes et exogènes, en y faisant apparaître les ordres indispensables à la survie. Pouvant être modélisée comme une pyramide hiérarchique constituée de ces identificateurs, la structure du réseau neuronal permet au cerveau de construire des images aussi adéquates que possible, correspondant aux «réalités » extérieures « permanentes » supposées correspondre aux messages sensoriels en émanant ; elle octroie au cerveau humain une grande plasticité et, de là, une intéressante capacité d'adaptation à un environnement complexe. Les algorithmes de deep learning s'opposent aux algorithmes d'apprentissage peu profond du fait du nombre de transformations réalisées sur les données entre la couche d'entrée et la couche de sortie, où une transformation correspond à une unité de traitement définie par des poids et des seuils ; ils peuvent être supervisés ou non supervisés et leurs applications comprennent la reconnaissance de modèles ou la classification statistique.

Ainsi, la figure 16 montre que, de la même manière que le système nerveux tente d'établir des connexions en fonction des messages reçus, de la réponse neuronale consécutive et du poids des connexions entre les neurones du cerveau, le deep-learning s'intéresse à la manière de modéliser, avec un haut niveau d'abstraction, des données grâce à des architectures articulées de différentes transformations non linéaire ; au lieu d'un traitement séquentiel des données, le système les traite toutes parallèlement.

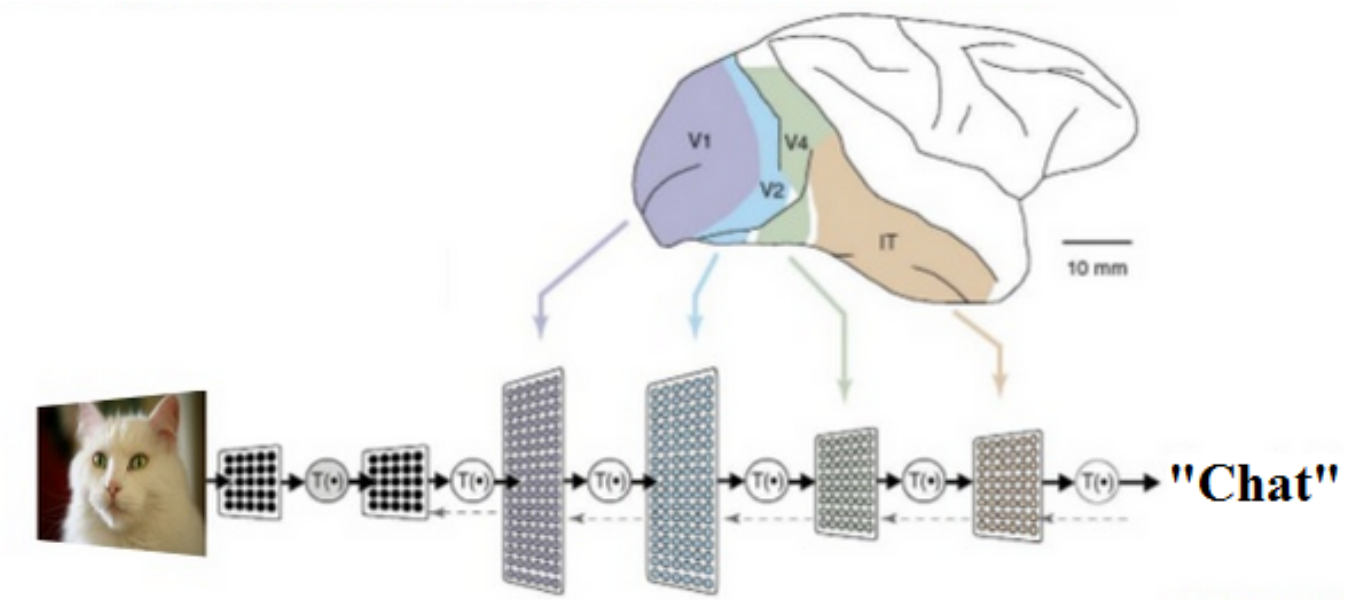


Figure 16. Structure du deep-learning

Les méthodes de "deep-learning" fonctionnement sur la base d'un apprentissage à plusieurs niveaux de détails ou de représentations des données, utilisant différentes couches d'unités de traitement non-linéaire pour l'extraction et la transformation des caractéristiques ; à travers ces différentes couches, les

algorithmes traitent le passage de paramètres de bas niveau à des paramètres de plus haut niveau. Les différentes couches composant un réseau profond sont donc organisées de façon hiérarchique ; elles peuvent être par exemple constituées de neurones artificiels. Lorsque le réseau est entrainé, les couches inférieures (reliés aux entrées du réseau) apprennent à représenter au mieux la structure des données. L'entraînement des couches inférieures est appelé le pré-apprentissage et nécessite un grand nombre de données brutes (images, texte, extraits sonores...). L'avantage de ce pré-apprentissage est double, puisqu'il :

- Permet de booster significativement les performances du modèle,
- Assure une bonne généricité, supérieure à ce que proposent les algorithmes classiques d'apprentissage ; il sera ainsi possible de réutiliser le même modèle pré-entrainé sur des tâches différentes. Par exemple, pour reprendre le projet de reconnaissance faciale, il est avantageux d'entraîner les couches inférieures à décrire les éléments fondamentaux constituant un visage (bouche, yeux, nez…). Au niveau suivant, ces parties seront organisées en descripteurs de plus haut niveau (paire d'yeux avec une distance variable entre les deux, alignement entre le nez et la bouche…). Le réseau aura alors appris de lui-même les attributs basiques lui permettant de reconnaitre efficacement ce qu'est un visage.

Dans le cas d'une image à traiter par exemple, à chaque étape du raisonnement -c'est-à-dire à chaque couche - le réseau de neurones approfondit sa compréhension de l'image avec des concepts de plus en plus précis. Pour une couche donnée correspond un niveau de détail spécifique. De chaque unité d'une couche c émergera le traitement d'un élément précis, que la couche supérieure c+1 prendra en considération et associera à toutes les réponses émises par les différentes unités de c, de manière à effectuer à second traitement, plus global. Ainsi, l'identification d'une personne s'opère par la décomposition de l'image : la méthode permet d'isoler la tête, puis les cheveux, la bouche, le nez... et de traiter chaque élément séparément. La couche c traite les éléments de détail du visage, la couche c+1 réunit toutes les réponses issues de la couche c pour traiter le visage en entier, et ainsi de suite. La figure 17 donne un exemple du mode de fonctionnement du deep-learning.

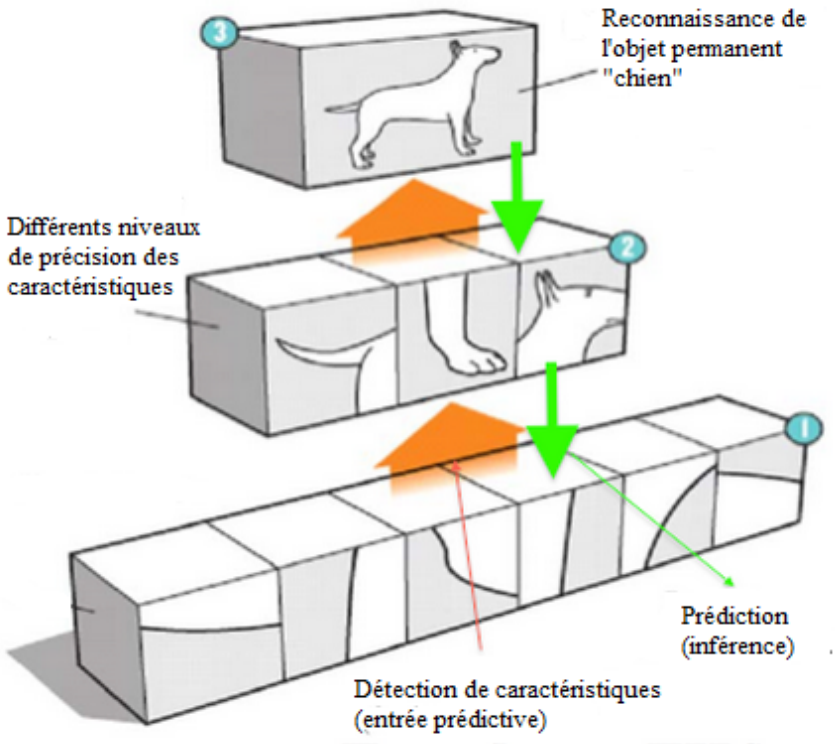


Figure 17. Reconnaissance d'une image représentant un chien

Dans le cas de la classification supervisée, il s'agit d'apprendre, à différents niveaux de précision, les caractéristiques de chacune des classes connues, à partir d'un certain nombre d'exemples. Le cas de la classification non supervisée est plus complexe puisqu'alors, le système ne connait pas les classes a priori et doit déterminer lui-même quelles sont les différents objets à reconnaître, et en déterminer les spécificités : il travaille à partir de données non labelisées. Comme le montre la figure 18, dans le cas d'un apprentissage facial, la construction de la connaissance de l'objet s'opère selon les étapes :

1. Détection et localisation de l'objet à reconnaître,
2. Division en zones de spécialisation,
3. Représentation des caractéristiques à chaque niveau de traitement,
4. Labelisation.

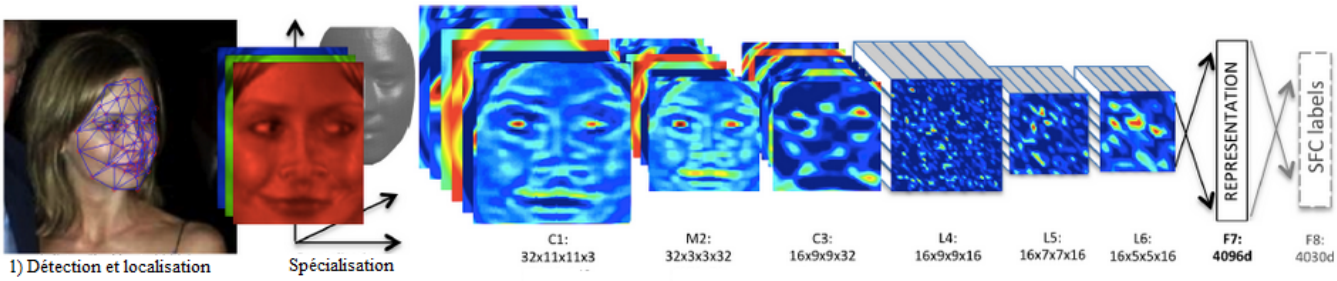


Figure 18. Architecture d'une technique de reconnaissance faciale par deep-learning

Pour expliquer le deep-learning dans le contexte particulier de la classification supervisée, prenons un exemple concret : il s'agit d'apprendre une relation entre des observations x en entrée (les images de chiffres) et des labels y en sortie (les chiffres '0' à '9'). La stratégie la plus courante consiste à construire un modèle discriminant, c'est-à-dire un modèle capable d'observer, pour les probabilités conditionnelle p(y | x), un label y lorsqu'on connait l'entrée x. La régression logistique - qui consiste en la construction un modèle permettant de prédire ou expliquer les valeurs prises par une variable cible qualitative y à partir d'un ensemble de variables explicatives quantitatives ou qualitatives x_1, x_2,... x_n [Rakotomalala, 2011] - rentre par exemple dans ce cadre. Par contraste, la stratégie de Hinton consiste à élaborer un modèle génératif : un modèle pour la distribution de probabilité conjointe p(y,x) des images x et des labels y. En d'autres termes, on cherche à construire un RNA capable d'apprendre puis de générer simultanément les images x et les labels y avec une distribution de probabilité proche de celle observée dans un ensemble d'entraînement. Une fois le RNA entraîné il est ainsi possible de l'utiliser d'une part pour reconnaître des images, c'est-à-dire associer un chiffre y à une image x - on cherche alors le y qui maximise p(y | x)) - ou, inversement, pour générer des images x correspondant à un chiffre y - auquel cas on échantillonne p(x | y)). L'architecture de RNA utilisée pour réaliser le modèle génératif de Hinton s'appelle un Deep Belief Network (DBN). Elle contient deux parties, représentées dans la figure 19, qui sont :

1. Un ensemble de couches de compression qui convertissent les données entrées x en une représentation abstraite,
2. Un ensemble de couches de conversion qui associent cette représentation en labels de classification y.

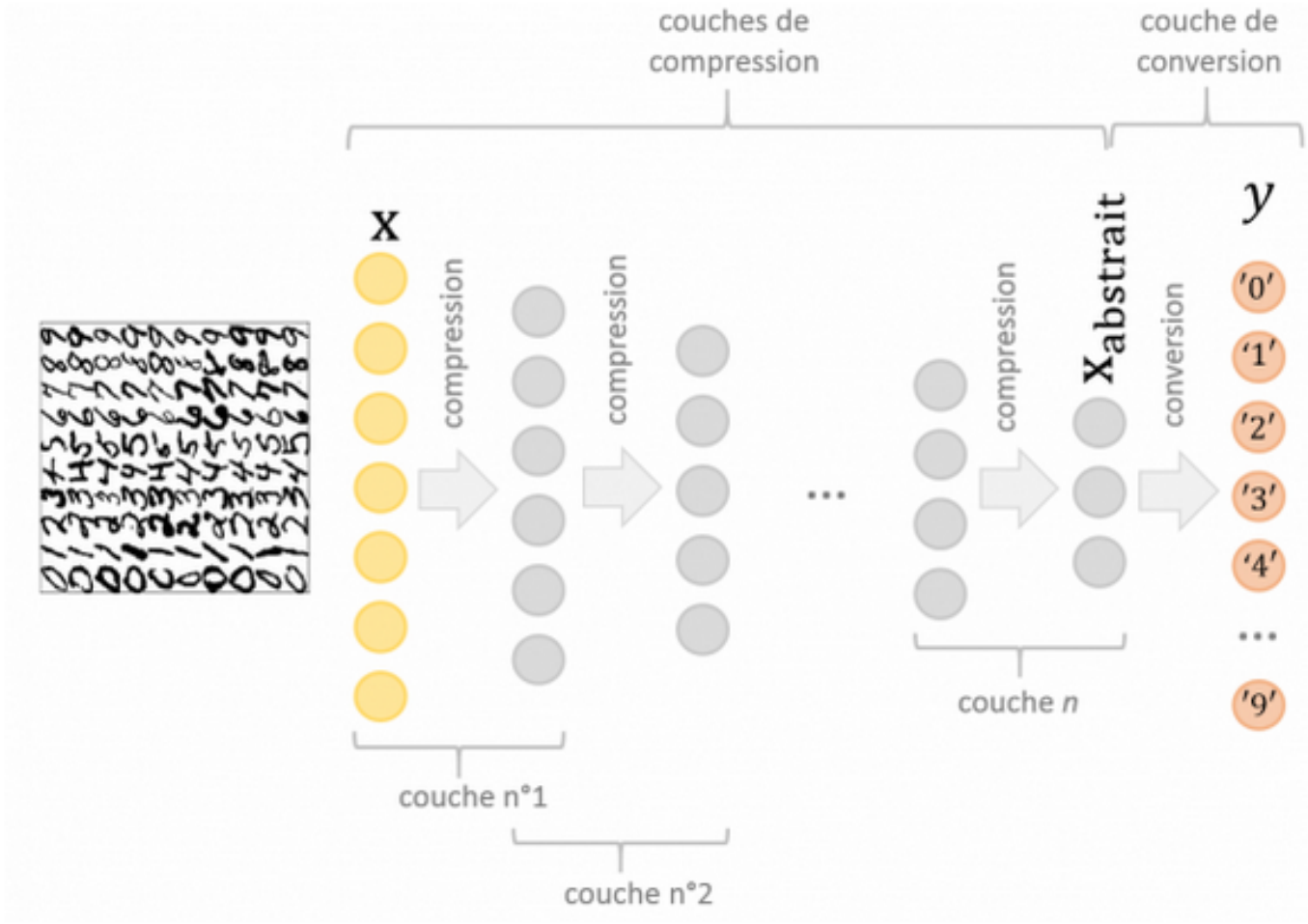


Figure 19. DBN identifiant des images de digits

La première partie a pour objectif d'apprendre la distribution des données x présentées en entrée sans tenir compte des labels y. Elle est constituée d'une succession de couches dont chacune contiendra une représentation plus abstraite (ou compressée) que la précédente. Considérons, par exemple, un RN chargé de classifier des images. La première couche stockera les niveaux de gris de l'image (l'équivalent de la rétine), la seconde contiendra, par exemple, un encodage des lignes ou des zones de contraste de l'image, la troisième détectera l'existence de certaines formes géométriques simples comme des cercles, la quatrième identifiera certains agencements particuliers de ces figures comme celles représentant un '8' formé de deux cercles juxtaposés et ainsi de suite. Une fois entraînée, cette première section du RNA contiendra une représentation hiérarchique des données en entrée, la dernière couche encodant la représentation la plus abstraite et aussi, finalement, la plus utile. Cette première phase peut être conçue comme une initialisation efficace du DBN, on l'appelle aussi pré-entraînement. Ainsi, la fonction de l'étape de compression est d'implémenter les couches de la première section. La principale caractéristique attendue de cette section réside dans sa capacité à apprendre rapidement une distribution de probabilité spécifiée empiriquement par un jeu d'exemples. Il existe pour cela différentes alternatives, comme l'emploi d'auto-encodeurs ou de réseaux de convolution, mais le modèle de réseau profond proposé par [**?**] est basé sur un empilement de machines de Boltzmann restreintes (MBR).

La machine de Boltzmann est un modèle de réseau de neurones qui a été défini en 1985 par Ackley, Hinton et Sejnowski [Ackley *et al.*, 1985]. Ce réseau comporte des unités visibles et des unités cachées, complétement interconnectées, et son apprentissage est basé sur des évaluations statistiques, par méthode de recuit simulé [Kirkpatrick *et al.*, 1983], lors de deux phases distinctes : l'éveil (sous l'influence de stimuli d'entrée) et le repos (relaxation en système fermé). Alors que les neurones des RNA ont des niveaux d'activation déterministes compris entre 0 et 1, les neurones des MBR sont des variables aléa-

toires binaires. Elles sont réparties sur deux couches :
- La couche visible, encodant les exemples de l'ensemble d'entraînement L,
- La couche cachée, stockant une forme compressée ou abstraite des données de la couche visible.

Les machines de Boltzmann restreintes constituent un modèle probabiliste à part entière : une RBM modélise la probabilité d'une entrée v avec la probabilité jointe p(v, h). Elle est définie par la distribution de probabilité conjointe :

$$p(v,q) = \frac{e^{-E(v,q)}}{Z}$$

où E est une fonction d'énergie donnée par la formule :

$$E(v,q) = \sum_{i=1}^{V} a_i v_i + \sum_{j=1}^{Q} b_j h_j + \sum_{i=1}^{V}\sum_{j=1}^{Q} w_{ij} v_i h_j$$

avec
- $v = (v_1, \ldots, v_V)$: Niveaux d'activation des V neurones de la couche visible,
- $h = (q_1, \ldots, q_Q)$:Niveaux d'activation des Q neurones de la couche visible,
- w_{ij} : Poids de la connexion entre les unités v_j et q_i,
- $Z = \sum_{v,h} e^{-E(v,q)}$:Facteur de normalisation approprié pour que p somme à 1 sur l'ensemble des configurations,
- a, b et w : Paramètres à optimiser durant l'entraînement.

Ainsi, par construction de la fonction d'énergie, les probabilités d'activation des unités v_i sont indépendantes sachant q et réciproquement, les q_i sont indépendants sachant v. Soit $sgm(t) = \frac{1}{1+e^{-t}}$ une fonction sigmoïde, les probabilités conditionnelles individuelles pour une machine de Boltzmann restreinte sont données par :

$$\forall i \leq V, p(q_i|v) = sgm(\sum_j w_{ij} v_j),$$

$$\forall j \leq Q, p(v_j|q) = sgm(\sum_i w_{ij} q_i)$$

Le rôle de la seconde partie du DBN est de convertir la représentation abstraite de la dernière couche en labels y, qui pourront être par la suite utilisés dans le cadre d'un apprentissage supervisé. Dans l'exemple des digits cette représentation sera constituée d'une couche de sortie de dix neurones, un neurone par digit. Cette conversion peut être réalisée au moyen d'une couche logistique entraînée par une SGD classique.

L'entraînement du DBN est considéré comme achevé lorsque la performance du DBN évaluée sur un ensemble de validation distinct de l'ensemble d'entraînement ne progresse plus significativement. Cette seconde étape est appelée le fine-tuning, elle est généralement beaucoup plus lente que l'initialisation.

L'approche du deep-learning nous a semblé intéressante, de sorte que nous pensons réinvestir l'idée de base dans notre méthode ; néanmoins dans notre cas, nous n'avons pas à traiter de grosses masses de données, aussi sa mise en oeuvre est-elle trop lourde pour l'architecture du système que nous voulions concevoir.

2.2.4 La méthode LAMDA

La méthode LAMDA (Learning Algorithm for Multivariate Data Analysis) est mixte : elle peut être aussi bien employée pour un problème de classification supervisée que non supervisée, ce qui est très intéressant dans notre cas. Son principe repose sur la comparaison des degrés d'adéquation des individus à chaque classe et se base sur la théorie de la logique floue.

Cette méthode a été développée par Joseph Aguilar-Martin en collaboration avec d'autres chercheurs [Aguilar-Martin et López de Mántaras, 1982] [Desroches, 1987] [Piera *et al.*, 1989], et a connu des nombreuses améliorations au fil des années. Elle est capable de prendre en compte différents types d'attributs, ce qui la rend apte à s'adapter à de nombreux types de situation ; elle a d'ailleurs été appliquée à des domaines très variés : la psychologie [de Ariza *et al.*, 2004], les procédés biotechnologiques [Waissman *et al.*, 1998], les procédés industriels [Kempowsky *et al.*, 2003], la médecine...

Ce projet a été motivé par la volonté de fonder des diagnostics en se basant sur des historiques de données ; les problèmes rencontrés lors de l'emploi de telles données concernent deux aspects :

1. La quantité des données employées et d'attributs considérés : plus leur nombre est élevé plus les éléments risquent d'être différents les uns des autres,
2. L'inexactitude potentielle des données liée à l'imprécision de certaines informations ou mesures.
3. L'évolution des catégories en fonction des réponses des parcelles aux traitements appliqués.

Dans ce contexte, l'expression de l'appartenance des individus aux différentes classes par la logique floue, et non par une réponse binaire, permet de rompre avec une partition logique stricte de l'espace des données.

Dans le cadre de notre projet, cette méthode est particulièrement adaptée puisqu'elle permet de :

1. Intervenir aussi bien en mode supervisé que non supervisé,
2. Gérer les attributs intervallaires, qualitatifs et quantitatifs des données que nous traitons sans que nous ayons besoin de les convertir préalablement, et donc sans risque de perte d'information,
3. Manipuler rapidement un nombre élevé d'individus,
4. Conserver un bon aperçu des mécanismes en jeu de manière à garder la maîtrise, si besoin est, de la classification et de sa qualité.

LAMDA intègre également, si besoin, un algorithme de classification non floue basé sur une approche statistique.

2.3 Conclusion

Dans ce chapitre, nous avons présenté les enjeux de la classification et les différentes méthodes les plus régulièrement employées, en les répartissant en fonction du mode de classification (supervisé ou non supervisé) qu'elles assurent. Ces méthodes ont été développées selon différentes approches, dont les trois principales sont l'approche statistique (k-plus proches voisins, SVM), l'apprentissage (arbres de décision), et le réseau de neurones (ANN, cartes de Konohen, K-Moyennes et K-Moyennes floue). Nous nous sommes également intéressés au cas particulier de l'apprentissage, pouvant impliquer aussi bien un mode de classification supervisé que non supervisé, et, de manière plus spécifique, au deep-learning, qui bénéficie d'une efficacité remarquable du fait de son architecture hiérarchisée. Nous verrons dans le chapitre IV en quoi l'intérêt porté à cette approche a ouvert de nouvelles perspectives dans le traitement de données opéré la méthode que nous avons choisie.

Au cours de notre étude comparative entre les différentes méthodes de classification, dont nous avons dressé un tableau récapitulatif détaillant les différents avantages et inconvénients de ces méthodes (annexe Tableau A.6), nous avons pu étudier l'intérêt que chacune d'elles présentait pour notre projet de manière à sélectionner celle qui nous paraissait la plus adaptée. Nous avons ainsi pu opter pour la méthode LAMDA, qui intègre les deux modes de classification et permet un traitement simultané des données qualitatives et quantitatives, ce qui était nécessaire dans le projet agronomique auquel nous

participions. Cette méthode base son algorithme sur les principes de la logique floue ; elle opère en attribuant à tous les individus un degré d'adéquation spécifique à chaque classe, permettant ainsi de s'affranchir d'une vision binaire - et donc irréaliste dans le contexte qui nous occupe - de la partition de l'espace de données.

Dans le prochain chapitre, nous décrivons la méthode LAMDA et son fonctionnement avec davantage de précision.

3 La méthode LAMDA

La logique floue, qui est une des particularités de la méthode LAMDA, et a été développée et formalisée en 1965 par Lofti Zadeh. Se basant sur sa théorie mathématique des ensembles flous, elle permet, dans une certaine mesure, la modélisation du raisonnement humain. En effet, un humain est capable de conceptualiser des notions abstraites, approximatives, et intégrant différents niveaux de vérité ; la logique floue étend en cela la logique classique : elle introduit l'idée selon laquelle un objet n'est pas nécessairement totalement A ou totalement distinct de A, mais qu'il peut être considéré comme A jusqu'à un certain point [Zadeh, 1997]. Si, dans la réalité, la logique classique peut avoir des applications directes, la logique floue permet de couvrir un plus grand nombre de concepts en prenant en compte leur ambiguïté. Par exemple, pour modéliser le fonctionnement d'un feu tricolore, la logique classique semble tout à fait adaptée : la couleur du feu appartient à l'ensemble vert, orange, rouge et, à un instant t, est strictement l'une de ces trois couleurs. Par contre, en considérant la vitesse à laquelle une voiture roule, la logique classique est en cela limitée qu'elle a besoin de critères précis pour fonctionner, comme par exemple : "si la voiture roule à 90 km/h ou au-delà, sa vitesse est élevée. Sinon, elle est faible.". Mais l'esprit humain ne fonctionne pas de cette manière : il considère avec certitude la vitesse élevée à 130 km/h ou la vitesse basse à 20 km/h, mais ne traite pas l'information de la vitesse selon ce schéma binaire parce qu'il intègre une évaluation contextuelle qui fonctionne par degré. Ainsi, à 90 km/h, la vitesse d'une voiture peut être, par exemple, considérée comme élevée à un degré de 0,6 et basse à un degré de 0,4. La figure 20 illustre schématiquement la différence entre la perception classique et la perception floue.

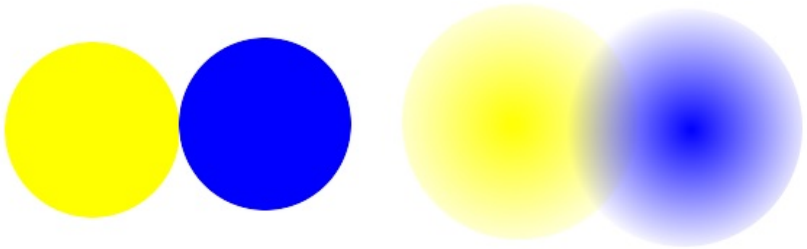


Figure 20. Représentation classique/ Représentation floue

La logique floue est ainsi une extension de la logique classique, puisqu'elle permet de traiter des cas ambigus ou approximatifs. Elle rend notamment possible la gestion de données bruitées ou basées sur des concepts imprécis. C'est un outil particulièrement utilisé en intelligence artificielle, de manière à bénéficier d'une représentation modulée de données réelles, et trouve de nombreuses applications dans l'automatique, la robotique, le contrôle aérien, la météorologie, l'aide au diagnostic médical,... Les données qu'elle traite sont des variables floues dites "variables linguistiques", chacune d'elles constituant un ensemble flou de l'univers du discours. Ainsi, à la différence d'avec la logique classique, les variables ne sont plus considérées en logique floue comme appartenant ou n'appartenant pas à A mais appartenant avec un certain degré à A.

La méthode LAMDA a été conçue dans le but de traiter des données représentant au mieux les informations issues du monde réel. Là encore, l'esprit humain gère de manière automatique les données qualitatives ou numériques par exemple. Si, dans la vie quotidienne, il décrira un objet en nommant sa couleur et évaluera son poids à l'aide d'une valeur chiffrée - encore que, si l'information ne requiert pas de précision, il pourra se contenter d'une appréciation adjectivale. Ainsi, outre le fait que son fonctionnement se base sur la logique floue, la méthode LAMDA intègre un système de gestions de données hétérogènes, qui lui permet de traiter des informations de types différents : quantitatifs, qualitatifs, intervallaires. Le choix du type employé pour décrire une donnée est à l'initiative de l'utilisateur ; cette décision relève du contexte particulier de la classification, de ses besoins, et des informations disponibles.

Nous allons maintenant décrire cette méthode de classification, en expliquant tout d'abord comment le principe SMSP (Simultaneous Mapping for Single Processing) lui permet de traiter des données hétérogènes, puis en détaillant le fonctionnement de LAMDA dans ces différents emplois (classification non supervisée et supervisée, sélection des descripteurs). Nous terminerons en évoquant l'algorithme d'évaluation de la qualité de la classification qui lui est intégré.

3.1 Traitement de données hétérogènes

Beaucoup de situations d'apprentissage impliquent des données décrites par des types différents et appartenant à des espaces hétérogènes. Dans certains cas, les données d'un type A peuvent être converties en un type B, mais ce n'est pas toujours possible sans perte d'informations ; aussi l'absence d'analogie entre les différents espaces de données peut-elle affecter la qualité des connaissances extraites. Le problème de représentation de données par des types multiples a été traité dans différents travaux à travers des frameworks d'apprentissage [Michalski et Stepp, 1983] [Mohri et Tanaka, 1994] [Hu *et al.*, 2007], mais les méthodes proposées ne permettent pas la gestion simultanée de données hétérogènes ; elles se basent respectivement sur la distance entre les différentes mesures pour traiter séparément des données qualitatives et quantitatives à travers des stratégies de réduction de dimensionnalité [Kira et Rendell, 1992], et la distance de Hamming pour gérer des données qualitatives dans des situations de classification [Aha, 1989]. D'autre approches permettent de ne traiter que des données quantitatives et, dans cette optique, proposent des méthodes pour transformer des données qualitatives en données quantitatives [Cover et Hart, 1967], ce qui peut occasionner des lourdes pertes d'information. De même, la qualité de l'information peut s'en trouver affectée dans le cas de transformations de données quantitatives en valeurs qualitatives par la discrétisation de l'espace de données en différents intervalles [Hall, 1999].

Pour pallier cette difficulté, LAMDA intègre le principe SMSP[Hedjaz *et al.*, 2012] afin de gérer simultanément des caractéristiques de types différents, de sorte qu'un même individu i peut être décrit par des critères très variés, quelles que soient leurs contraintes, leur incertitude, et le moyen employé pour les évaluer. Il permet d'opérer une partition des individus en un seul processus, en projetant les différents descripteurs - quel que soit le type de chacun d'eux - dans un espace homogène et unique, de sorte à les traiter tous ensemble. Il s'agit là d'un des principaux avantages de LAMDA, puisque, parmi les autres méthodes de classification, la solution la plus couramment employée consiste à traiter chaque type de données séparément et présenter les différentes classifications obtenues comme autant de réponses possibles à la question initiale. Cependant, cette approche ne résout pas complètement le problème initial puisque la question relative au moyen de rassembler les solutions partielles pour faire émerger une décision finale globale, basée sur l'ensemble des données, se pose à nouveau.

De ce fait, le choix des différentes caractéristiques et de leurs types respectifs est une étape importante de l'analyse, et nécessite une prise en compte rigoureuse des besoins et des objectifs de la classification. Par exemple, pour décrire une couleur, on choisira d'opter pour un type qualitatif ("rouge", "vert", "bleu") pour une discrimination stricte, et un type quantitatif - c'est-à-dire la longueur d'onde correspondante - si l'idée de progression dans le spectre lumineux est importante (considérer que le violet est plus proche du rouge que du vert, par exemple). Ainsi, à l'intérieur d'une même classification, l'espace de données employé peut être très hétérogène, et la transcription de données issues d'un capteur en un type compatible peut s'avérer complexe et occasionner de la perte d'informations importantes.

Le principe SMSP, représenté sur la figure 21, propose une solution alternative en intégrant toutes les données - qu'elles soient de type quantitatif, qualitatif, ou intervallaire - dans un même espace unifié de données et les traiter simultanément, afin de n'établir qu'une seule classification, prenant en compte l'ensemble des données en entrée. A chaque type de données correspond une fonction de traitement bien spécifique afin d'éviter toute distorsion ou perte d'informations. A la suite de tous ces traitements particuliers, les données de l'espace homogène obtenu sont traitées au moyen d'une fonction unique pour fournir la classification voulue. Deux fonctions sont donc consécutivement employées pour permettre le traitement simultané de données représentées par des types de données différents :

- Une fonction d'appartenance floue, spécifique à chaque type de données,
- Une fonction d'aggrégation floue, valant pour toutes les données -quel que soit leur type - une fois projetées.

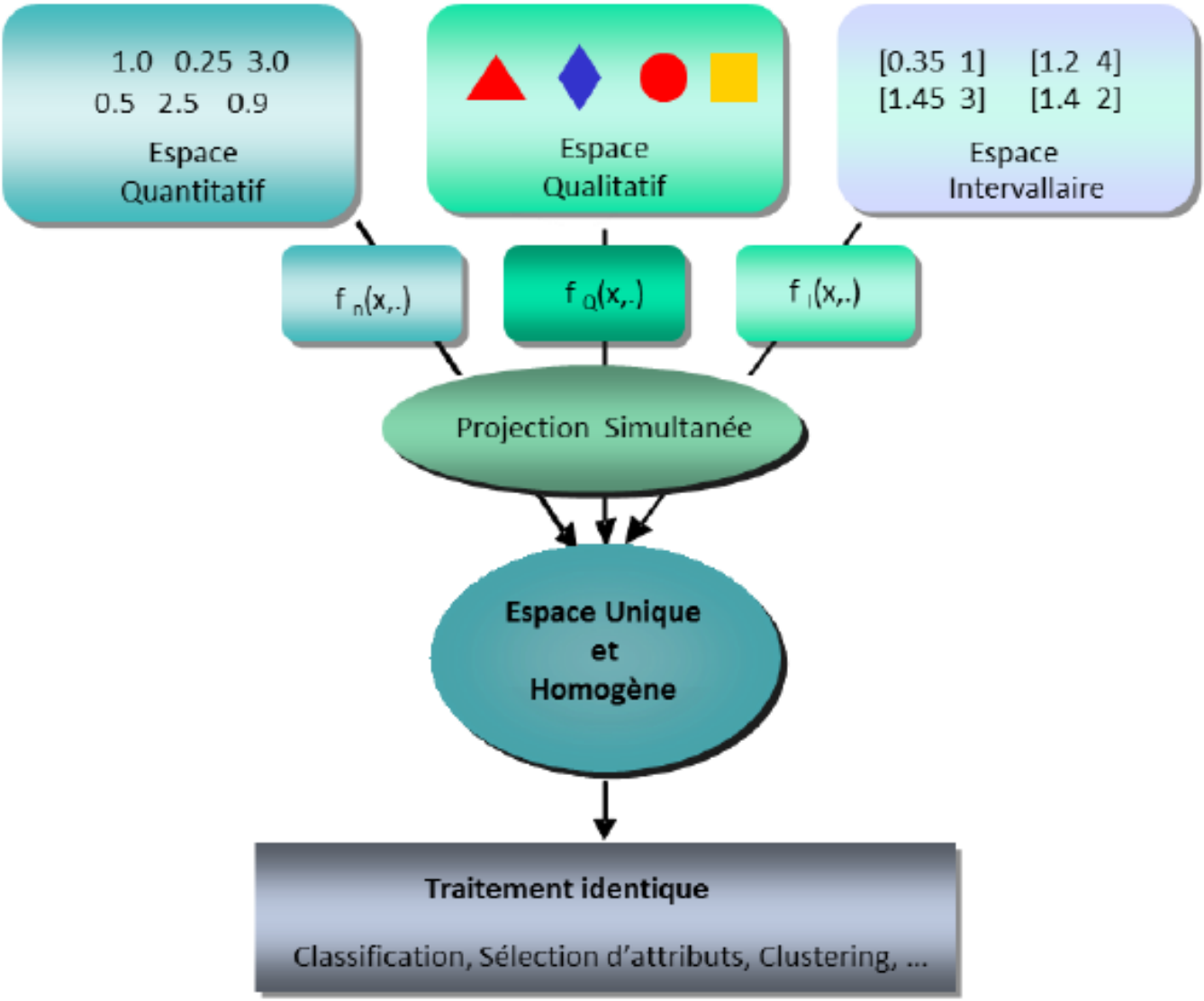


Figure 21. Le principe SMSP

De manière à permettre une représentation plus réaliste de l'espace de données, le principe SMSP intègre la théorie floue dans son raisonnement. Les différentes fonctions pour la projection des données initiales dans un même espace de données leur attribuent systématiquement un vecteur de Degrés d'Adéquation Marginal (MAD) basé sur la théorie floue. Le traitement des données homogènes, en suivant, s'appuie sur la théorie des ensembles flous et affecte à chaque individu un vecteur de Degrés d'Adéquation Global (GAD).

3.1.1 Degré d'Adéquation Marginale (MAD)

Un individu j décrit par d caractéristiques est représenté par un vecteur X de dimension d tel que $\overrightarrow{v_1}$ = {x_1, x_2,... x_d}. Chacune de ces caractéristiques peut être de type qualitatif, quantitatif, ou intervallaire. A chacun de ces types correspond un calcul de MAD spécifique. Pour classer i dans un espace de K classes, ce traitement projectif permet d'affecter à chaque donnée x_i un vecteur de K MAD, décrivant le degré d'appartenance de x_i à la caractéristique i de chacune des classes. On cherche à obtenir μ_k^i tel que :

$$\mu_k^i(x_i) = f_i(x_i, \theta_{ki})$$

avec

– f(x) : fonction spécifique au type de la caractéristique i,
– k : numéro de la classe tel que $k \in [1, K]$,
– θ_{ki} : valeur de la i^e caractéristique de la classe C_k.

3.1.1.1 Descripteurs quantitatifs

Les variables quantitatives correspondent à une valeur chiffrée. Il peut s'agir d'une mesure de capteur, d'un numéro de rang,... Elles présentent deux principales caractéristiques :
– Elles sont précises,
– Elles ont, entre elles, un rapport de grandeur.

Pour le calcul du degré d'appartenance, on considère que l'ensemble des valeurs possibles de chaque caractéristique de type quantitatif appartient à un intervalle compact. Si les bornes de l'intervalle ne sont pas connues a priori, il est toujours possible de les déclarer à partir de l'ensemble de données, de sorte à permettre une normalisation sans perte d'information de toutes les valeurs quantitatives d'une caractéristique données. Soit $[x_{imin}, x_{imax}]$ l'intervalle déduit à partir de l'ensemble des données d'entraînement ou fixé a priori. La normalisation de la valeur de la caractéristique x_i^j est calculée selon la formule :

$$x_i^j = \frac{\widehat{x}_i^j - x_{imin}}{x_{imax} - x_{imin}}$$

avec :
– $\widehat{x}_i^j$ la valeur initiale de la i^e caractéristique de l'individu j,
– x_i^j la valeur normalisée de la i^e caractéristique de l'individu j.

Plusieurs fonctions peuvent alors être employées. Le choix de la fonction se fait en fonction du problème auquel la classification cherche à faire face. La première fonction ayant été proposée est la guaussienne [Aguado et Aguilar-Martin, 1999] ; analytique et de limite nulle en infini, la fonction gaussienne est une fonction propre de la transformée de Fourier continue, ce qui la rend particulièrement pertinente dans le domaine de la physique, où de nombreux phénomènes suivent une distribution de type gaussien. Pour la i^e caractéristique, le calcul de son degré d'appartenance au descripteur correspondant de la k^e classe est décrit selon la formule :

$$\mu_k^i(x_i) = \exp^{\frac{-\frac{1}{2}(x_i - \rho_k^i)^2}{\sigma_i^2}}$$

où :
– $\rho_k^i \in [0, 1]$ = Moyenne de la i^e caractéristique de la classe k, calculée selon les valeurs de la i^e caractéristique des individus de l'échantillon d'entraînement ayant été affectés à cette classe,
– $x_i \in [0, 1]$ = i^e caractéristique normalisée,
– σ_i = écart type de la i^e caractéristique obtenue à partir des individus de l'espace d'entraînement ayant été affectés à la classe k.

Une autre fontion très usitée est la binomiale. Cette loi de probabilité discrète est décrite par deux paramètres : le nombre n d'expériences réalisées et la probabilité de succès p ; elle modélise le nombre de succès obtenus lors de la répétition indépendante de plusieurs expériences aléatoires identiques. Ainsi, elle est utilisée dans divers domaines d'étude, notamment à travers des tests statistiques qui permettent d'interpréter des données et de prendre des décisions dans des situations dépendant de l'aléa : la génétique [Morgenthaler, 2008], la linguistique [Cossette, 1994], ... Elle permet de calculer le MAD selon la formule :

$$\mu_k^i(x_i) = \rho_k^{i^{x_i)}}.(1 - \rho_k^i)^{(1-x_i))}$$

où :
– $\rho_k^i \in [0, 1]$ = Moyenne de la i^e caractéristique de la classe k, calculée selon les valeurs de la i^e caractéristique des individus de l'échantillon d'entraînement ayant été affectés à cette classe,

– $x_i \in [0,1] = i^e$ caractéristique normalisée.

On peut également employée la fonction d'appartenance basée sur la loi binomiale centrée, c'est-à-dire la loi binomiale de moyenne nulle et d'écart type unitaire. Une fonction de similitude établit la relation entre l'individu à classer et un prototype de la classe k. Elle est définie selon la formule :

$$\mu_k^i(x_i, \mid \nu_k^i, \rho_k^i) = \rho_i^{1-|x_i-\nu_k^i|)}.(1-\rho_k^i)^{|x_i-\nu_k^i|}$$

où :
– ν_k^i correspond au centre de la classe k,
– $\rho_k^i \in [0,1]$ = Fonction de dispersion remplissant les conditions selon lesquelles $\forall x_i \neq \nu_k^i, \mu_k^i(\nu_k^i \mid \nu_k^i, \rho_k^i) \geq \mu_k^i(x_i \mid \nu_k^i, \rho_k^i)$ et que pour $\rho_1 \leq \rho_2 \forall x_i \neq \nu_k^i$, les degrés d'adéquation ont des valeurs telles que $\mu_k^i(x_i, \mid \nu_k^i, \rho_1) \geq \mu_k^i(x_i, \mid \nu_k^i, \rho_2)$,
– $x_i \in [0,1] = i^e$ caractéristique normalisée.

3.1.1.2 Descripteurs qualitatifs

Les variables quantitatives sont décrites par des mots. Concrètement, avoir recours à des variables qualitatives nécessite déjà une première classification -qui peut-être inconsciente par le biais du langage. Au lieu de décrire un objet par une valeur chiffrée exacte, on le caractérise par un terme. Néanmoins, contrairement à ce que l'on pourrait croire, l'emploi d'un mot n'est pas nécessairement moins précis qu'une évaluation numérique : si c'est effectivement le cas lorsqu'on choisi de parler par couleurs et non par longueurs d'onde, évoquer des triangles ou des quadrilataires au lieu de figures à 3 ou 4 côtés n'induit pas de perte d'information. Leurs intérêts principaux résident dans le fait qu'au contraire des variables quantitatives :
– elles ne sont pas triables et, de ce fait, n'entraînent pas nécessairement d'échelonnage - bien que nous l'ayons rendu possible par la prise en compte de la proximité des mots entre eux comme nous le verrons par la suite,
– elles peuvent décrire de manière approximative une variable dont la valeur numérique est incertaine ou difficilement évaluable.

Pour calculer le MAD des descripteurs qualitatifs, on considère les différentes valeurs que peut prendre la caractéristique comme un ensemble de modalités, telles que $D_i = Q_1^i, Q_2^i, ...Q_m^i$. Le MAD correspond alors à la fréquence de la modalité Q_m^i de la caractéristique x_i^j dans la classe k, c'est-à-dire la proportion d'individus de l'échantillon d'entraînement ayant été affectés à la classe k et dont la i^e caractéristique affichait la modalité Q_m^i [Dubois et Prade, 1997]. Ainsi, chaque modalité $Q^i \in D_i$ a une fréquence associée. Soit Θ_{kj}^i la fréquence de la modalité Q_j^i pour la classe k. La fonction d'appartenance concernant la i^e caractéristique est décrite par la formule :

$$\mu_k^i(x_i) = (\Theta_{k1}^i)^{q_1^i} * (\Theta_{k2}^i)^{q_2^x} * ... * (\Theta_{km}^i)^{q_m^i}$$

où :
– $q_j^i = 1$ si $x_i = Q_j^i$
– $q_j^i = 0$ sinon

3.1.1.3 Intervalles

Les intervalles sont décrits par deux valeurs numériques : une borne minimale et une borne maximale. Ils permettent notamment de prendre en compte l'incertitude de certaines données, en particulier l'incertitude liée aux capteurs [Hedjazi, 2011].

La fonction d'appartenance employée pour les descripteurs de type intervallaire correspond à la similarité $S(x_i, \rho_k^i)$ entre la valeur de l'intervalle symbolique pour la i^e caractéristique x_i et l'intervalle

$\rho_k^i = [\rho_k^{i-}, \rho_k^{i+}]$, représentant la valeur de l'intervalle moyen pour la i^e caractéristique de la classe k, c'est-à-dire :

$$\mu_k^i(x_i) = S(x_i, \rho_k^i)$$

Soit ϖ le cardinal vectoriel d'un ensemble flou d'un univers discret tel que $\varpi[X] = \sum_{\xi \in V} \mu_x(\xi_i)$. Son extension dans un univers continu vaut $\varpi[X] = \int_V \mu_x(\xi_i).d\xi$. Dans un interval fixe, la formule est alors $\varpi[X]$ = borne.minimale(X)-borne.maximale(X). Soient deux intervalles A=$[a^-, a^+]$ and B=$[b^-, b^+]$, la distance est définie comme valant :

$$\delta[A, B] = max[0, (max\{a^-, b^-\} - min\{a^+, b^+\})]$$

et la définition de la mesure de similarité entre deux intervalles flous s I_1 et I_2 est donnée par :

$$S(I_1, I_2) = \frac{1}{2}(\frac{\sum_V \mu_{I_1 \cap I_2}(\xi_i)}{\sum_V \mu_{I_1 \cup I_2}(\xi_i)} + 1 - \frac{\delta[I_1, I_2]}{\varpi[V]})$$

dans le cas discret, et :

$$S(I_1, I_2) = \frac{1}{2}(\frac{\int_V \mu_{I_1 \cap I_2}(\xi_i).d\xi}{\int_V \mu_{I_1 \cup I_2}(\xi_i)d\xi} + 1 - \frac{\delta[I_1, I_2]}{\varpi[V]})$$

dans le cas continu. La similarité combine la mesure de similarité de Jaccard [Jaccard, 1908], qui calcule la similarité de deux intervalles se chevauchant, et un second terme qui perme de prendre compte deux intervalles strictement distincts.
Si seuls les intervalles non flous sont considérés, la mesure de similarité peut être simplifiée selon :

$$S(I_1, I_2) = \frac{1}{2}(\frac{\varpi[I_1 \cap I_2]}{\varpi[I_1 \cup I_2]} + 1 - \frac{\delta[I_1, I_2]}{\varpi[V]})$$

3.1.2 Degré d'Appartenance Global (GAD)

Le GAD représente le degré d'appartenance d'un individu à une classe. Lorsque tous les MADs sont calculés pour un individu j, c'est-à-dire, lorsqu'à chacun de ses descripteurs a été affecté un vecteur de dimension K, un vecteur de K GADs est calculé pour j, de manière à décider de quelle classe il est le plus proche. Pour cela, considérant les vecteurs $\overrightarrow{v_j^1}$, $\overrightarrow{v_j^2}$, ... $\overrightarrow{v_j^d}$ attribués respectivement aux d descripteurs de j, le vecteur $\overrightarrow{v^j}$ représentant les GAD de j aux classes $k_1, k_2, ... k_K$ est calculé en appliquant systématiquement la fonction d'agrégation η sur les MADs de j aux différentes classes. Ainsi, le calcul du GAD de j à la classe k est opéré selon la formule :

$$GAD_j^k = \eta(MAD(\mu_j^1 \mid k), MAD(\mu_j^2 \mid k), ..., MAD(\mu_j^d \mid k))$$

avec $MAD(\mu_j^i \mid k)$ = MAD du j^e individu à la classe k pour la i^e caratéristique.

Les connectifs de la logique floue sont les versions floues des connectifs de la logique binaire, en particulier l'intersection (ET) et l'union (OU). η correspond ainsi à une interpolation entre l'opérateur logique d'intersection (la T-norme) et celui de l'union (la T-conorme) qui en est la fonction duale. Concrètement, en introduisant un paramètre $\alpha \in [0, 1]$ décrivant l'indice d'exigence de la classification, on obtient la formule :

$$GAD_j(MAD_1, MAD_2, ...MAD_d) = \alpha TN(MAD_1, MAD_2, ...MAD_d) + (1-\alpha)TC(MAD_1, MAD_2, ...MAD_d)$$

TN et TC sont définis respectivement comme la T-norme et la T-conorme.

Il est possible de choisir entre la fonction probabiliste et la fonction possibiliste, auquel cas les fonctions sont respectivement :

– TN(a,b)=a.b et sa fonction duale TC(a,b)=a+b-a.b,

– TN(a,b)=min(a,b) et sa fonction duale TC(a,b)=max(a,b).

Ainsi, plus α est élevé, plus le GAD calculé sera faible, puisqu'alors un poids plus important sera attribué à la fonction TN, décrivant l'intersection. A l'inverse, une classification réalisée avec α faible sera beaucoup plus permissive.

3.2 Classification selon LAMDA

La méthode LAMDA, dont le fonctionnement est représenté figure 22, permet de classer des données hétérogènes tant en mode supervisé que non supervisé selon le fonctionnement décrit dans la figure

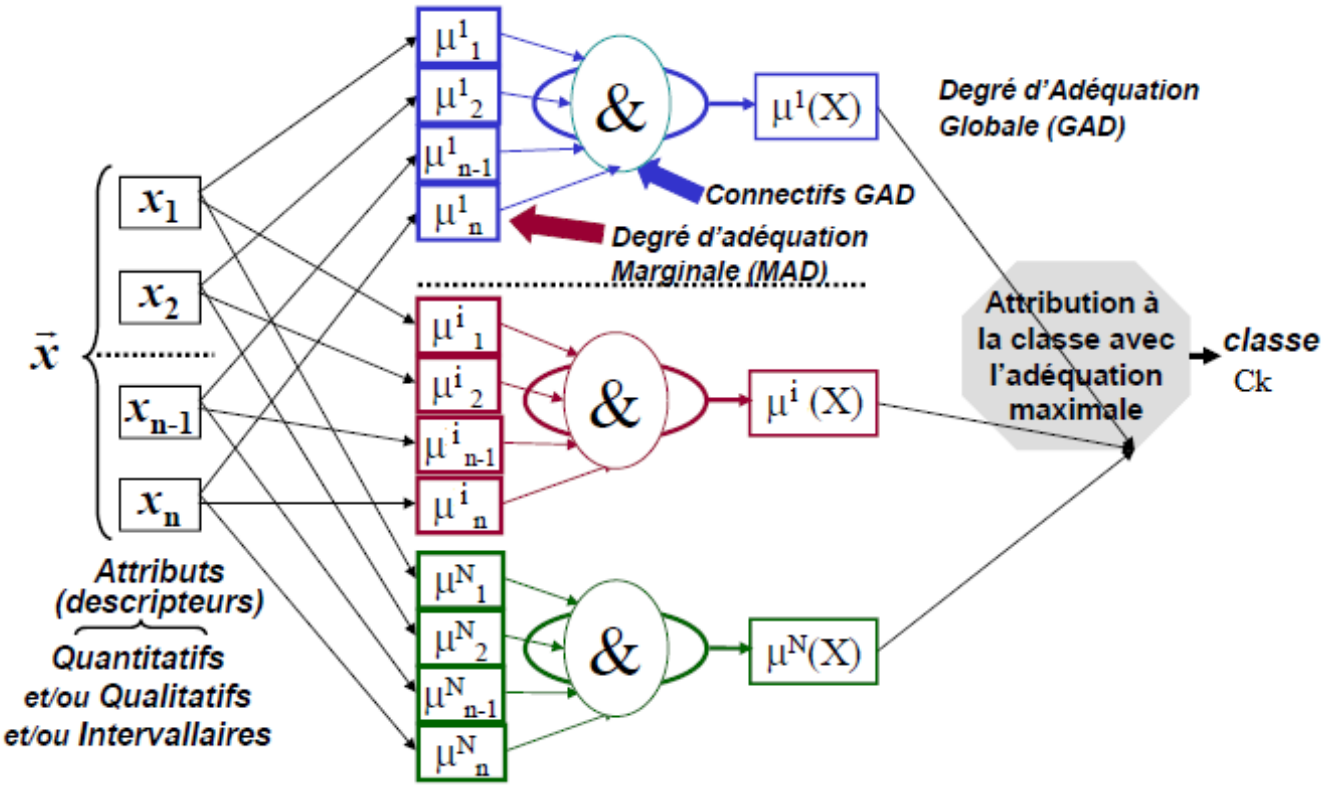


Figure 22. Fonctionnement de LAMDA

3.2.1 Classification non-supervisée

Dans le cas d'un problème de classification non-supervisée ou de coalescence, il s'agit d'opérer des groupements des individus d'une liste donnée selon leur similarité ; le nombre d'individus J de la liste doit être au moins égal à trois, pour qu'une telle classification ait un sens. Soit λ une variable préalablement définie, décrivant la valeur minimale de proximité, c'est-à-dire la valeur de GAD en-dessous de laquelle un individu est considéré comme trop éloigné d'une classe.La méthode LAMDA décrit l'algorithme suivant :

1. Création d'une classe à partir de la description du premier élément de la liste
2. Pour chaque élément j qui suit, calcul du GAD à chacune des classes déjà existantes. Soit K le nombre de classes existantes (K vaut 1 lors du traitement du deuxième élément de la liste) :
   – $\forall k \in K$, si $GAD_j(k) \geq \lambda$ alors j est affecté à la classe k pour laquelle GAD_j est maximisé et la classe k est redéfinie en fonction de tous les éléments qui la composent.
   – Si non création d'une classe à partir de la description de j.
3. Lorsque tous les éléments de la liste sont classés, calcul des K GAD pour tous les individus de la liste et ré-affectation si nécessaire.

3.2.2 Classification supervisée

Dans le cas d'une classification supervisée, on dispose d'une liste de J individus dont on connait la partition a priori mais pas la description précise des classes. Elle vise donc à définir les différentes classes à partir des caractéristiques des individus qui les composent. Ce type de classification consiste à calculer les moyennes, les variances, les fréquences,... des différentes modalités afin de créer des profils types, c'est-à-dire des classes représentatives de la population et de sa répartition et décrites par des descripteurs types.

3.2.3 Reconnaissance

Lors d'une reconnaissance, le nombre des individus de la liste n'est pas contraint : les classes sont déjà définies, le but est simplement d'affecter les individus à la classe dont chacun d'eux est le plus proche. Ainsi, l'algorithme décrit par LAMDA consiste à calculer les GADs de chaque individu j de la liste aux K classes définies et d'affecter j à la classe dont il est le plus proche - c'est-à-dire celle qui permet de maximiser son GAD. Il est possible d'accepter la potentielle existence d'une classe annexe, dans laquelle seront placés tous les individus j pour lesquels $\forall k \in K$, $GAD_j(k) < \lambda$. Cette classe est appelée NIC, pour Non Informative Class. Il est possible, à partir de cette classe et des éléments qui la composent, d'effectuer ultérieurement un apprentissage dit spécialisé, en créant une nouvelle classe pour ces individus.

3.3 Sélection des descripteurs

Toutes les caractéristiques n'ont pas la même importance dans un problème de classification, mais il peut être difficile de déterminer a priori quelles sont les informations réellement discriminantes. La méthode MEMBAS (MEmbership Margin Based feAture Selection), intégré à LAMDA, permet d'évaluer le rôle que chacun d'eux a joué dans la classification ; cette connaissance permet à l'utilisateur de saisir les mécanismes de la classification ayant eu lieu [Hedjazi *et al.*, 2010]. L'idée générale de MEMBAS est d'attribuer un poids plus important aux descripteurs dont la valeur moyenne dans nk est la plus différente de celle dans $n\bar{k}$, puisque ce sont eux qui ont joué le plus grand rôle dans la formation de deux classes distinctes. Cette méthode repose sur la maximisation de la marge entre deux classes voisines. L'algorithme s'applique sur une classification opérée à partir d'une liste L de J données d'entraînement et s'appuie sur le principe de marge d'appartenance, réinvestissant ainsi la notion de marge maximale développée à travers les SVM. Le poids d'un descripteur d est défini à partir de la différence entre les MAD de d pour tous les individus sélectionnés par l'algorithme à la classe dont chacun d'eux est le plus proche, et le MAD de d pour ces mêmes individus à leur deuxième classe la plus proche. Plus cette différence est élevée, plus le descripteur témoigne d'un poids élevé pour la classification ; il permet une distinction d'autant plus facile entre deux classes proches.

La maximisation de la marge est donnée par la formule :

$$\max_{w_f} \sum_{j=1}^{J} \beta_j(w_f) = \frac{1}{N} \sum_{j=1}^{J} \begin{cases} \sum_{i=1}^{N} w_{fi} \mu_c^i(x_i^j)) \\ -\sum_{i=1}^{N} w_{fi} \mu_{\bar{c}}^i(x_i^j)) \end{cases}$$

sous les contraintes : $|| w_f ||_2^2 = 1, w_f \geq 0$
La première contrainte correspond à la borne normalisée pour le modulo de w_f, de sorte que la maximisation s'achève avec des valeurs finies. La seconde contrainte, quant à elle, assure que le vecteur poids obtenu ne soit défini que par des valeurs positives. Elles peuvent être simplifiés ainsi :

$$\max_{w_f} (w_f)^T s$$

$$Soumis \quad à : || w_f ||_2^2 = 1, \quad w_f \geq 0$$

$$où \quad s = \frac{1}{m} \sum_{n=1}^{N} \{U_{nc} - U_{n\bar{c}}\}$$

On considère qu'il existe au moins une caractéristique i<D telle que s_i>0.

La gestion de ce problème classique d'optimisation par la méthode des multiplicateurs de Lagrange a permis d'obtenir finalement une solution analytique :

$$w_f* = \frac{s^+}{|| s^+ ||}$$

avec $s^+ = [max(s_1, 0), ..., max(s_d, 0)]^T$.

MEMBAS suit les étapes suivantes :

1. Initialisation du vecteur de poids à zéro.
2. Pour t=1 jusqu'à T :
   (a) Sélection aléatoire d'un individu j de L,
   (b) Détermination de la partition floue de chaque caractéristique,
   (c) Calcul des vecteurs U_{nk} et $U_{n\bar{k}}$ pour l'individu j,
   (d) Mise à jour du vecteur s tel que s= s+ $\{U_{nc} - Un\bar{c}\}$
3. Calcul du vecteur de poids flou optimal tel que $w_f^* = \frac{s^+}{||s||}$.

avec :
- T : nombre d'itérations tel que $T \leq J$,
- U_{nk} : Vecteur des MADs de j envers la classe k telle que $GAD_j(k)$ est maximisé,
- $U_{n\bar{c}}$: Vecteur des MADs de j envers la classe $\bar{c}$ telle que $GAD_j(c) - GAD_j(\bar{c})$ est minimisé,
- $s^+ = [max(s_1, 0), max(s_2, 0), ...max(s_m, 0)]^T$.

3.4 Evaluation de la qualité de la partition de l'espace de données

Dans le domaine de l'intelligence artificielle, l'analyse de la validité d'une partition est généralement basée sur la qualité de la séparation et de la cohésion des classes [Bezdek, 2013]. Les critères retenus pour évaluer la qualité d'une partition concernent la compacité des classes, leur séparation, et le nombre de classes - car une classification affectant chaque élément n'apporterait pas de réponse significative au problème initial mais serait d'une compacité idéale. Dans cette optique, la qualité d'une partition sera fonction de :
- La proximité des individus intra-classes : plus les individus ayant intégré une classe seront semblables, plus elle sera compacte.
- L'éloignement des individus inter-classes : plus les individus appartenant à deux classes différentes seront dissemblables, plus ces classes seront séparées.

Ainsi, en appliquant ces présupposés à la méthode LAMDA, on considère que dans une partition de bonne qualité, les individus ont un GAD très élevé avec la classe à laquelle ils appartiennent, et un GAD le plus faible possible avec les autres. La formule employée pour évaluer la qualité de la partition est donc [Narvaez, 2007] :

$$CV = \frac{Dis}{N}.D^*_{min}.\sqrt{K}$$

Dis représente la dispersion obtenue par la formule :

$$Dis = \sum_{k=1}^{K} 1 - \frac{\sum_{n=1}^{N} \delta_{kn}.exp(\delta_{kn})}{N.GAD_{Mk}.exp(GAD_{Mk})}$$

où

- $\delta_{kn} = \mu_{Mk} - \mu_{kn}$,
- $\mu_{Mk} = max[\mu_{kn}]\ \forall n \in [1, D]$, D étant le nombre de descripteurs d'un individu,
- N : nombre total d'individus ayant participé à la définition des classes,
- K : nombre de classes
- D^*_{min} : distance minimale entre deux classes, calculée à partir de la distance d*(A,B) entre deux intervalles flous selon la formule :

$$d^*(A,B) = 1 - \frac{M[A \bigcap B]}{M[A \bigcup B]} = 1 - \frac{\sum_{n=1}^{N} min(MAD_A^n \bigcap MAD_B^n)}{\sum_{n=1}^{N} max(MAD_A^n \bigcap MAD_B^n)}$$

3.5 Conclusion

La méthode LAMDA permet de gérer aussi bien l'apprentissage supervisé que la coalescence, et est capable de gérer des types de données hétérogènes grâce au principe SMSP, qui permet :

- La projection des données hétérogènes dans un espace de données homogène par le calcul des MADs.
- La classification des données par calcul des GAD.

La projection des différentes données dans un même espace homogène permet leur traitement simultané et équitable, de sorte que chaque donnée peut initialement avoir le même poids. Par la suite, si certaines caractéristiques s'avèrent plus discriminantes que d'autres, MEMBAS permet de connaître le poids de chacune d'elles dans la classification obtenue.

Son utilisation nécessite un travail en amont puisqu'il est important de :

- Sélectionner les données et leur type de manière pertinente et appropriée au problème soumis,
- Définir la valeur d'α en fonction des contraintes et des objectifs de la classification.

Dans le prochain chapitre, nous expliquerons quelles améliorations nous avons apportées à la méthode LAMDA et les raisons qui nous y ont poussés.

4 Traitement de données manquantes ou multidimensionnelles

En tant que méthode de classification, LAMDA vise à dessiner une partition la plus représentative possible de la réalité. Plusieurs solutions peuvent généralement être proposées relativement à chaque problème et chaque situation, sans qu'aucune d'elles ne puisse décisivement se targuer d'être la meilleure, ni parfaitement conforme à la réalité ; il s'agit d'une manière d'organiser les informations connues pour en fournir une interprétation possible et intelligible. [Roux *et al.*, 2015]

Dans le but d'élire une représentation particulière, des méthodes permettent d'évaluer sa pertinence - comme celle décrite dans la section III.4, qui mesure la qualité mathématique de la partition. Néanmoins, elle ne dit rien de sa représentativité sémantique puisqu'elle prend en compte les propriétés géométriques de la classification et non les propriétés particulières des individus classés. Or, la qualité d'une partition ne dépend pas que de la compacité intra-classe et de la distance inter-classes : elle est directement liée au choix des attributs sélectionnés pour décrire les individus. La représentativité et la pertinence des descripteurs vis-à-vis du problème auquel prétend répondre la classification sont les deux critères décisifs de l'intérêt que peut présenter la partition obtenue et sa significativité. Pour cela, il est important de lever les écueils pouvant échoir à la sélection des attributs :

- L'abandon d'informations manifestement importantes mais dont l'obtention systématique est douteuse,
- La séparation de données pourtant indubitablement liées,
- La perte d'informations due au typage.

Un type décrit un ensemble de caractères d'abord connus comme distincts puis structurés en un tout servant d'instrument de connaissance à la suite d'une abstraction rationnelle et permettant d'élaborer ou de distinguer des catégories. Quel que soit le domaine concerné, le typage décrit déjà une étape de classification, puisqu'il s'agit de décrire les différents éléments par types - typer un objet, c'est le marquer d'une empreinte correspondant à certaines caractéristiques : la biologie catégorise les différentes espèces vivantes par types, la médecine définit le typage cellulaire, et le traitement de données, en informatique, nécessite un typage dans bien des langages. Dans ce dernier cas, les types de base qu'offrent les langages de programmation correspondent aux données qui peuvent être traitées directement par le processeur — c'est-à-dire sans conversion ou formatage préalable. Par la suite, les types composés permettent de grouper plusieurs champs de types distincts dans une même variable (qui contient alors des "sous-variables"). En programmation orientée objet, on nomme classe un type composé associé à du code spécifique — la définition des méthodes de la classe — propre à la manipulation de variables de ce type. La programmation orientée objet étend le paradigme précédent en organisant hiérarchiquement les classes de telle manière qu'une variable d'une sous-classe puisse être utilisée de manière transparente à la place de n'importe quelle classe située à un niveau supérieur dans la hiérarchie définie.

Ici, définir des types pour les données consiste à les mouler dans des types contraints, ce qui peut occasionner la perte de l'information de leurs contraintes particulières. C'est pourquoi il nous a sem-

blé intéressant d'étendre les types existants afin de pouvoir disposer de modèles plus complets pour représenter des données complexes sans perte d'information.

Dans ce chapitre, nous allons présenter les solutions que nous proposons pour pallier ces contraintes, et limiter leur incidence sur la significativité du processus de classification :

- Un nouveau type de données permettant la représentation de données multidimensionnelles par un traitement multicouche,
- Une comparaison enrichie des données qualitatives, de manière à discerner leur proximité entre elles de manière floue et non plus binaire,
- La possibilité de prendre en compte des données incomplètes.

4.1 Classification multicouche

Dans le monde réel, au sein d'un même système, on peut observer des entités distinctes et complexes. Ainsi, lors d'une classification, les caractéristiques peuvent elles aussi être complexes et ne pas pouvoir être simplement décrites par un simple nombre, mot, ou intervalle. Elles peuvent être un objet signifiant du monde réel, ou le résultat d'une combinaison de différents facteurs. De même, les caractéristiques qui la composent peuvent être liées entre elles, et être par exemple obtenues à partir d'un ensemble d'observations conjointes, ou au contraire être relativement hétérogènes et être transmises par des sources diverses.

Par exemple, un individu pourrait être décrit par la couleur de ses cheveux, de ses yeux, de sa peau... Chacune de ces caractéristiques proviendraient ainsi d'observations simples et pourraient être représentées par un type simple - quantitatif, qualitatif, ou intervallaire en fonction de la précision et des contraintes spécifiques au problème de classification. Si, par la suite, on veut ajouter à sa description une information relative à son état de santé, il faudra ajouter à ses caractéristiques un attribut décrivant cet aspect de son individualité ; cet attribut résulte manifestement d'un ensemble de caractéristiques variées, relatives, par exemple, à sa tension artérielle, son taux de cholestérol,.. qu'il faudra exprimer et combiner ensemble pour parvenir à une description personnalisée de sa santé biologique. Chacune de ces informations biologiques auraient été obtenues à partir de différentes analyses réalisées dans un hôpital ou un centre d'analyses, par exemple, donc provenir d'une seule base de données - mais distincte de celle permettant de déterminer les caractéristiques physiques observables à l'oeil nu. Par la suite, il est à nouveau possible de séparer la caractéristique "état de santé" en deux groupes, qui seraient respectivement les attributs "santé biologique" et "santé psychologique", ces dernières pouvant par exemple résulter de tests effectués par des spécialistes dans des cabinets de psychologues, et proviendraient ainsi d'une nouvelle base de données. Ainsi, toutes ces différentes informations proviendraient d'instruments de mesure variés, qui n'auraient pas nécessairement la même précision, et demanderaient néanmoins à être traitées ensemble. Cette contrainte nécessite l'introduction d'un nouveau type de données, capable de représenter dans un même objet des caractéristiques différentes, d'un niveau de détail varié, et obtenues à partir de types de mesure distincts.

La classification multicouche, dont l'architecture est présentée figure 23, permet de classer des individus dont les caractéristiques peuvent être décrites par un des trois types de base ou par une agrégation de plusieurs caractéristiques. Ces caractéristiques constitutives d'une caractéristique plus complexe peuvent, à leur tour, être de type quantitatif, qualitatif, intervallaire, ou encore être des entités composites. Ce type composite permet la prise en compte de données multi-dimensionnelles par la combinaison de plusieurs données, de types potentiellement différents, en une même entité. Chaque attribut d'une caractéristique composite est constitué d'une valeur typée associée à un poids. Une classe obtenue à la couche c-1 ne sera considérée que comme un seul descripteur à la couche c. Si, à la couche c, les individus possèdent un attribut i_c décrit par un type composite, cet attribut sera nécessairement traité auparavant lors d'une classification opérée au niveau de la couche c-1. Si cet attribut i_c est lui-même décrit par au moins un attribut de type composite i_{c-1}, une couche c-2 sera conséquemment impliquée pour la classification de i_{c-1}, et ainsi de suite. En d'autres termes, il y a autant de couches que de descripteurs composites imbriqués.

Le calcul du degré d'appartenance de la caractéristique composite i_c d'un individu j à la caractéristique i_c^k d'une classe k, au niveau de la couche c, est décrit par la formule :

$$\mu_k^{i_c}(x_{i_c}) = \sum_{i_{c-1}}^{I_{c-1}} (MAD_k^{i_{c-1}}.\bar{w}_{i_{c-1}})$$

avec :

- I_{c-1} : Nombre de caractéristiques constituant la caractéristique composite i_c,

– $\bar{w}_{i_{c-1}} \in [0,1]$: Poids normalisé du poids $w_{i_{c-1}}$ de la i_{c-1}e caractéristique de i_c, obtenu par MEMBAS,
– MAD_k^i : MAD de la caractéristique i_{c-1} de i_c à la caractéristique i_{c-1}^k de la caractéristique i_c^k de la classe k.

Ce type composite présente deux principaux avantages, puisqu'il permet de :
– Considérer les attributs distincts d'une même entité comme un ensemble consistant : il s'agit de représenter un objet signifiant du monde réel par une combinaison pondérée de caractéristiques,
– Traiter au cours d'une même classification un ensemble d'informations issues de base de données différentes.

L'intégration d'une caractéristique composite à une classification nécessite une première classification des différents individus décrits par les attributs de cette caractéristique.

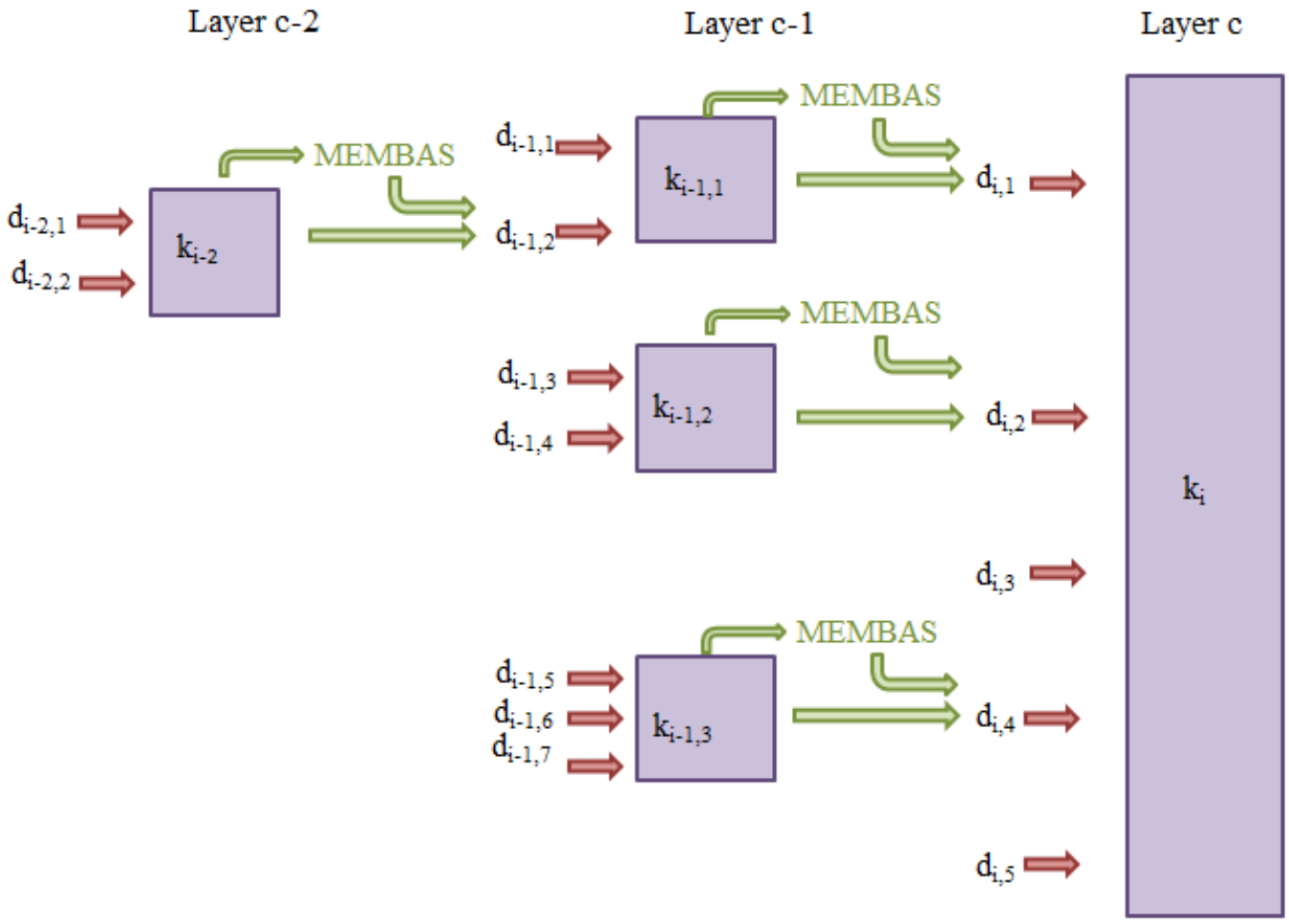


Figure 23. Architecture de la classification multicouche.

Soit c une classification à réaliser, d'une liste L_c de J individus à classer, et soit $i_{c,1}$ une des I_c caractéristiques des J individus constituant L_c, $i_{c,1}$ étant de type composite, pour chaque individu j_c de L_c, la valeur de $i_{c,1}$ est considérée comme un individu $j_{c-1,1}$ d'une liste $L_{c-1,1}$ contenant J individus à classer. Il y a autant de listes L_{c-1} que C contient de caractéristiques de type composite. La classification C-1 a lieu avant C ; elle mène à la définition de K classes K_{c-1} et son résultat est transmis directement à la couche C. MEMBAS permet de calculer le poids de toutes les caractéristiques constituant $i_{c,1}$, de manière à les combiner en un objet pouvant être intégré directement à la classification C. Soit $I_{c-1,i_{c,1}]}$ le nombre de descripteurs constituant le descripteur $i_{c,1}$, MEMBAS fournit donc, pour la classification des J descripteurs $i_{c,1}$ un vecteur poids de dimension $I_{c-1,i_{c,1}]}$. Un descripteur composite $i_{c,1}$ correspond

donc à l'association du vecteur poids de dimension $I_{c-1,i_{c,1}]}$ et de la définition de la classe $K_{c-1,k}$ la plus proche de cet attribut. Le principe de fonctionnement de la classification multicouche, considérant deux couches, est décrite figure 24.

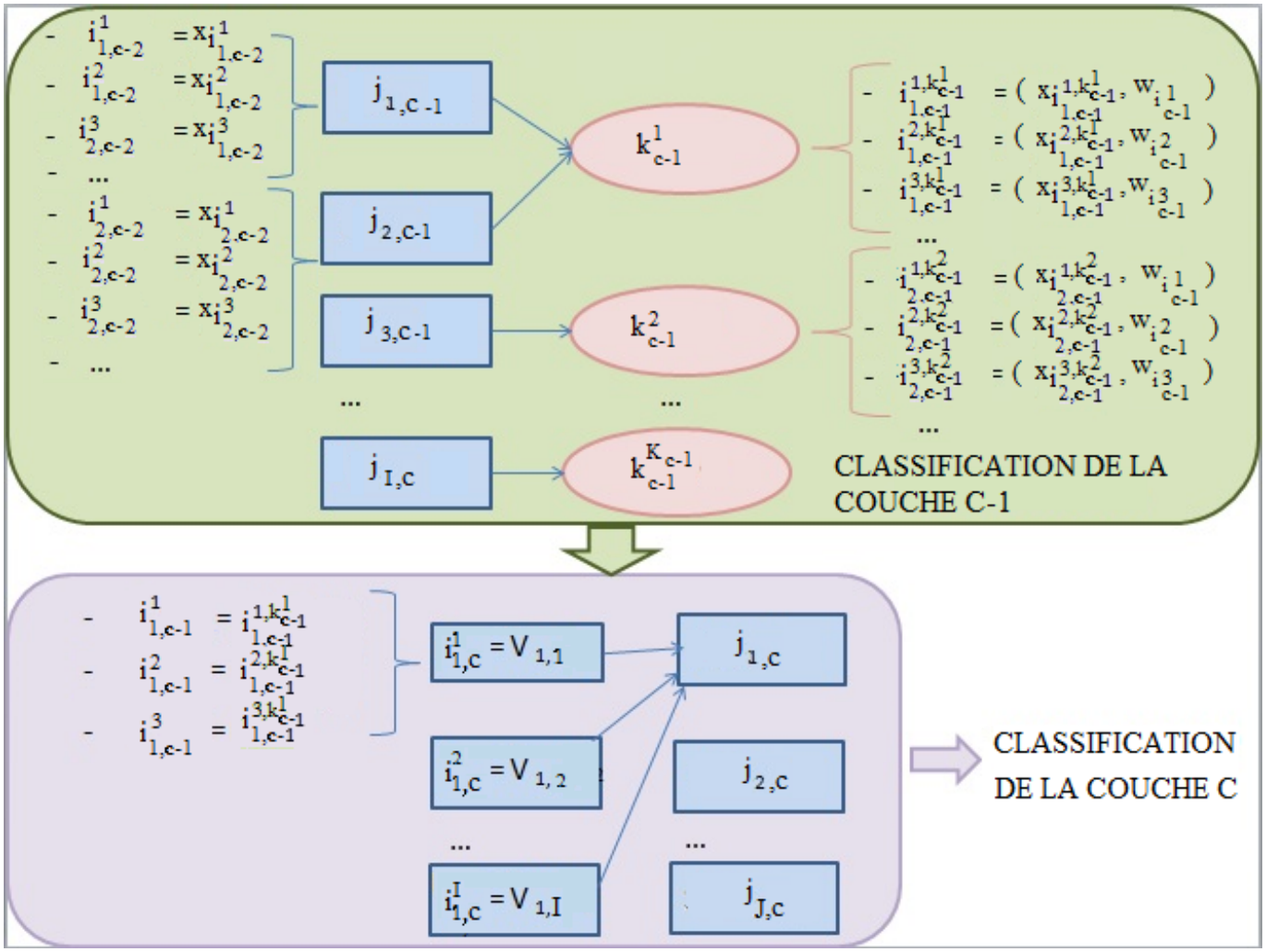


Figure 24. Description du fonctionnement de la classification multicouche.

Pour bien comprendre l'intérêt de cette approche, prenons comme exemple la population présentée dans le tableau 2, décrivant, de manière simplifiée, les conditions et les résultats d'une expérience de psychologie visant à comprendre les mécanismes impliqués dans l'échec ou la réussite à une épreuve. Il s'agissait d'un test en deux temps :

- Faire subir une première épreuve à tous les participants, une moitié en ayant une facile et l'autre moitié étant placée en face d'un examen impossible à réaliser -sans qu'ils soient mis au courant de l'insolubilité du problème. Les résultats à cette première épreuve sont indiqués par les modalités "oui" et "non" de la variable "Réussite_epreuve1",
- Placer tous les participants en face d'une même seconde épreuve résolvable, la moitié des étudiants effectuant cette tâche dans le calme, et l'autre moitié avec une musique désagréable. La présence ou non de l'atmosphère musicale est précisée respectivement par les modalités "oui" et "non" relativement à la variable "Musique". Les résultats des sujets à cette seconde épreuve sont reportés dans la colonne "Réussite_epreuve2", avec "oui" s'ils ont réussi et "non" dans le cas contraire.

Des informations relatives à chaque sujet sont renseignées à la fin du tableau : leur âge, le domaine des études qu'ils suivent ou ont suivies ("L" pour Littéraire", "S" pour Scientifique, "ES" pour "Economique et Social"), et leur sexe ("H" pour Homme et "F" pour Femme"). L'hypothèse de cette étude est que

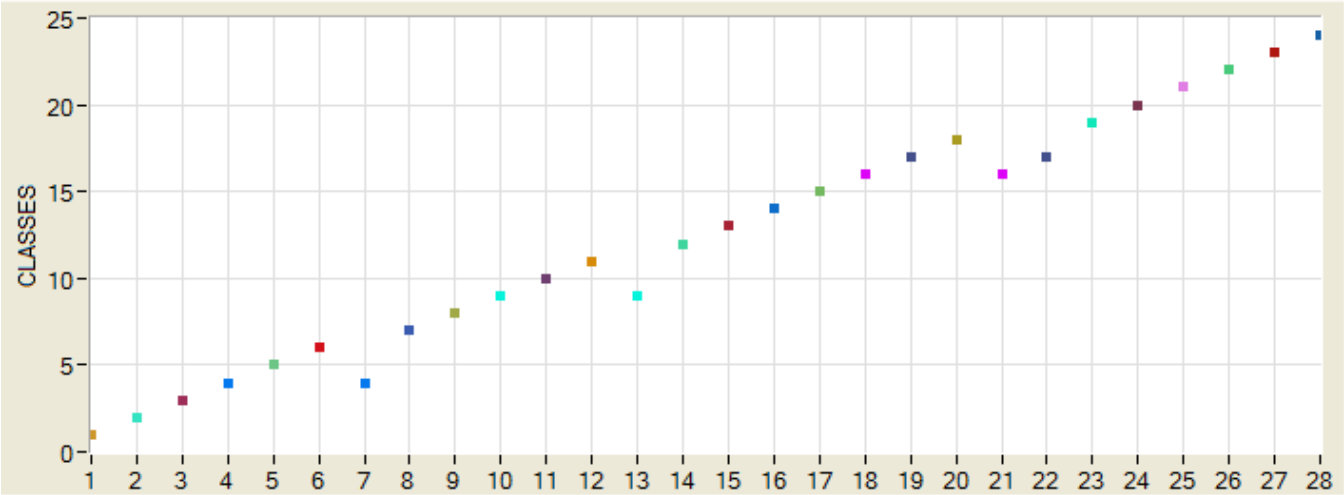
le résultat à l'épreuve 2 est indifférent aux caractéristiques individuelles indiquées, mais est directement lié au sentiment d'efficacité personnelle flatté ou atteint respectivement par la réussite ou l'échec lors de la première épreuve : plus un sujet se croit capable de réussir, plus il se donne les moyens d'y parvenir. La musique, envisagée comme étant un facteur pénalisant puisque déconcentrant, peut avoir un effet positif sur les résultats puisqu'alors la responsabilité de l'échec peut être attribuée à une cause externe. Le résultat de l'expérience confirme cette hypothèse.

Musique	Réussite_epreuve1	Réussite_epreuve2	Age	Filière	Sexe
oui	non	oui	20	S	H
oui	non	non	25	ES	H
oui	non	oui	22	L	H
oui	non	oui	18	S	F
oui	non	oui	34	ES	F
oui	non	oui	29	L	F
oui	non	oui	38	S	F
oui	oui	oui	29	ES	F
oui	oui	non	41	L	F
oui	oui	oui	19	S	H
oui	oui	non	21	ES	H
oui	oui	non	39	L	H
oui	oui	oui	24	S	H
oui	oui	oui	43	ES	H
non	oui	oui	22	L	H
non	oui	oui	19	S	H
non	oui	oui	48	ES	H
non	oui	oui	43	L	F
non	oui	oui	28	S	F
non	oui	non	22	ES	F
non	oui	oui	47	L	F
non	oui	oui	30	S	F
non	non	non	18	ES	F
non	non	non	53	L	H
non	non	non	23	S	H
non	non	non	18	ES	H
non	non	oui	53	L	H
non	non	oui	45	S	H

Tableau 2. Exemple d'un échantillon d'individus à classer pour la comparaison entre classification simple et classification multicouche.

La figure 25 montre le résultat de la classification opérée de manière simple, c'est-à-dire en considérant chaque caractéristique comme égale à elle-même : qu'il s'agisse de la musique, des résultats lors de la première épreuve, ou des caractéristiques individuelles, tout est envisagé simultanément comme les conditions initiales équivalentes de l'expérience. La qualité de la partition obtenue est évaluée à 0.16 - α

ayant été paramétré pour que l'indice de qualité soit maximisé. Les caractéristiques individuelles étant relativement hétérogènes, le classifieur n'a pas pu opérer de groupements cohérents et a livré une partition de 24 classes pour 28 individus ; les résultats de l'expérience sont difficilement analysables dans ces conditions.



a. Répartition des individus dans les classes

	Musique	Réussite_epreuve1	Réussite_epreuve2	Age	Filière	Sexe
Classe 1	oui	non	oui	20	S	H
Classe 2	oui	non	non	25	ES	H
Classe 3	oui	non	oui	22	L	H
Classe 4	oui	non	oui	28	S	F
Classe 5	oui	non	oui	34	ES	F
Classe 6	oui	non	oui	29	L	F
Classe 7	oui	oui	oui	29	ES	F
Classe 8	oui	oui	non	41	L	F
Classe 9	oui	oui	oui	21.5	S	H
Classe 10	oui	oui	non	21	ES	H
Classe 11	oui	oui	non	39	L	H
Classe 12	oui	oui	oui	43	ES	H
Classe 13	non	oui	oui	22	L	H
Classe 14	non	oui	oui	19	S	H
Classe 15	non	oui	oui	48	ES	H
Classe 16	non	oui	oui	45	L	F
Classe 17	non	oui	oui	29	S	F
Classe 18	non	oui	non	22	ES	F
Classe 19	non	non	non	18	ES	F
Classe 20	non	non	non	53	L	H
Classe 21	non	non	non	23	S	H
Classe 22	non	non	non	18	ES	H
Classe 23	non	non	oui	53	L	H
Classe 24	non	non	oui	45	S	H

b. Description des classes

Figure 25. Résultats de la classification simple

L'échantillon d'apprentissage est trop hétérogène pour constituer une partition significative, puisque les caractéristiques "Réussite_épreuve1", "Réussite_épreuve2", "Âge", "Filière", "Sexe" sont totalement indépendantes les unes des autres et témoignent d'une distribution aléatoire. Ainsi, une classification prenant en compte tous les attributs, sans réelle possibilité de déterminer quels descripteurs sont finalement liés, si certains devraient être privilégiés, etc... Typiquement, les descripteurs "Réussite_épreuve1", "Réussite_épreuve2", et "Sexe" sont des descripteurs de type qualitatif pouvant prendre deux valeurs distinctes, et la fréquence d'apparition de leurs deux modalités est à peu près équitable dans cet échantillon, aussi le classifieur n'est-il pas en mesure de décider de privilégier un attribut par rapport aux autres. Dans ce cas, en fonction du α choisi, le classifieur peut suivre différentes stratégies :

- Si la distribution de la population le permet, choisir un ou plusieurs attributs au détriment des autres, de manière à maximiser la qualité de la partition,
- Si la distribution des individus est trop dispersée et qu'aucun choix n'est réellement meilleur que les autres, réaliser cette classification en fonction d'un ou plusieurs attributs sélectionnés aléatoirement et laisser les autres de côté,
- Ne faire aucun choix et prendre tous les attributs en compte.

La première méthode présente l'avantage de proposer la meilleure partition possible, mais n'est pas toujours applicable. Concernant la deuxième méthode, elle propose une solution de partition mais n'est basée sur aucun critère rationnel et risque d'occulter des paramètres cruciaux. Quant à la dernière stratégie, la partition fournie se montre bien souvent très pauvre en informations - si ce n'est celle de l'hétérogénéité de l'échantillon. Dans notre situation, c'est cette méthode qui a été choisie, la première n'étant pas applicable, et la seconde n'apportant rien d'intéressant.

La figure 26 montre le résultat de la classification de l'échantillon présenté dans le tableau 2 en utilisant le mode multicouche. Il s'agit d'un exemple de partition pouvant être générée, mais le classifieur est capable de fournir diverses partitions, permettant de chercher celle qui répondra au mieux aux besoins de l'analyse des résultats de l'étude - ce qui n'était pas vraiment ajustable lors de la classification simple. Par ailleurs, l'indice de qualité de la partition obtenue vaut 0.51, ce qui est largement supérieur à la qualité maximale obtenue précédemment. La représentation des caractéristiques individuelles par le type composite a permis au classifieur de les interpréter comme liées, et de les traiter isolément lors d'une première classification au niveau de la couche $c-1$ pour ensuite les intégrer à la classification ayant cours dans la couche c.

La possibilité de considérer les qualités individuelles comme une condition initiale unique de l'expérience permet une meilleure visibilité des résultats et de les classer plus clairement. Ici, la couche c concerne la classification des 28 individus tels qu'ils sont présentés dans le tableau 2, et caractérisés par 4 attributs :

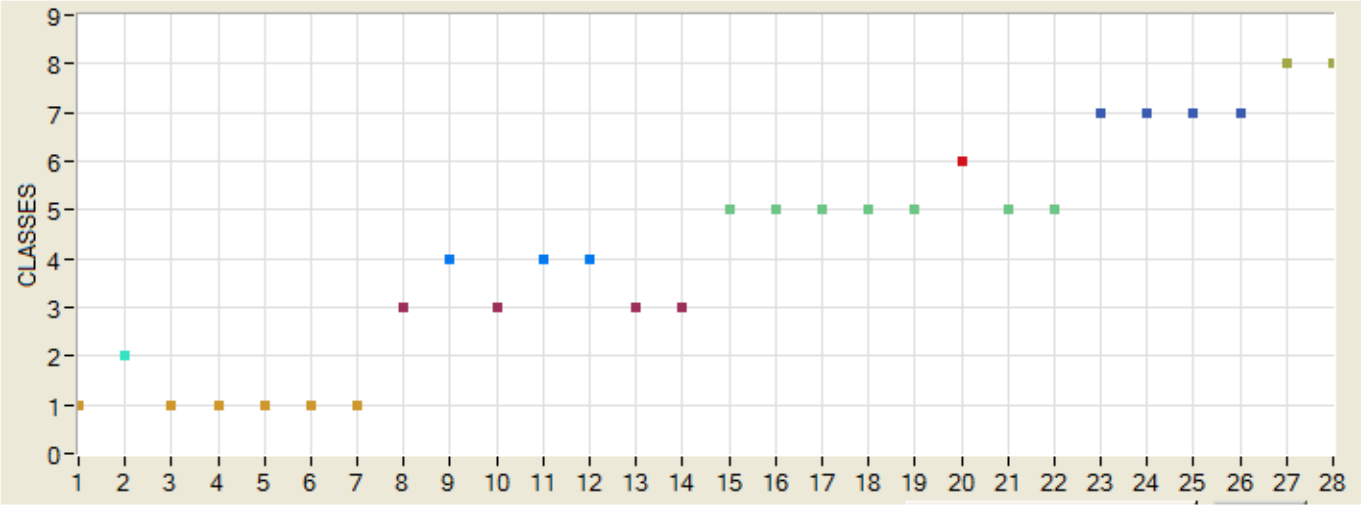
- La présence, ou non, d'une musique gênante ("Musique", attribut qualitatif),
- La réussite, ou non, à la première épreuve ("Réussite_epreuve1", attribut qualitatif),
- La réussite, ou non, à la deuxième épreuve ("Réussite_epreuve2", attribut qualitatif),
- Ses caractéristiques individuelles ("Age", "Filière", "Sexe", attribut composite).

Quant à la couche c-1, elle s'applique aux 28 individus obtenus à partir de la sélection de l'attribut relatif aux caractéristiques individuelles, pour opérer une première classification. Ainsi, chacun d'eux est intégralement constitué par l'attribut "caractéristiques individuelles" de l'individu de la couche c qui lui correspond $i_1(c-1)$ correspond à $i_1(c)$, $i_2(c-1)$ correspond à $i_2(c)$,... $i_28(c-1)$ correspond à $i_28(c)$). Chaque individu de c-1 est donc caractérisé par 3 caractéristiques qui sont :

- Son âge ("Age", attribut quantitatif),
- La filière des études suivies ("Filière", attribut qualitatif),
- Son sexe ("Sexe", attribut qualitatif).

La couche $c-1$ a ainsi pu fournir à la couche c le résultat de la partition obtenue, en termes de

définition de classes d'une part, et de pondération des attributs constitutifs de l'attribut "caractéristiques individuelles" d'autre part. Au niveau de la couche c, pour chaque individu, l'attribut "caractéristiques individuelles" a été remplacé par la définition de la classe la plus proche obtenue au niveau de la couche c-1, associée à un vecteur poids de dimension 3 - attribuant aux caractéristiques "Âge", "Filière", et "Sexe" une valeur indiquant leur influence dans la partition.



a. Répartition des individus dans les classes

	Musique	Réussite_epreuve1	Réussite_epreuve2	Age	Filiere	Sexe
Classe 1	oui: 6	non: 6	oui: 6	27.1	ES: 1 / L:2 / S: 3	F: 4 / H: 2
Classe 2	oui:1	non:1	non:1	25	ES: 1	H: 1
Classe 3	oui: 4	oui: 4	oui: 4	31.4	ES:2 / S: 2	F: 1 / H: 3
Classe 4	oui: 3	oui: 3	non: 3	33.3	L: 2 / ES: 1	F: 1 / H: 2
Classe 5	non:7	oui:7	oui:7	35	ES: 1 /L: 3 / S: 3	F: 4 /H: 3
Classe 6	non:1	oui:1	non:1	22	ES: 1	F: 1
Classe 7	non: 4	non: 4	non: 4	28	ES:2 / L: 1 / S: 1	F: 1 / H: 3
Classe 8	non:2	non:2	oui:2	49	L: 1 / S:1	H: 2

b. Description des classes

Figure 26. Résultats de la classification multicouche

4.2 Prise en compte de la proximité des données qualitatives

Le langage, pour thématiser le monde réel, nous permet de l'appréhender d'un point de vue particulier. Chaque langage a ses particularités, représentatives de la manière dont la société le perçoit, et permet d'interpréter la réalité en la catégorisant. Pourtant, comme nous l'avons vu en décrivant les principes de la théorie de la logique floue, ce n'est pas parce que deux mots ne sont pas strictement similaires qu'ils décrivent deux objets ou deux états nécessairement distincts. En accord avec cette considération, nous nous sommes intéressés au calcul du MADs entre variables qualitatives, puisque l'algorithme ne décrivant la proximité de deux modalités qu'en binaire nous semblait trop rigide, et peu représentatif de la réalité linguistique.

Dans de nombreux cas, cet algorithme est suffisant puisque le problème initial appelle à une distinction stricte des modalités représentant des traits totalement différents : "oui/non", "bleu/rouge/jaune", ... comme c'était le cas, par exemple, dans l'exemple académique cité dans la section IV.1. Il arrive également que la proximité des données qualitatives trouve une expressivité dans la représentation quantitative ou intervallaire : par exemple, s'il s'agit de nommer des couleurs tout en prenant en compte leur position sur le spectre lumineux, la solution la plus évidente consiste à les exprimer par longueur d'onde - une longueur d'onde précise si on dispose de l'information, et un intervalle correspondant à la couleur concernée sinon. Ainsi, dans le cas d'une variable qualitative pouvant être décrite par les modalités "rouge", "bleu", "vert", si la proximité entre "bleu" et "vert" apparaît comme présentant un intérêt particulier pour la présente classification, il suffit de les remplacer par les valeurs intervallaires de leurs longueurs d'onde respectives exprimées en nm, soit par exemple : [620-700], [478-483], [510-541].

Néanmoins, il est des cas où cette conversion n'est pas possible et où la proximité doit être évaluée au cours du traitement des données qualitatives directement. Cette proximité est établie en fonction des similitudes entre les mots constitutifs des différentes modalités - car une modalité peut être constituée de plusieurs mots et être un syntagme par exemple. Pour illustrer notre propos, prenons le cas d'une variable qualitative dont certaines modalités sont un mot unique et d'autres un syntagme nominal. Il peut être intéressant de rapprocher les modalités qui se ressemblent, c'est-à-dire celles dont le mot unique est un nom similaire au noyau de certains syntagmes, ou bien celles dont le noyau syntagmatique est identique - rappelons que le noyau d'un syntagme nominal est le nom. Un syntagme nominal peut être constitué également de syntagmes adjectivaux qualifiant le nom, c'est-à-dire d'un ou plusieurs adjectifs assortis ou non d'adverbes. Par exemple, si un attribut peut prendre les modalités "chaise", "table", "placard", "table basse", "chaise haute", considérer "chaise" comme plus proche de "chaise haute" que des autres modalités et "table" plus proche de "table haute" que des différentes chaises et de l'armoire peut permettre une classification plus fine, et plus représentative de la réalité concrète. Nous nommerons les mots constitutifs d'une modalité "unités syntaxiques minimales".

La formule permettant ce traitement est :

$$\mu_i(j,k) = \frac{n^2_{u(i,j) \bigcap u_{Q_j^i}(i,k)}}{N_u(i,j) * N_{u_{Q_j^i}(i,k)}} * \Theta^i_{kQ_j^i}$$

avec

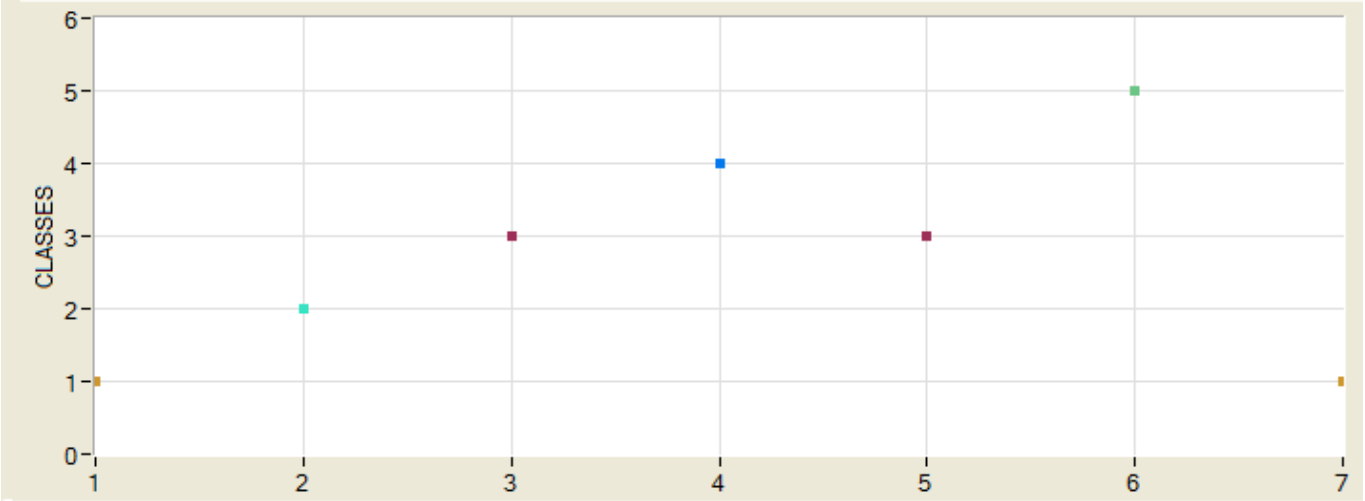
- $\mu_i(j,k)$: MAD pour la caractéristique i de l'individu j à la classe k,
- $n_{u(i,j) \bigcap u_m(i,k)}$: nombre d'unités syntaxiques minimales similaires entre la modalité de la caractéristique i de l'individu j et la modalité m de la caractéristique i de la classe k.
- $N_u(i,j)$: nombre d'unités syntaxiques minimales constituant la modalité de la caractéristique i de l'individu j,
- $N_{u_m(i,k)}$: nombre d'unités syntaxiques minimales constituant la modalité m de la caractéristique i de la classe k.
- Θ^i_{km} : fréquence de la modalité Q^i_j pour la classe k.

Pour illustrer le fonctionnement de traitement, un échantillon d'individus à classer est décrit dans le tableau 3, et les résultats de la partition correspondante sont représentés sur les figure 27 et 28 ; cet exemple montre comment il est possible, suivant cette méthode, d'opérer un rapprochement entre les animaux à poils, par opposition aux animaux à plumes ou à écailles. Ici, les unités syntaxiques minimales de cet exemple sont : "poils", "plumes", "écailles", "longs", "courts". Les noyaux des syntagmes nominaux sont "poils", "plumes", et "écailles" ; "poils" et "longs" ne sont que des adjectifs pour les qualifier. Les modalités "poils" et "poils longs" ou "poils" et "poils courts" ne sont pas considérées comme identiques, mais comme plus ressemblants que "poils longs" et "poils courts", qui eux-mêmes sont plus proches que "poils" et "écailles" ou que "poils" et "plumes".

Les figures 27 et 28 décrivent deux classifications obtenues, la première en considérant deux modalités comme totalement distinctes si elles ne sont pas strictement similaires, et la deuxième prenant en compte le principe de proximité entre modalités. α est paramétré tel que l'indice de la qualité de la partition de la classe soit maximisé, c'est-à-dire respectivement 1 et 0.6. Notons que dans le cas où α aurait été paramétré à un degré équivalent ou supérieur à 0.8, lors de la classification tenant compte de la proximité entre les modalités, le guépard aurait formé une classe à lui seul, du fait de la différence de son poids moyen par rapport aux autres mammifères à classer. Les indices de qualité de la partition sont respectivement 0.33 et 0.62. La matrice de coalescence du descripteur "revetement_externe" est indiquée dans le tableau 4.

	Poids moyen	Revetement externe	Mammifère
Guépard	40	poils-courts	Oui
Chat	3.5	poils	Oui
Faucon	0.9	plumes	Non
Thon rouge	75	ecailles	Non
Canard	1.2	plumes	Non
Alpaga	6	poils-longs	Oui
Renard	5	poils-courts	Oui

Tableau 3. Exemple d'un échantillon d'individus à classer pour la comparaison avec et sans la prise en compte de la proximité des modalités

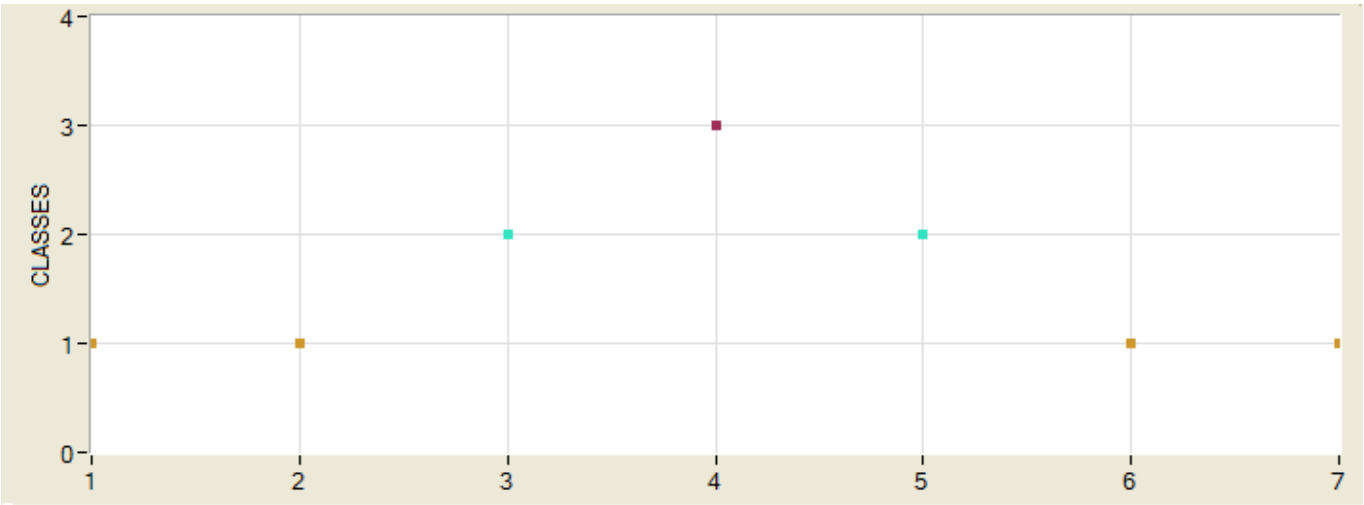


a. Répartition des individus dans les classes

	Poids_moyen	Revetement_externe	Mammifère
Classe 1	22.5	poils-courts: 2	Oui: 2
Classe 2	3.5	poils: 1	Oui: 1
Classe 3	1.05	plumes: 2	Non: 2
Classe 4	75	ecailles: 1	Non: 1
Classe 5	6	poils-longs: 1	Oui: 1

b. Description des classes

Figure 27. Classification sans prise en compte de la proximité des modalités



a. Répartition des individus dans les classes

	Poids_moyen	Revetement_externe	Mammifère
Classe 1	13.6	poils-courts: 2 / poils: 1 / poils-longs: 1	Oui: 4
Classe 2	1.05	plumes: 2	Non: 2
Classe 3	75	ecailles: 1	Non: 1

b. Description des classes

Figure 28. Classification avec prise en compte de la proximité des modalités

	Poils	Poils courts	Poils longs	Plumes	Ecailles
Poils	1	0.5	0.5	0	0
Poils courts		1	0.25	0	0
Poils longs			1	0	0
Plumes				1	0
Ecailles					1

Tableau 4. Matrice de coalescence de la caractéristique "revetement_externe"

La méthode LAMDA permet ainsi la détection automatique des unités syntaxiques minimales et les rapprochements de certaines modalités. La prise en compte de la similarité de certains mots constitutifs de la modalité peut n'être cependant pas souhaitable, soit de manière totale, soit partiellement ; il reste important de laisser à l'utilisateur la possibilité de désactiver cette option, car le choix de prendre cette proximité en compte ou non dépendra des contraintes et des objectifs de la partition à réaliser. Par exemple, dans le cas présenté dans le tableau 5, l'utilisateur peut ne vouloir voir rapprochés que "triangle" et "triangle-rectangle" et non "rectangle" et "triangle-rectangle" - sans pour autant que "triangle" et "triangle-rectangle" soient considérés comme strictement similaires puisque, le cas échéant, il aurait pu simplement décrire cette caractéristique en terme de nombre d'angles. Pour pallier cette difficulté, il reste nécessaire de laisser à l'utilisateur la possibilité de désactiver cette option, ou d'indiquer manuellement les termes qu'il veut sélectionner comme unités syntaxiques minimales. Dans l'exemple cité, il pourra ainsi ne sélectionner que "triangle" et la matrice de coalescence pour la caractéristique "forme" correspondra à celle représentée dans le tableau 6.

Forme	Couleur
triangle	Rouge
triangle-isocele	Bleu
cercle	Rouge
triangle-rectangle	Rouge
rectangle	Bleu
rectangle	Bleu
triangle-isocele	Rouge

Tableau 5. Exemple de classification requérant une saisie manuelle des unités syntaxiques minimales

	triangle	triangle-isocele	triangle-rectangle	rectangle	cercle
triangle	1	0.5	0.5	0	0
triangle-isocele		1	0.25	0	0
triangle-rectangle			1	0	0
rectangle				1	0
cercle					1

Tableau 6. Matrice de coalescence de la caractéristique "Formes"

Les résultats de la classification des individus décrits par les caractéristiques "Formes" et "Couleurs" sont indiqués dans la figure A.2 de l'annexe III.1 pour la classification ne prenant pas en compte la proximité des modalités et la figure A.3 pour la classification la prenant en compte. Les indices de partition sont respectivement 0.29 et 0.65. On constate que le fait de considérer la proximité entre les différents types de triangles permet d'augmenter la capacité du système à les reconnaître. Un triangle rectangle est un triangle particulier, et cette particularité doit être soulignée - d'où l'importance de les considérer comme distincts - mais leurs caractéristiques de base restent les mêmes, et cette information peut être primordiale dans certaines classifications. Dans cet esprit, le degré de particularité d'une modalité est prise en compte en fonction du nombre d'unités syntaxiques la composant. Par exemple, si nous avions eu un triangle isocèle rectangle, le degré de similarité avec le triangle aurait été de 0.33 et de 0.66 avec les triangles particuliers lui ressemblant davantage : triangle rectangle ou avec le triangle isocèle.

Cette fonctionnalité permet aux variables de type quantitatif d'admettre une extension : la liste. En effet, jusqu'à présent, il n'était pas possible de traiter des listes de taille indéfinie avec LAMDA car le nombre de descripteurs est fixe et strictement égal entre les différents individus d'un même échantillon. Avec la prise en compte de la proximité des mots constituant une modalité, une variable qualitative peut également être constituée d'une liste de variables qualitatives et de taille dynamique. Nous en voyons un exemple dans le tableau 7, décrivant une liste de sportifs à trier, associée à la matrice de coalescence de la variable "sports" du tableau 8.

Poids	Sports	Sexe
80	Rugby-Tennis	Homme
72	Marathon-PingPong	Homme
65	Tennis	Femme
88	Marathon-PingPong-Rugby	Homme
79	Rugby-Tennis	Homme
91	Marathon-Tennis	Homme
69	Tennis	Femme

Tableau 7. Exemple de classification dont les modalités d'une variable qualitative intègrent un type liste

	Rugby-Tennis	Marathon-PingPong	Marathon-PingPong-Rugby	Marathon-Tennis	Tennis
Rugby-Tennis	1	0	0.15	0.25	0.5
Marathon-PingPong		1	0.66	0.25	0
Marathon-PingPong-Rugby			1	0.15	0
Marathon-Tennis				1	0.5
Tennis					1

Tableau 8. Matrice de coalescence de la caractéristique "Sports"

La variable "sports" admet comme modalités des listes de sports. Les unités syntaxiques sont : "Marathon", "PingPong", "Rugby", "Tennis".
Les résultats de la classification des individus décrits par les caractéristiques "Poids", "Sports" et "Sexe" sont indiqués dans la figure A.4 de l'annexe III.2 pour la classification ne prenant pas en compte la proximité des modalités et la figure A.5 pour la classification la prenant en compte. Les indices de partition sont respectivement 0.31 et 0.62.

4.3 Traitement des données manquantes

Si un apprentissage à partir de données manquantes est le plus souvent problématique - car apprendre à partir d'informations incomplètes peut aboutir à une partition lacunaire ou faussée - il est intéressant d'être capable de gérer une classification supervisée d'individus dont toutes les caractéristiques ne sont pas nécessairement renseignées. Cette classification ne peut jamais garantir d'être parfaite, mais si les informations manquantes ne sont pas décisives, elle peut proposer des affectations réalistes. La justesse de la classification dépend essentiellement de l'importance des données à classer, ce qui peut être connu précisément par l'intermédiaire de MEMBAS.

Ainsi, comme le montre la figure 29, pour classer un individu dont certaines caractéristiques ne sont pas renseignées, nous permettons à LAMDA de calculer les vecteurs de MAD de toutes les autres caractéristiques aux classes existantes, et en suivant le vecteur GAD de l'individu à ces classes, de manière à le classer dans celle pour laquelle le GAD est maximisé - tout se passe en fin de compte comme lorsque toutes les caractéristiques de l'individu sont connues, seules les dimensions du vecteur MADs de l'individu changent : soient I le nombre de caractéristiques d'un individu j complètement connu et d'un autre individu j' partiellement connu, I' le nombre de caractéristiques de j' n'étant pas renseignées, et K le nombre de classes dans lesquelles j et j' devront être classés. Nous avons donc :

$$Dim(\overrightarrow{MAD_j}) = IxK$$

et

$$Dim(\overrightarrow{MAD'_j}) = (I - I')xK$$

Le calcul du vecteur $\overrightarrow{GAD'_j}$ peut donc être opéré directement à partir de $\overrightarrow{MAD'_j}$.

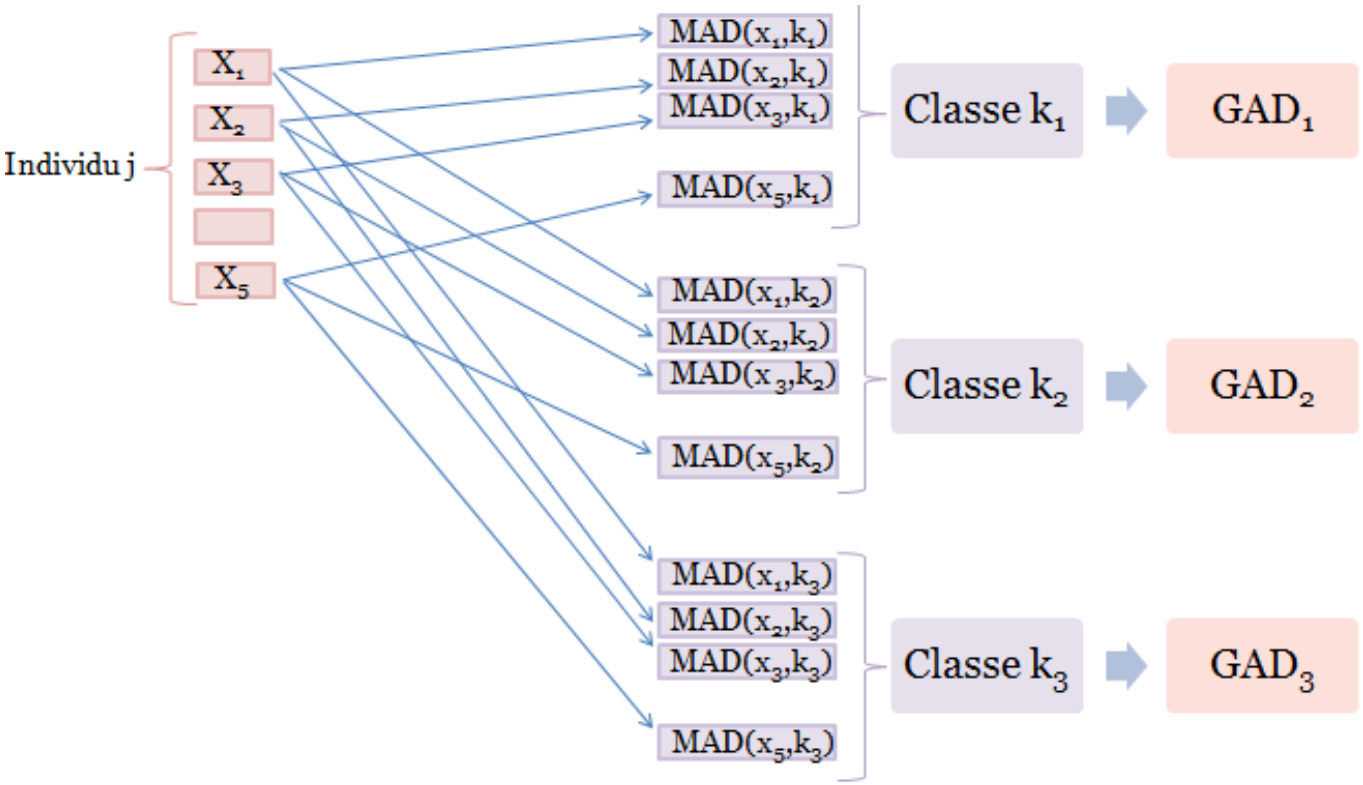


Figure 29. Calcul des GADs lors de la classification impliquant 3 classes et des individus normalement décrits par 5 caractéristiques dans le cas où le quatrième descripteur de l'individu j n'est pas renseigné

La probabilité pour que l'individu j soit classé dans k_j, qui est la classe de laquelle il est effectivement le plus proche en réalité, est fonction du poids de ses I descripteurs connus et est définie par la fonction :

$$P(j \in k_j) = \sum_{i}^{I^-} \bar{w}_{i_c}$$

avec

- I^- : I sans I',
- $\bar{w}_{i_c} \in [0,1]$: Poids normalisé du poids w_{i_c} dans la classification c de la ie caractéristique de j, obtenu par MEMBAS.

Par exemple, nous présentons dans le tableau 9 un exemple de classification de livres, réalisées d'après les caractéristiques "Genre", "Annee", "Roman". Les poids normalisés de ces caractéristiques pour cette partition sont respectivement 0.57, 0.06, et 0.37. Ainsi, les probabilités pour qu'un individu soit bien classé lorsqu'une caractéristique fait défaut sont : 0.43 lorsque c'est le genre qui n'est pas connu, 0.94 pour l'année, et 0.63 pour le fait qu'il s'agisse ou non d'un roman. On estime qu'une classification dont la probabilité d'exactitude se situe en-dessous de 0.5 ne peut plus délivrer d'information fiable : lorsqu'une caractéristique ou un groupement de caractéristiques dont le poids dépasse 0.5 vient à manquer, le résultat de la classification n'a plus vraiment de sens et est davantage tributaire du jeu du hasard que d'une réelle représentativité.

	Genre	Annee	Roman
Classe 1	Aventure	1844	oui
Classe 2	Historique	1831	oui
Classe 3	Drame	1856	oui
Classe 3	Drame	1782	oui
Classe 1	Aventure	1869	oui
Classe 4	Drame	1849	non
Classe 3	Drame	1846	oui
Classe 4	Drame	1944	non
Classe 3	Drame	1947	oui
Classe 3	Drame	1939	oui
Classe 2	Historique	1959	oui
Classe 3	Drame	1883	oui
Classe 3	Drame	1885	oui
Classe 5	Historique	1827	non
Classe 4	Drame	1838	non
Classe 2	Historique	1678	oui
Classe 1	Aventure	1771	oui
Classe 4	Drame	1637	non
Classe 4	Drame	1677	non
Classe 1	Aventure	1838	oui
Classe 6	Aventure	1456	non
Classe 7	Drame	1100	oui
Classe 7	Drame	1176	oui
Classe 8	Comédie	2013	non
Classe 8	Comédie	2014	non
Classe 8	Comédie	2010	oui
Classe 8	Historique	2010	oui

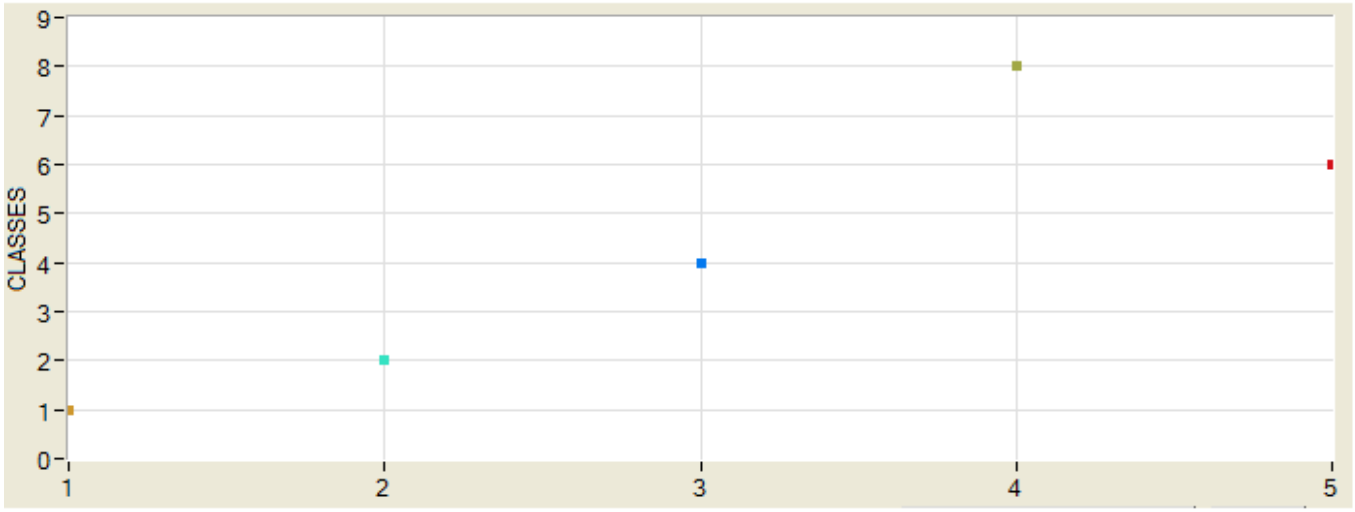
Tableau 9. Exemple d'une partition initiale

Nous pouvons observer dans la figure 31 les individus ayant été affectés à la bonne classe malgré leurs données manquantes, ainsi que les erreurs de classement comparativement à la figure 30.

Pour l'individu test 1, typiquement, la caractéristique "Genre" étant une "aventure", deux classes pourraient lui correspondre : la 1 et la 6. La caractéristique "Roman" étant "oui", ainsi que l'année "1800" penchent ensuite en faveur de la 1. Si la caractéristique "Roman" ou "Année" venaient à manquer, le résultat serait conséquemment inchangé. Par contre, si aucune des deux n'étaient renseignées, la méthode ne pourrait pas décider dans quelle classe affecter l'individu test 1, car rien ne pourrait l'amener à préférer la 1 à la 6. Par contre, si la caractéristique "Genre" faisait défaut, la caractéristique "Roman" permettrait à la méthode LAMDA de le rapprocher des classes 1, 2, 3, 7, et potentiellement 8. D'après la caractéristique "Date", la classe 2 serait élue, car elle affiche une moyenne de 1822 contre respectivement 1830, 1876, 1138, et 2012 pour les classes 1, 3, 7, et 8. Cette observation, au même titre que celles pouvant être établies à partir des exemples fournis par les autres individus tests du même échantillon, explique l'importance que présente la caractéristique "Genre" dans la classification et, de là, son poids relativement élevé : c'est elle qui permet la meilleure discrimination. Nous pouvons dire, en quelque sorte, que c'est elle qui fournit l'information décisive à la bonne classification d'un individu.

	Genre	Annee	Roman
Test 1	Aventure	1800	oui
Test 2	Historique	1300	oui
Test 3	Drame	1928	non
Test 4	Comédie	2000	non
Test 5	Aventure	2000	non

a. Echantillon à classer

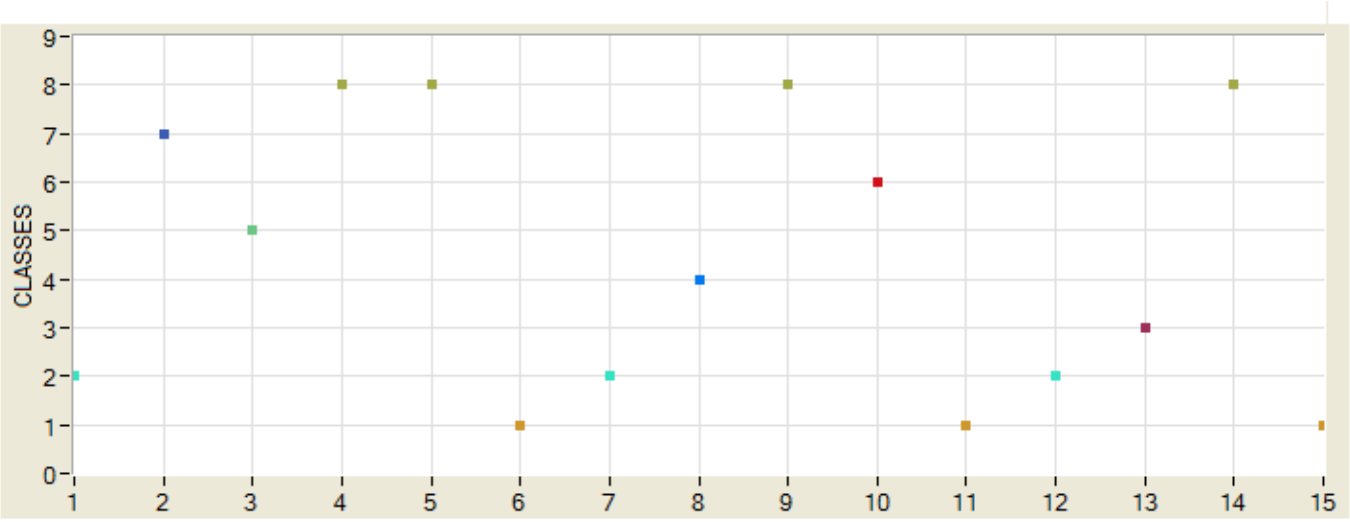


b. Répartition des individus dans les classes

Figure 30. Classification sans données manquantes

	Genre	Annee	Roman
Test 1		1800	oui
Test 2		1300	oui
Test 3		1928	non
Test 4		2000	non
Test 5		2000	non
Test 6	Aventure		oui
Test 7	Historique		oui
Test 8	Drame		non
Test 9	Comédie		non
Test 10	Aventure		non
Test 11	Aventure	1800	
Test 12	Historique	1300	
Test 13	Drame	1928	
Test 14	Comédie	2000	
Test 15	Aventure	2000	

a. Echantillon à classer



b. Répartition des individus dans les classes

Figure 31. Classification avec données manquantes

On observe ainsi que, dans cet exemple, tous les individus pour lesquels l'année est inconnue ont été correctement classés, car cette caractéristique n'entre que très peu en jeu dans la classification. Concernant les individus dont le champ "Roman" n'est pas renseigné, le taux de réussite est de 60%, et chute à 20% pour ceux dont le genre n'est pas indiqué. De fait, on constate dans le tableau 10, que peu s'en est fallu que l'individu test 1 soit affecté à la classe 1 - et soit donc bien classé : seul a joué le fait que la classe 2 affiche une date moyenne de 1822 contre 1830 pour la classe 1. Ainsi, lorsqu'une information aussi importante que la caractéristique "Genre" ici vient à manquer, le résultat de la classification doit beaucoup sa justesse à l'aléatoire.

	Genre	Annee	Roman
Classe 1	Aventure : 4	1830	oui: 4
Classe 2	Historique : 3	1822	oui: 3
Classe 3	Drame: 7	1877	oui: 7
Classe 4	Drame: 5	1789	non: 5
Classe 5	Historique: 1	1827	non: 1
Classe 6	Aventure: 1	1456	non: 1
Classe 7	Drame: 2	1138	oui: 2
Classe 8	Comédie:3 /Historique: 1	2012	oui: 2 / non: 2

Tableau 10. Description des classes

En outre, il est important de garder à l'esprit que, concernant ces probabilités d'échec ou de réussite, il ne peut s'agir que d'approximations, car la réussite de la classification est dépendante de la représentativité des données d'apprentissage et de la normalité des individus à classer. Evidemment, un individu pour lequel tous les champs des attributs ne sont pas renseignés ou se montrent sensiblement différents de ceux ayant permis l'apprentissage sera difficilement classable.

4.4 Conclusion

Pour que la classification soit réaliste, il faut lui fournir les types adéquats, c'est-à-dire les plus adaptés possible à la description d'objets réels. Ainsi, si SMSP apporte une première solution satisfaisante en permettant la gestion simultanée de données qualitatives, quantitatives, et intervallaires, l'algorithme nous est apparu comme étant encore insuffisant pour représenter certaines informations observables sans perte d'information.

C'est pourquoi nous avons ajouté aux trois types de base le type composite, dont le traitement est rendu possible par la classification multicouche, afin d'y intégrer des caractéristiques multi-dimensionnelles, permettant de gérer simultanément les trois types de base dans une même entité signifiante qui peut elle-même être constituée d'autres attributs composites.

Nous avons en outre ajouté le traitement du squelette des modalités des données qualitatives afin de rompre avec le calcul binaire du degré d'appartenance de ce type d'informations, et d'introduire l'usage de listes d'unités syntaxiques. Les variables qualitatives ne se limitent donc plus nécessairement à n'être qu'un mot, mais peuvent être un syntagme -grammatical ou sémantique- consistant, ou une liste de mots.

Enfin, nous avons permis à LAMDA de gérer des données lacunaires, de manière à fournir une partition qui - s'il n'est pas garanti qu'elle soit totalement exempte d'erreurs - pourra constituer une première approche du problème de classification auquel l'utilisateur est soumis tout en l'informant de la probabilité relative à la justesse des résultats obtenus. De cette manière, le choix des descripteurs pour un apprentissage en vue d'une reconnaissance ultérieure est moins contraint par la nécessité de ne sélectionner que des données systématiquement accessibles. En effet, si certaines données viennent à manquer au moment de la reconnaissance, une classification peut tout de même être opérée, informant l'utilisateur du degré de probabilité de son efficience. Néanmoins, dans le cas où il s'agit d'informations réellement décisives pour la classification, nous considérons que le résultat de cette dernière n'est pas significatif.

Dans le prochain chapitre, nous montrerons comment nous avons appliqué LAMDA et les améliorations que nous lui avons apportées au problème de classification auquel le projet MAISEO nous avait soumis.

5 Application au projet MAISEO

Pour le travail de classification nécessaire au bon déroulement du projet MAISEO, nous avons sélectionné LAMDA - méthode qui paraissait particulièrement adaptée aux constraintes auxquelles nous savions devoir faire face :

- Assurer la gestion simultanée de données de types hétérogènes - qualitatifs et quantitatifs au minimal,
- Etablir une classification d'un ensemble d'individus n'en comptant pas deux strictement similaires,
- Faire face à l'inégale disponibilité des informations relatives aux individus.

En effet, il est difficile de prévoir le comportement exact d'une parcelle au cours de l'année à venir. Et pour cause, malgré le fait que certaines parcelles témoignent de caractéristiques communes -ce qui rend possible la définition de profils - chacune d'elles a ses particularités et il n'existe pas deux exploitations similaires. Les données météorologiques affichent en outre une relative incertitude, à plus forte raison lorsque la période anticipée concerne un futur éloigné de plusieurs mois. Le fait que LAMDA base sa classification sur la logique floue constitue donc un point fort, puisqu'il tolère une marge d'incertitude.

Concernant le risque inhérent à l'incertitude de disposer systématiquement de toutes les caractéristiques des cultures, il s'agissait de définir des profils génériques de parcelles compte tenu des informations dont nous pouvions disposer et sur lesquelles nous pensions pouvoir régulièrement compter, sans en avoir une garantie systématique. De ce fait, notre classifieur devait n'être pas trop rigide et - s'il n'avait pas été conçu en vue d'être capable de traiter les données manquantes - pouvoir s'y adapter. En ajoutant cette capacité à LAMDA, nous avons pu pallier cette difficulté.

La méthode LAMDA est en outre capable, nous l'avons vu, de traiter simultanément les données qualitatives, quantitatives, et intervallaires, ce qui la rend particulièrement adaptée aux spécificités des classifications à opérer.

Dans ce chapitre, nous allons montrer comment LAMDA nous a permis de réaliser les deux parties de notre travail de classification, en commençant par la partie relative à la classification des parcelles pour le conseil aux agriculteurs et poursuivant en détaillant la phase dévolue à la classification des îlots d'exploitations pour l'aide à la décision du gestionnaire du bassin versant. Dans les deux cas, nous développerons d'abord les choix que nous avons faits en matière de représentation de données, puis expliquerons comment nous avons fait face aux différentes difficultés qui se sont présentées à nous, avant d'en exposer les résultats.

5.1 Classification des parcelles

La première étape dans notre travail de classification des parcelles a été de sélectionner les caractéristiques à prendre en compte pour obtenir une partition représentative des profils en accord avec la double question qui nous occupait : quel type de maïs semé permettrait d'optimiser de concert le rendement et la quantité d'eau d'irrigation à apporter ? Quand le semer ?

Le deuxième problème auquel il a fallu répondre concerne le traitement particulier qu'exigeait cette situation. En effet, il s'agissait d'une classification contrainte, du fait que certaines valeurs - en l'occurrence celles relatives à la date de semis et la précocité du grain - ne devaient pas être mêlées à d'autres. En d'autres termes, à un couple "date de semis"/"précocité" devait correspondre au minimum un profil de classe. Dans le cas d'une classification non supervisée non contrainte, les classes auraient été définies en fonction de la proximité des différentes valeurs de l'ensemble des attributs et chaque profil serait décrit par un ensemble de moyennes. Cette contrainte a occasionné des questionnements supplémentaires, du fait de la difficulté à recueillir des informations complètes pour étayer notre travail de définition de classes. En effet, pour s'assurer de la qualité et de la représentativité de la partition, il était primordial que chaque classe concerne un nombre suffisant de parcelles ; mais nos partenaires ont rencontré divers obstacles au cours de la collecte de données. Nous avons finalement pu constituer un échantillon d'apprentissage constitué de 150 parcelles.

Nous avons détaillé section 1.2.2.2 les attributs que nous avons retenus. Nous présentons ci-dessous comment nous les avons représentés, puis en quoi la capacité à traiter des données lacunaires a permis d'améliorer les performances de la classification et d'éprouver notre système.

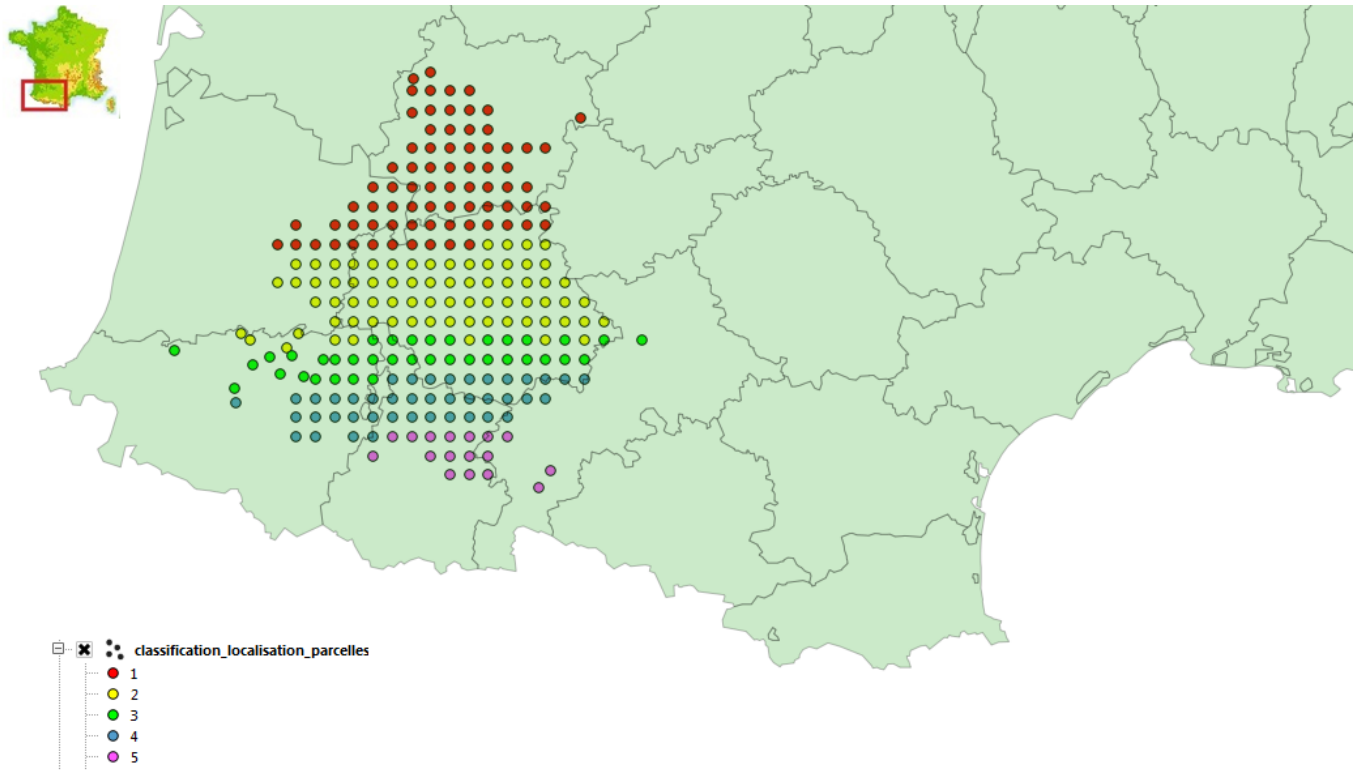
5.1.1 Représentation des données parcellaires

5.1.1.1 Données météorologiques

Les données météorologiques ne concernent qu'un aspect des facteurs entrant en considération dans les besoins en eau et le rendement d'une culture de maïs, il s'agit d'un attribut unique, au même titre que le type de sol, la réserve utile, ou le fait que la parcelle soit irriguée ou non. En outre, les différentes caractéristiques des conditions météorologiques sont clairement interdépendantes : la température, les précipitations, le rayonnement, l'évaporation sont des expressions phénoménales d'un contexte atmosphérique et sont intrinsèquement liées entre elles. Dans cet optique, leur représentation par un type composite ne faisait aucun doute. Chacun des attributs de la caractéristique "Météo" correspond à un type simple, en l'occurrence le type quantitatif puisqu'il s'agissait d'une valeur chiffrée précise.

Le diagramme de classes du classifieur que nous avons implémenté est présenté dans l'annexe IV.1. Nous avons utilisé le même outil pour la section suivante - relative à la classification des îlots d'irrigation - la principale différence entre les deux traitements résidant dans le choix du mode de classification employé : contraint ou non.

Pour valider notre modèle, nous avons tout d'abord réalisé une classification multicouche basée sur notre propre définition des profils météorologiques. Pour cela, nous avons défini chaque individu selon ses caractéristiques propres, tant au niveau pédoloqique (type de sol, réserve utile) que cultural (date de semis, précocité du grain choisi, irrigation), du rendement, et météorologique. Pour ce faire, nous avons attribué à chaque individu les caractéristiques météorologiques de la maille à laquelle il appartient. Les caractéristiques météorologiques employées ici sont au nombre de 11, car nous avons réalisé la classification à partir de notre calcul du Q50 des valeurs extraites à partir du relevé journalier des dix dernières années, pour les informations présentées dans le tableau A.2 de l'annexe I.2, qui nous avaient été transmises par Météo France. Le résultat de la classification météorologique qui en découle est présenté à la figure 32.



a. Représentation des différentes parcelles en fonction de la partition météorologique

	Disponibilités en eau					Chaleur	Froid	Besoins en eau			
Classe	Hiver	Semis	Flo1	Flo2	Flo3	Remp	CC	Min	Flo1	Flo2	Flo3
1	220.3	169.7	19.24	15	14.15	121.5	2404	6.4	75.78	79.78	73.6
2	240.3	198.2	22.45	18.88	14.89	149. 5	2398	6	72.82	77.63	71.31
3	275.2	204.9	22.07	19.77	18.63	158.3	2325	5.9	70.77	74.78	68.61
4	306.2	245.1	26.14	21.01	23.92	174.4	2189	5.4	63.86	67.26	62.54
5	308.2	249.1	27.02	25.76	27.89	181.5	2014	5.2	66.23	62.53	59.29

b. Définition des classes météorologiques

Figure 32. Partition de la zone test en fonction des données météoroloqiques.

Empiriquement, nous pouvons observer sa cohérence car la zone test apparaît bien comme découpée en différentes sous-zones consistantes. Nous avons ensuite pu comparer, aux côtés des experts météo, nos classes à celles qu'ils nous ont ultérieurement présentées pour valider notre modèle. Sa représentativité a été confirmée, néanmoins, par la suite, nous avons remplacé nos classes météorologiques par celles définies par Météo France à partir de leur modèle, car elles bénéficiaient d'un plus haut niveau de détail. Les cartes que nous ont fourni Météo France affichaient 7 classes distinctes contre 5 pour celles que nous avons produites.

Initialement, notre classification ne prenait en compte que 50 parcelles car nous ne disposions alors pas d'individus supplémentaires, ni la répartition des parcelles en fonction de la précocité du grain et de la date de semis ; il s'agissait d'une classification non supervisée non contrainte, dont les résultats sont présentés dans le tableau A.7 de l'annexe IV.2. Nonobstant le nombre restreint d'individus - alors insuffisant pour qu'une classification non supervisée contrainte puisse fournir une partition pertinente - cette première classification nous a permis de valider l'efficacité de notre système et la cohérence du résultat qu'il était en mesure de nous fournir. En effet, parmi les 50 individus alors disponibles présen-

taient 4 degrés de précocité différents ("semi-précoce", "semi-tardif", "tardif", "très tardif") et 5 plages de dates de semis, et totalisaient 14 couples différents, imposant du même coup un nombre égal de classes au minimum en cas de classification non supervisée contrainte.

5.1.1.2 Types de sol

L'attribut "Type de Sol" concerne la composition physique des sols, en terme de structure minérale. De fait, la structure du sol dépend de l'assemblage des particules minérales et organiques en agrégats qui, eux-mêmes, forment des mottes. Ce sont les microorganismes, la petite faune, les filaments de champignons et les racines, qui en déplaçant et organisant les particules minérales et organiques échafaudent la structure du sol. La matière organique sert de liant entre les particules minérales (limon, sable...). Chaque sol possède des proportions de particules minérales qui lui sont propres et permettent de les regrouper en familles : c'est la texture d'un sol. Cette dernière donne des informations sur les caractéristiques importantes d'un sol comme sa richesse chimique, sa résistance physique, etc. Il s'agit donc, pour cet attribut, d'énoncer les particules minérales présentes en grande proportion dans le sol, car ce sont d'elles dont dépendront principalement les caractéristiques pédologiques, en terme de capacité à retenir l'eau, à absorber la chaleur... Pour s'en convaincre, il suffit d'observer le tableau A.8 de l'annexe IV.2 qui dresse un portait des types de sol les plus fréquents.

La connaissance de la composition précise des sols, en terme de proportions, ainsi que celle de la composition chimique nécessitent une analyse technique que peu d'agriculteurs sont enclins à solliciter, et ne présentent pas de réel intérêt ajouté ; déterminer quels sont les minéraux les plus présents dans le sol d'une parcelle constitue l'étude la plus intéressante : elle ne requiert pas de réelle expertise et l'agriculteur peut le faire lui-même. C'est en outre la seule information relative à la composition pédologique réellement décisive quant au diagnostic des besoins en eau et du rendement de la parcelle.

Pour la représenter, puisqu'il s'agissait alors de citer les principaux composants d'un sol, nous avons eu recours au type liste, tel que nous l'avons présenté dans la section IV.2. Cette décision a été arrêtée du fait que le type élu devait témoigner des caractéristiques suivantes :

- Enumérer des termes d'un même type, à savoir la dénomination du composant minéral énoncé,
- Admettre un nombre indéfini de termes différents,
- Permettre une comparaison basée tant sur le nombre de termes présents que sur leur proximité.

Pour cela, les mots -c'est-à-dire les unités syntaxiques minimales - possibles représentent les différents composants minéraux pouvant être présents dans le sol, c'est-à dire : "limon", "calcaire", "argile", "sable". Il est néanmoins important de noter que, même si l'échantillon d'apprentissage a été sélectionné dans l'optique d'être le plus complet possible, c'est-à-dire de couvrir le maximum de cas pouvant être rencontrés, d'autres unités syntaxiques minimales pourront être ajoutées ultérieurement, ce qui nécessitera, le cas échéant, une phase d'apprentissage préalable pour que ces parcelles puissent être reconnues et diagnostiquées efficacement. Cet apprentissage ayant été opéré sur une zone géographique restreinte, sa représentativité à l'échelle du territoire est bien entendue limitée.

Sans l'emploi de la liste, deux individus dont la composition pédologique est proche seraient considérés comme aussi éloignés que deux individus dont les sols ne seraient constitués d'aucun minéral principal en commun. Par exemple, la proximité entre les valeurs "limon" et "limon-argile" serait égale à celle entre "limon" et "calcaire", c'est-à-dire à zéro. Les tableaux 11 et 12 exposent les matrices de coalescence à l'origine de la différence de classification observée lors d'un apprentissage avec utilisation de variables qualitatives simples et avec utilisation de listes.

	limon	limon-argile	sable-argile	calcaire	argile-calcaire	argile	limon-sable	limon-sable-argile
limon	1	0	0	0	0	0	0	0
limon-argile		1	0	0	0	0	0	0
sable-argile			1	0	0	0	0	0
calcaire				1	0	0	0	0
argile-calcaire					1	0	0	0
argile						1	0	0
limon-sable							1	0
limon-sable-argile								1

Tableau 11. Matrice de coalescence entre les différentes modalités pour la variable "Type de sol" lorsqu'elle est de type quantitatif simple.

	limon	limon-argile	sable-argile	calcaire	argile-calcaire	argile	limon-sable	limon-sable-argile
limon	1	0.5	0	0	0	0	0.5	0.33
limon-argile		1	0.25	0	0.25	0.5	0.25	0.33
sable-argile			1	0	0.25	0	0.25	0.66
calcaire				1	0.5	0	0	0
argile-calcaire					1	0.5	0	0.15
argile						1	0	0.15
limon-sable							1	0.66
limon-sable-argile								1

Tableau 12. Matrice de coalescence entre les différentes modalités pour la variable "Type de sol" lorsqu'elle est de type liste.

5.1.1.3 Autres caractéristiques

Pour la représentation des autres caractéristiques, à savoir le rendement espéré, la RU, la possibilité d'irriguer, le degré de précocité du grain employé, et la date de semis, nous avons employé les types de base. En effet, le rendement espéré et la RU, exprimés respectivement en q/ha et en cm, sont décrits chacun par une valeur numérique et leur conversion en variables de type quantitatif se fait tout naturellement.

Concernant la possibilité d'irriguer, décrite par les deux seules modalités "oui" et "non, elle peut être représentée sans perte d'information par un type qualitatif simple, puisque l'évaluation de leur similarité doit être de type booléen : soit le mot est identique, soit il ne l'est pas.

Enfin, pour la représentation du degré de précocité du grain et la date de semis, nous avons eu recours au type quantitatif à nouveau. Pour le degré de précocité, il s'agissait de représenter six modalités s'étalant de "très précoce" à "très tardif", aussi le type quantitatif nous est-il apparu comme le plus adapté à l'expression de cette gradation. Quant à la date de semis, il nous fallait représenter des plages temporelles qui se succédaient. Ces plages étant de dimension égale (15 jours), l'emploi d'intervalles ne nous a pas semblé utile ; seule la notion d'échelonnage nécessitait d'être conservée. Les tableaux 13 et 14 retranscrivent respectivement les différentes valeurs initiales des degrés de précocité et des dates de semis et leur conversion en type quantitatif.

Valeur initiale	très tardif	tardif	semi-tardif	semi-précoce	précoce	très précoce
Valeur convertie	1	2	3	4	5	6

Tableau 13. Correspondance entre la variable "Précocité" et sa représentation quantitative.

Valeur initiale	15/03 - 31/03	01/04 - 14/04	15/04 - 30/04	01/05 - 14/05	15/05 - 31/05	01/06 - 14/06
Valeur convertie	1	2	3	4	5	6

Tableau 14. Correspondance entre la variable "Date de semis" et sa représentation quantitative.

Notons que, si pour l'apprentissage, la variable rendement n'est pas prise en compte et est représentée par un type quantitatif, elle est ensuite stockée sous forme intervallaire afin d'être présentée en sortie du diagnostic parcellaire afin d'aiguiller au mieux les agriculteurs ; le minimum correspond au plus petit rendement des parcelles affectées à la classe, et la maximum au plus élevé. De cette manière, lorsque les scenarii seont présentés à l'utilisateur, les profils associés à sa classe présenteront un intervalle de rendement espéré et non une valeur précise, qui serait bien trop difficile à évaluer.

5.1.2 Utilisation de données lacunaires

Pour la partie apprentissage, notre système ne tolère pas de données lacunaires : l'absence de certaines informations empêchent catégoriquement toute tentative de partition significative. L'aspect reconnaissance, par contre, peut être envisagé malgré le défaut de certaines données.

Cependant, comme nous l'avons vu dans la section IV.3, toutes les données ne sont pas égales en importance dans l'affectation d'une parcelle à une classe. Typiquement, si l'attribut "type de sol" n'est pas renseigné, le résultat de la classification s'en trouve sensiblement moins fiable, car il s'agit d'une information primordiale pour le diagnostic de la parcelle. De fait, cet attribut est celui dont le poids dans la classification est le plus élevé, comme le montre le tableau 15 ; aussi un individu classé sans cette information témoigne-t-il d'une probabilité de seulement 0.58 d'être bien classé. Or l'information concernant le type de sol, telle que la demande le système, est aisément accessible - au moins visuellement à l'agriculteur, l'aspect et la texture étant assez éloquents. Si, malgré tout, ils ne parviennent pas à estimer eux-mêmes la composition physique de leur parcelle, ils peuvent recourir à une analyse de leur sol par un technicien et, dans le pire des cas, c'est-à-dire si l'information vient à manquer au moment de la reconnaissance, ils seront avertis de la probabilité d'occurrence des scénarii que leur seront, en suivant, soumis.

	Poids normalisé
TypeSol	0.42
RU	0.14
Irrigation	0.2
Meteo	0.24

Tableau 15. Poids normalisé des caractéristiques dans la classification

A l'inverse, si l'information relative à la taille de la RU fait défaut, cette parcelle a une probabilité de 0.9 d'être bien classée ; ce qui est intéressant dans le cadre de son utilisation par les agriculteurs puisqu'il s'agit de l'information la plus difficile a obtenir pour eux puisqu'elle se calcule à partir de données qui demandent une analyse préalable du sol [BADEAU et ULRICH, 2008] :

$$RU = \epsilon.\Theta.(1 - C_x)$$

avec :
- ϵ : Epaisseur du sol,
- Θ : Teneur en eau volumique,
- C_x : Charge en éléments grossiers, la réserve utile de ces éléments étant considérée égale à 0.

L'étude pédologique permettant d'accéder à ces informations n'est pas lourde mais nécessite évidemment que le cultivateur en fasse la demande. Ainsi, cet attribut apporte une information intéressante mais non capitale, facilement accessible dès lors que l'agriculteur demande une analyse de son sol, mais dont tous ne disposent pas.

Quant à sa capacité ou non à irriguer, le cultivateur peut en avoir une idée systématique - puisqu'il connait l'état de son matériel d'irrigation - ce qui n'aurait évidemment pas été le cas si nous lui avions demandé un volume précis. Le but étant en outre d'économiser l'eau, l'irrigation ne doit intervenir que comme étai, pour aider à assurer un rendement optimal en maïs, mais l'idée demeure assurément d'y avoir recours le moins possible.

Enfin, en ce qui concerne les données météorologiques, il est difficile pour l'utilisateur d'y avoir accès, mais notre système peut obtenir l'information sans difficultés dès lors que la parcelle est géolocalisée. Si elle ne l'est pas et que le cultivateur n'est pas en mesure de fournir l'information de manière précise, le système peut pourvoir des données approximatives à partir du code postal, en calculant la classe météorologique à laquelle ce code postal appartient en moyenne - pour un code postal donné, il n'y a généralement qu'une classe météorologique correspondante, mais des cas plus ambigus peuvent subvenir dans le cas où des mailles limitrophes couvrent la même ville. Néanmoins, les cas de réelle ambiguïté demeurent rares : les différences ne sont pas sensibles au point que la reconnaissance s'en trouve affectée ; si une parcelle ne se voit pas assigner la bonne maille, mais une maille voisine, les valeurs seront de toutes façons dans le même ordre de grandeur. On considère que les villes des zones agricoles sont suffisamment petites pour que l'écart pouvant occasionnellement apparaître dans les pires cas ne soit pas significativement gênant pour la classification de la parcelle.

Ainsi, les descripteurs permettant de dresser le profil de la parcelle ont bien été sélectionnés dans la mesure où :

- Ils permettent une classification bien représentative de leurs différents profils et validée par les experts agronomes,
- Ils sont disponibles pour un nombre suffisant de données d'entraînement,
- Les données les plus importantes sont systématiquement disponibles pour les agriculteurs au moment de la reconnaissance,
- Les données les moins importantes sont :
  - Soit régulièrement disponibles pour les agriculteurs au moment de la reconnaissance,
  - Soit accessibles, au moins avec une bonne approximation, par le système.

5.1.3 Résultats

Le tableau A.9 de l'annexe IV.2 présente un extrait de la population utilisée pour l'apprentissage. La première année, en 2013, nous n'avions que 50 individus sur lesquels appuyer notre classification. Nous en comptions 88 la seconde année, pour atteindre les 150 présentés ici courant 2015. Ainsi, nous n'avons commencé à expérimenter la classification non supervisée contrainte qu'au cours de l'année 2014. Cet effectif était encore insuffisant et, si les résultats apparaissaient comme pertinents, le système n'était pas utilisable en l'état pour un diagnostic en situation réelle ; chaque classe étant définie à partir d'un nombre d'individus trop faible, aucune d'elles ne pouvait prétendre à une réelle représentativité. Basée sur la population concernant les 88 premiers individus des données d'apprentissage (dont le tableau A.9 sus-cité donne un extrait), elle devait comptabiliser au minimum 16 classes puisque les données d'apprentissage représentent autant de couples "date de semis"/"précocité" possibles.

La classification que nous avons pu réaliser en 2015, comptabilisant donc 150 individus, imposait un minimum de 23 classes. Les résultats sont présentés dans les tableaux 16 et 17 : le tableau 16 décrit les 23 classes obtenues, tandis que le tableau 17 présente la répartition des différents individus dans ces classes.

Pour certaines classes, le nombre d'individus les décrivant est encore trop faible pour que le scenario qu'elles décrivent puisse être considéré comme totalement fiable - il sera nécessaire de préciser, dans un premier temps, à l'agriculteur que le scenario qui lui est soumis est basé sur l'exemple de parcelles similaires à la sienne mais en nombre très faible, et adapter le degré de confiance fourni en conséquence.

La classification complète des parcelles intégrant notre propre classification météorologique figure dans le tableau 18 ; elle a été réalisée à titre purement comparatif, pour éprouver la fiabilité des deux classifications. Nous observons peu de différences entre les deux classifications, ce qui tend à prouver la robustesse de la classification multicouche. Elle offre la possibilité de traiter ensemble des données issues de bases différentes - ce qui nous a permis d'intégrer dans un même processus des données recueillies par des agronomes et une partition préalablement réalisée par Météo France - mais assure également une classification significative des données brutes : si nous n'avions pas pu bénéficier de la partition de Météo France, notre système aurait fourni un modèle quasiment identique, et tout à fait utilisable.

La colonne "classe créée" désigne le numéro de la classe à laquelle appartient l'individu lors de l'apprentissage - ce rapprochement est directement tributaire de la date de semis et de la précocité, puisque la proximité entre un individu et une classe n'est calculé que s'ils disposent du même couple. La colonne "classe reconnue" indique de quelle classe l'individu est le plus proche, sans prise en compte de la date de semis ni de la précocité, ce résultat étant obtenu au cours de la reconnaissance. De nombreuses différences sont observables, parce qu'une classe k_r montre des caractéristiques plus proches de celles de l'individu i à diagnostiquer que sa classe k_a d'origine : cette dernière a évolué lorsque, pendant l'apprentissage, d'autres individus l'ont rejointe, et il a pu se confronter à de nouvelles classes, dont la classe k_r lors de la reconnaissance, puisque la barrière imposée par le couple "date de semis/précocité" a pu être levée. Cette différence est importante et même nécessaire dans un but d'amélioration des pratiques culturales, car il s'agit de proposer à l'agriculteur la configuration la mieux adaptée à sa parcelle, même si elle ne correspond pas à ses habitudes ; le but n'est pas de confirmer la situation actuelle, mais d'en favoriser une plus adaptée à chaque cas particulier.

Pour évaluer la stabilité du système et sa fiabilité dans le temps, nous avons également réalisé l'apprentissage en substituant aux données d'entraînement huit parcelles sélectionnées au hasard - le seul critère de sélection exigé était qu'aucune de ces parcelles ne décrive une classe à elle seule. Les résultats de cette classification sont présentés tableau 19 et 20, le tableau 19 décrivant les classes obtenues, et le tableau 20 la classification des parcelles.

Date Semis	Précocite	Rendement	TypeSol	RU	Irrigation	PluieHiv_Q20	PluieSem_Q20	PluieFlo_Q20	PluieFlo2_Q20	PluieFlo3_Q20	PluieRempQ20	STM6_Q20	TSemQ20	ETPFlo1_Q20	ETPFlo2_Q20	ETPFlo3_Q20	PluieHiv_Q80	PluieSem_Q80	PluieFlo_Q80	PluieFlo2_Q80	PluieFlo3_Q80	PluieRempQ80	STM6_Q80	TSemQ80	ETPFlo1_Q80	ETPFlo2_Q80	ETPFlo3_Q80
5	1	[115,126]	limon:2	208.5	non:2	244.75	181.35	19.5	16.35	15.6	134	2388	5.95	78.15	79.9	75.15	376	270.2	59.15	43.55	56.15	225.45	2634	7.55	91.7	98.95	89.45
5	2	[107;141]	limon:5	206	non:4/oui:1	244.75	181.35	19.5	16.35	15.6	134	2388	5.95	78.15	79.9	75.15	376	270.2	59.15	43.55	56.15	225.45	2634	7.55	91.7	98.95	89.45
5	2	[120;156]	argile:1/sable-argile:3	197	oui:4	244.75	181.35	19.5	16.35	15.6	134	2388	5.95	78.15	79.9	75.15	376	270.2	59.15	43.55	56.15	225.45	2634	7.55	91.7	98.95	89.45
5	3	[90;128]	argile:2/argile-calcaire:1/limon-argile:3	238.75	non:6	310.238	222.684	21.3656	20.825	20.3969	163.725	2262.53	5.70937	67.5094	70.4281	65.3969	505.969	369.072	60.4281	56.2812	60.2281	263.428	2520.47	7.14063	83.3125	88.8187	78.6969
4	1	[112;123]	limon-argile:2	180	oui:2	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
5	1	[125;145]	limon-argile:2	250	oui:2	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
6	2	[115;155]	sable-argile:3	177.5	oui:3	271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
4	2	[115;137]	sable-argile:3/limon-argile:1	161.25	oui:4	278.9	187.7	19.875	17.675	13.425	138.65	2397.62	6.275	74.1375	76.1125	70.5	410.387	293.438	55.075	47.775	55.0875	236.487	2629	7.5875	87.8625	95.8125	86.3375
6	3	[111;128]	sable-argile:2/limon-argile:1	110	oui:3	271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
6	3	[100;115]	limon-argile:2	250	non:2	306.2	245.05	26.15	21	23.95	174.4	2188.5	5.45	63.85	67.25	62.55	521.5	400.25	63.75	64.2	66.65	272.55	2459.5	6.95	80.3	85.1	74.95
4	3	[116;119]	sable-argile:3	120	oui:3	252.675	178.825	18.05	16.475	12.925	129.275	2377.75	6.1	74.625	76.75	71.5	380.75	283.625	53.75	49.025	54.25	224.125	2611.5	7.45	88.725	96.775	87.6
4	3	[126;126]	limon:1	180	non:1	271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
4	2	[121;121]	limon:1	180	non:1	271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
4	3	[112;112]	limon-argile:1	250	non:1	344.4	267.4	32.6	23.8	26.5	211.6	2251	5.8	59.4	66.7	61.1	599	428	67.2	71.5	77.3	320	2549	7.4	78.1	81	71.1
3	3	[118;145]	calcaire:1/argile-calcaire:1/argile-limon:1/sable-argile:5	159.344	oui:8	201.977	163.561	13.0219	14.8266	13.4203	115.592	2312.11	5.56563	74.3859	77.6031	73.3031	343.572	277.602	51.8563	53.2109	53.1875	205.111	2561.72	7.0625	90.2891	97.6328	89.0281
3	1	[134;180]	limon-argile:4/limon-sable:1	185.688	oui:5	264.256	194.444	19.675	17.3813	16.3125	137.05	2325.44	5.85625	74.1	75.4063	70.5688	407.844	305.469	57.9687	48.5312	55.7625	230.137	2565.88	7.3	87.9187	95.25	85.525
3	3	[139;139]	limon-argile:1	250	oui:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
3	2	[120;151]	limon:4/limon-argile:1	226.5	non:5	248.712	180.087	18.775	16.4125	14.2625	131.638	2382.88	6.025	76.3875	78.325	73.325	378.375	276.913	56.45	46.2875	55.2	224.787	2622.75	7.5	90.2125	97.8625	88.525
3	2	[90;136]	sable-argile:3	150	oui:3	233.55	172.95	16.1	15.85	13.05	123.95	2352.5	5.9	74.45	77	72.1	366.7	281.65	52.9	50.75	53.8	216.85	2592	7.3	89.25	97.05	88.1
4	2	[126;139]	sable-argile:9/sable-limon:1	157.48	oui:10	263.433	182.13	19.1469	16.8266	12.8547	132.27	2391.95	6.2125	74.7234	76.6094	71.1625	388.653	284.736	54.2281	48.0547	54.5031	228.217	2622.47	7.53437	88.4297	96.6203	87.3187
2	2	[126;161]	calcaire:2/sable-argile:5	62.6719	oui:7	258.506	181.081	18.8562	16.6781	13.1328	131.548	2387.39	6.15469	75.0156	76.9594	71.6359	385.134	283.045	54.5484	47.9187	54.5828	226.75	2620.31	7.51094	88.8234	96.8844	87.6016
2	2	[119;164]	argile:1/limon:1/limon-argile:4	220	non:1/oui:5	269.684	192.391	19	18.4594	18.1187	149.15	2366.91	5.83437	75.875	78.6594	73.9156	433.278	298.628	61.2156	45.0969	57.6937	244.156	2626.69	7.5	90.6531	96.525	86.7375
5	3	[133;169]	calcaire:1/argile:1	104.5	oui:2	300.2	196.7	19.5	19.4	15.3	150.8	2381.5	6.2	72.15	74.95	69.3	457.15	316.95	56.5	49.2	56.25	251.75	2623	7.55	86.85	93.75	84.05
1	3	[52;52]	argile-calcaire:1	124	non:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
2	3	[54;88]	argile-calcaire:1/limon-argile:1	145.5	non:2	217.7	178	19	15.6	18.4	133.4	2373	5.6	81.5	83.3	79.4	357.2	254.8	63.7	39.8	57.6	219.5	2637	7.5	95.2	101.4	91.8
1	2	[124;161]	calcaire:1/argile:2/argile-calcaire:1/argile-limon:2	171.094	oui:6	187.269	157.987	13.6562	13.3562	12.7938	108.575	2415.06	6.19375	73.6	80.8188	71.9625	325.462	257.156	57	45.5313	52.9875	204.75	2686.12	7.99375	95.4188	103.819	92.225
3	3	[117;134]	limon:2	179.5	non:2	273.15	193.35	19	18.65	18.1	150.2	2366.5	5.85	75.5	78.35	73.55	438.35	301.55	61.05	45.45	57.7	245.8	2626	7.5	90.35	96.2	86.4
1	3	[153;153]	argile-calcaire:1	160	non:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
2	1	[106;192]	argile:1/limon:1/limon-argile:1	218.5	oui:3	245.425	185.675	19	17.125	18.25	141.8	2369.75	5.725	78.5	80.825	76.475	397.775	278.175	62.375	42.625	57.65	232.65	2631.5	7.5	92.775	98.8	89.1
1	2	[138;138]	limon:1	194	oui:1	344.4	267.4	32.6	23.8	26.5	211.6	2251	5.8	59.4	66.7	61.1	599	428	67.2	71.5	77.3	320	2549	7.4	78.1	81	71.1
2	2	[132;159]	argile-calcaire:1/limon:1	246	oui:2	300.2	196.7	19.5	19.4	15.3	150.8	2381.5	6.2	72.15	74.95	69.3	457.15	316.95	56.5	49.2	56.25	251.75	2623	7.55	86.85	93.75	84.05
5	1	[175;175]	limon-sable-argile:1	227	oui:1	344.4	267.4	32.6	23.8	26.5	211.6	2251	5.8	59.4	66.7	61.1	599	428	67.2	71.5	77.3	320	2549	7.4	78.1	81	71.1
3	2	[150;150]	calcaire:1	273	oui:1	217.7	178	19	15.6	18.4	133.4	2373	5.6	81.5	83.3	79.4	357.2	254.8	63.7	39.8	57.6	219.5	2637	7.5	95.2	101.4	91.8
2	1	[182;182]	calcaire:1	273	non:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
4	4	[127;148]	argile-calcaire:3	193.5	oui:3	233.55	172.95	16.1	15.85	13.05	123.95	2352.5	5.9	74.45	77	72.1	366.7	281.65	52.9	50.75	53.8	216.85	2592	7.3	89.25	97.05	88.1
6	3	[146;146]	limon:1	303	oui:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
5	2	[148;148]	limon-argile:1	303	non:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
2	2	[139;148]	argile-calcaire:4	125.625	oui:4	181.59	155	13.7	12.7	12.1	104.6	2460	6.5	72.7	81.8	70.8	317.8	249.9	58.5	42.9	52.6	204.3	2739	8.4	97.3	106.3	93.4
2	3	[143;153]	calcaire:1/argile-limon-sable:1	180	oui:2	233.55	172.95	16.1	15.85	13.05	123.95	2352.5	5.9	74.45	77	72.1	366.7	281.65	52.9	50.75	53.8	216.85	2592	7.3	89.25	97.05	88.1
1	3	[138;156]	sable-argile:4	143.75	oui:4	214.42	167.075	14.15	15.225	13.175	118.625	2327.25	5.7	74.275	77.25	72.7	352.65	279.675	52.05	52.475	53.35	209.575	2572.5	7.15	89.775	97.325	88.6
3	4	[126;131]	argile-calcaire:2/argile-limon:1/sable-argile:1	165	oui:4	214.42	167.075	14.15	15.225	13.175	118.625	2327.25	5.7	74.275	77.25	72.7	352.65	279.675	52.05	52.475	53.35	209.575	2572.5	7.15	89.775	97.325	88.6
2	4	[128;149]	sable-argile:5/sable-limon:1	178.125	oui:6	271.801	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
2	5	[124;132]	argile-calcaire:1/sable-argile:4/sable-limon:1	171.562	oui:6	231.159	172.216	15.8562	15.7719	13.0656	123.284	2349.34	5.875	74.4281	77.0312	72.175	364.944	281.403	52.7937	50.9656	53.7438	215.941	2589.56	7.28125	89.3156	97.0844	88.1625
3	5	[124;125]	argile-calcaire:1/sable-argile:1	180	oui:2	244.75	181.35	19.5	16.35	15.6	134	2388	5.95	78.15	79.9	75.15	376	270.2	59.15	43.55	56.15	225.45	2634	7.55	91.7	98.95	89.45
1	3	[143;144]	calcaire:1/argile-limon:1/argile-limon-sable:1	80	oui:3	195.3	161.2	12.2	14.6	13.3	113.3	2302	5.5	74.1	77.5	73.3	338.6	277.7	51.2	54.2	52.9	202.3	2553	7	90.3	97.6	89.1
2	4	[136;140]	argile-limon-sable:1/sable-argile:2	75	oui:3	233.55	172.95	16.1	15.85	13.05	123.95	2352.5	5.9	74.45	77	72.1	366.7	281.65	52.9	50.75	53.8	216.85	2592	7.3	89.25	97.05	88.1
1	4	[146;149]	argile-calcaire:1/sable-argile:2	135	oui:3	195.3	161.2	12.2	14.6	13.3	113.3	2302	5.5	74.1	77.5	73.3	338.6	277.7	51.2	54.2	52.9	202.3	2553	7	90.3	97.6	89.1
4	4	[127;127]	argile-calcaire:1	150	oui:1	195.3	161.2	12.2	14.6	13.3	113.3	2302	5.5	74.1	77.5	73.3	338.6	277.7	51.2	54.2	52.9	202.3	2553	7	90.3	97.6	89.1

Tableau 16. Définition des classes avec 150 individus en classification supervisée contrainte (Source Météo France - Vivadour)

Parcelle	Classe créée	Classe reconnue
1	1	12
2	2	12
3	3	7
4	3	7
5	4	14
6	4	10
7	5	5
8	6	6
9	7	11
10	5	5
11	8	6
12	9	11
13	2	12
14	2	12
15	7	7
16	3	7
17	10	14
18	4	10
19	11	40
20	9	9
21	11	11
22	12	12
23	13	12
24	8	7
25	8	7
26	14	14
27	15	40
28	15	40
29	16	5
30	17	6
31	11	11
32	8	11
33	18	12
34	18	12
35	19	7
36	15	7
37	20	40
38	21	40
39	22	5
40	6	6
41	23	23
42	9	11
43	2	36
44	19	7
45	7	7
46	4	14
47	10	10
48	23	23
49	21	11
50	24	24
51	15	11
52	16	16
53	25	25
54	25	25
55	26	26
56	27	27
57	28	28
58	16	44
59	22	5
60	16	5
61	26	33
62	29	16
63	22	36
64	18	1
65	4	28
66	27	1
67	16	16
68	22	37
69	22	5
70	29	29
71	18	14
72	29	29
73	3	29
74	30	30
75	31	35
76	18	1
77	32	32
78	2	1
79	1	1
80	33	33
81	34	34
82	21	3
83	35	35
84	31	36
85	4	4
86	22	6
87	36	36
88	37	37
89	26	38
90	26	38
91	38	38
92	38	38
93	26	41
94	39	39
95	40	40
96	39	39
97	20	11
98	40	40
99	20	11
100	20	40
101	41	48
102	41	48
103	42	7
104	42	11
105	41	26
106	42	7
107	42	42
108	41	40
109	43	48
110	43	7
111	44	35
112	44	7
113	38	38
114	26	26
115	20	7
116	15	23
117	45	45
118	15	40
119	45	45
120	46	45
121	47	40
122	47	40
123	46	11
124	48	48
125	42	7
126	46	11
127	47	48
128	43	7
129	43	11
130	43	42
131	43	40
132	21	39
133	38	38
134	21	11
135	20	11
136	15	41
137	35	48
138	45	45
139	20	40
140	15	48
141	40	11
142	20	7
143	20	11
144	20	42
145	40	40
146	35	35
147	42	7
148	19	11
149	21	21
150	21	21

Tableau 17. Répartition des 150 individus dans les 48 classes

Date Semis	Précocité	Rendement	TypeSol	RU	Irrigation	PluieHiv_Q20	PluieSem_Q20	PluieFlo_Q20	PluieFlo2_Q20	PluieFlo3_Q20	PluieRempQ20	STM6_Q20	TSemQ20	ETPFlo1_Q20	ETPFlo2_Q20	ETPFlo3_Q20	PluieHiv_Q80	PluieSem_Q80	PluieFlo_Q80	PluieFlo2_Q80	PluieFlo3_Q80	PluieRempQ80	STM6_Q80	TSemQ80	ETPFlo1_Q80	ETPFlo2_Q80	ETPFlo3_Q80
5	1	[115,126]	limon:2	208.5	non:2	311.25	199.15	18.65	19.45	15.4	159.5	2382.45	6.5	71	73.9	69.4	392.15	280.1	56.6	48.5	52	224	2645.6	7.45	92.6	99.2	90.85
5	2	[107,141]	limon:5	206	non:4/oui:1	311.25	199.15	18.65	19.45	15.4	159.5	2382.45	6.5	71	73.9	69.4	392.15	280.1	56.6	48.5	52	224	2645.6	7.45	92.6	99.2	90.85
5	2	[120,156]	argile:1/sable-argile:3	197	oui:4	254.4	495.4	18.65	13.25	15.7	137.4	2381.3	6.4	74.05	78.8	73.05	383	281.55	70.2	43.2	59.3	236.4	2651.35	8	90.25	98.4	87.65
5	3	[90,127]	argile:2/argile-limon:3	249.5	non:5	330.531	235.562	23.825	20.975	20.0125	189.338	2251.69	6.03125	62.4625	67.3563	64.2562	543.369	392.231	59.9563	61.275	66.5188	277.131	2522.82	7.6	79.5375	83.4437	75.0688
4	1	[112,123]	argile-limon:2	180	oui:2	327.7	209.9	17.6	21.7	17.4	179	2363.3	6.5	64.8	70.1	67.7	518.3	368.2	55	53	56.5	252.2	2647	8	82.9	86.5	78.5
5	1	[125,175]	argile-limon:2/limon-sable-argile:1	238.5	oui:3	336.4	238.8	25.05	22.8	22.05	195.35	2302.95	6.2	62.1	68.4	64.05	558.6	394.9	61.35	62.3	66.05	288	2600.95	7.65	80.4	83.55	74.8
6	2	[115,155]	sable-argile:3	177.5	oui:3	273.575	181.125	20.85	16.05	12.975	132.625	2445.65	6.35	75	77.025	71.375	396.45	297.45	66.6	49.8	54.1	236.075	2686.32	7.8	87.65	96.725	87.45
4	2	[115,115]	argile-limon:1	250	oui:1	330.7	212	17.7	21.9	17.5	180.7	2329.6	6.5	64	69.1	66.5	522.2	364.6	55.5	53.6	57.2	258.1	2621.4	7.9	82.1	85.5	77.7
6	3	[111,128]	argile-limon:1/sable-argile:2	110	oui:3	285.2	185.7	19.5	16.2	13.2	137.8	2439.8	6.5	77.7	78.9	72.8	426.3	314.7	52.5	56.4	54.7	220.1	2664.2	7.8	89.9	96.2	88.5
6	3	[100,115]	argile-limon:2	250	non:2	328.85	257.2	29	19.2	21.5	195.2	2104.05	5.45	60.3	64.15	60.45	558.25	410.95	64.65	68.5	74.75	290.25	2346.3	7.1	75.6	79.95	71.45
4	3	[116,119]	sable-argile:3	120	oui:3	271.45	184.75	19	16.025	13.775	132.475	2386.73	6.25	76.5	77.025	71.6	413.175	317.05	52.325	56.65	54.45	216.725	2625.45	7.575	89.55	95.925	87
4	3	[126,126]	limon:1	180	non:1	293.6	187.3	19.7	17.1	13.3	139.2	2407.7	6.5	77.4	78	71.7	430.2	317.8	52.6	56.7	54.8	228.8	2627.1	7.8	89	96	87.8
4	2	[121,121]	limon:1	180	non:1	293.6	187.3	19.7	17.1	13.3	139.2	2407.7	6.5	77.4	78	71.7	430.2	317.8	52.6	56.7	54.8	228.8	2627.1	7.8	89	96	87.8
4	2	[126,137]	sable-argile:3	155	oui:3	277.45	182.65	20.4	16.1	13.05	134.35	2443.7	6.4	75.9	77.65	71.85	406.4	303.2	61.9	52	54.3	230.75	2678.95	7.8	88.4	96.55	87.8
4	3	[112,112]	argile-limon:1	250	non:1	355.7	276.9	33.6	24.6	26.8	218.3	2141.4	5.5	57.6	64.4	59.9	619	442.2	66.9	73.9	83.1	330.3	2427.7	7.1	75.9	78.9	69
3	3	[118,145]	argile-calcaire:1/sable-argile:5/sable-argile-limon:1	153.688	oui:7	213.878	169.472	15.3938	15.3719	13.0625	112.006	2314.31	5.725	73.95	76.6625	71.9156	350.541	284.819	54.3031	44.8312	47.7625	202.225	2587.46	7.28438	92.95	96.6875	87.625
3	1	[147,147]	argile-limon:1	180	oui:1	327.7	209.9	17.6	21.7	17.4	179	2363.3	6.5	64.8	70.1	67.7	518.3	368.2	55	53	56.5	252.2	2647	8	82.9	86.5	78.5
3	3	[139,139]	argile-limon:2	215	oui:2	277.2	195.05	17.45	18.7	16.35	147.6	2315.05	6.1	69.2	71.9	68.45	444	341.05	53.1	54.35	54.25	230.15	2592.3	7.45	86.2	91.55	81.15
3	2	[120,151]	limon:4/limon-argile:1	226.5	non:5	263.6	182.525	18.7125	15.9125	14.5125	130.863	2436.99	6.325	79.3	80.3375	75.125	396.137	298.813	56.95	51.8125	57.475	224.15	2669.76	7.625	92.3875	98.95	89.8375
3	2	[90,136]	sable-argile:3	150	oui:3	247.85	176.75	19.25	15.7	14.05	123.35	2356.65	5.95	73.9	74.7	70.15	378.2	306.15	60.4	52.15	53.5	221.3	2616.9	7.4	88.1	96.55	85.3
2	3	[126,139]	sable-argile:9/sable-limon:1	157.48	oui:10	266.82	180.916	20.1947	16.1506	13.3572	130.018	2419.29	6.22051	73.9961	76.3695	71.059	384.793	290.228	64.5549	48.2896	53.4008	234.785	2661.15	7.65566	87.6145	96.8264	86.9695
2	2	[126,164]	argile-limon:1/limon-argile:2/sable-argile:5	91.4688	oui:8	250.388	181.752	18.9789	16.8094	13.7883	131.077	2355.68	5.95781	75.093	76.9477	71.882	382.849	288.69	54.5578	47.4281	55.2078	221.741	2599.95	7.29141	89.8906	97.0375	86.9484
5	3	[133,159]	argile:1/sable-argile-limon:1	104.5	oui:2	310.25	200.2	18.7	19.25	15.45	160.5	2358.45	6.45	70.1	73.45	69	477.95	341.8	54.4	55.3	56.3	240.85	2618.65	7.75	85.45	90.3	82.2
1	3	[52,52]	argile-calcaire:1	124	non:1	322.6	207.3	17.3	21.3	17.1	176.2	2423	6.6	66.2	71.2	69	509.4	355.2	54	52	54.8	252	2709	8.1	83.9	88.1	81.2
3	1	[134,180]	limon-argile:3/limon-sable:1	182.625	oui:4	270.9	192.3	19.225	16.2	15.15	142.175	2329.18	5.95	74.725	76.375	70.55	415.15	310.875	56.35	49.65	55.4	225.975	2565.62	7.25	88.85	94.5	86.225
2	3	[54,88]	argile-calcaire:1/limon-argile:1	145.5	non:2	259.5	180.5	18.45	16.25	14.25	126.9	2340.4	6.05	75.65	75.45	70.6	399.5	318.9	51.9	56.65	53.9	216.5	2586.55	7.4	89.25	96.35	85.8
1	2	[124,161]	argile:2/argile-calcaire:1/argile-limon:2/sable-argile-li	171.094	oui:6	185.574	156.809	13.5812	13.2687	12.225	104.172	2418.53	6.23125	73.2781	80.6813	71.8188	321.769	256.403	57.4969	42.2344	50.4	201.016	2699.9	8.10312	97.05	103.669	92.0781
3	3	[117,134]	limon:2	179.5	non:2	283.45	196.85	18.15	18.05	17.75	154.05	2356.05	6.2	72.65	76.25	73.35	436.9	312.4	64.6	47.05	63.05	247.25	2628.5	7.6	89.85	94.05	84.3
5	2	[153,153]	argile-calcaire:1	160	non:1	335.3	214.7	17.9	22.3	17.7	183.2	2277.1	6.4	62.5	68	65.2	529.6	368.9	56.3	54.2	57.9	261.6	2573.1	7.7	81	84.4	75.9
2	1	[106,192]	argile:1/limon:1/limon-argile:1	218.5	oui:3	256.575	344.25	17.525	15.025	17.875	141.85	2398	6.275	76.6	80.275	76.725	394.875	287.2	67.35	42.7	64.75	237.525	2668.15	7.825	93.65	98.825	88.325
2	2	[136,159]	argile:1/limon:2	217.5	oui:3	341.325	216.975	18.9	21.575	18	187.275	2416.4	6.45	67.675	72.625	69.525	552.9	372.075	55.375	52.425	57.125	280.125	2693.5	8.05	85.65	89.425	80.9
5	3	[128,128]	argile-calcaire:1	207	non:1	249.7	208.6	14.4	20.1	18.1	102	2166.4	4.5	67.7	68.4	65	376.8	328.1	47.8	43.2	50.1	183.7	2355	5.7	82.2	87.9	79
2	2	[119,119]	limon-argile:1	178	non:1	335.7	215.1	18.6	22.3	17.9	183.4	2279.6	6.4	62.6	68	65.1	529.5	376.7	56.4	54.4	57.9	261.7	2558	7.7	81	84.4	75.8
1	2	[138,138]	limon:1	194	oui:1	342.1	265.6	32.4	23.7	26.6	210	2276.3	5.9	60.2	67.7	61.6	595	425.2	67.2	71	74.9	317.9	2580.5	7.4	78.7	81.6	71.9
2	2	[132,148]	argile-calcaire:5/sable-argile-limon:2	101.062	oui:7	239.592	174.042	17.7281	15.5047	12.5547	123.869	2424.06	6.375	74.0906	78.4422	70.8938	367.62	273.225	55.9687	45.8609	53.9453	221.484	2670.33	7.89062	91.4969	100.017	89.3859
3	2	[150,150]	sable-argile-limon:1	273	oui:1	288.3	187.3	19.7	16.5	13.3	139.2	2407.7	6.5	77.4	78	71.7	359.8	259.7	65.6	37.2	61.9	222	2677.9	8	97.1	102.9	91.1
2	1	[182,182]	sable-argile-limon:1	273	non:1	331.9	212.6	17.7	21.9	17.6	181.5	2314	6.4	63.8	68.8	66.1	524.6	372.7	55.7	53.6	57.4	257.3	2605	7.9	81.9	85.4	77.4
2	1	[167,167]	cailloux:1	273	oui:1	243	208.6	14.4	19	18.3	102.4	2150.3	4.5	67.6	68.1	65	379.2	327.4	48.2	43.2	50.4	185.4	2340.8	5.7	81.9	87.8	78.8
2	3	[127,148]	argile-calcaire:3	193.5	oui:3	226.275	174.175	15.6	16.225	14	116.825	2334.98	5.9	74.85	74.95	71.975	361.675	294.55	53.25	51.7	50.35	201.15	2591.75	7.275	89.275	97.525	86.725
6	3	[146,146]	limon:1	303	oui:1	345.2	221.3	18.4	22.9	18.2	189.1	2182.1	6.1	60.1	66.3	62.2	545.3	380.2	58.1	55.9	58.8	269.7	2457.2	7.6	78.6	81.6	71.9
5	2	[148,148]	limon-argile:1	303	non:1	395.2	245.8	23	22	20.5	220.1	2440	6.2	72.3	77.7	71	679.7	398.6	61.3	52.7	61	363.4	2686	8.1	90.8	93.7	84.2
1	3	[143,153]	argile-limon-sable:1/sable-argile-limon:1	180	oui:2	249.31	179.2	18.6	16.25	14	124.65	2337.2	6	74.3	74.7	70.2	382.05	302.85	52.9	51.45	53.85	217.8	2587.9	7.3	88.85	96.55	85.3
1	3	[138,156]	sable-argile:4	143.75	oui:4	229.22	173.125	16.65	15.575	14.125	119.175	2317.65	5.75	74.325	75.2	70.95	367.3	300.475	52	53.175	53.3	210.275	2571.35	7.15	89.575	97.075	85.95
3	4	[126,131]	argile-calcaire:2/argile-limon:1/sable-argile:1	165	oui:4	238.6	178.375	18.0875	15.9	14.7875	121.088	2282.81	5.725	73.4375	72.9625	69.15	377.8	313.525	52.3625	54.525	53.5875	212.25	2547.77	7.0875	88.7125	95.7625	83.975
2	4	[128,149]	sable-argile:5/sable-limon:1	178.125	oui:6	272.264	184.119	20.2563	16.8938	12.9063	133.925	2411.13	6.3	74.5688	76.475	70.9125	393.212	286.594	57.7062	47.3375	54.4	233.356	2642.6	7.6375	87.9625	96.575	87.0687
2	5	[124,132]	argile-calcaire:1/sable-argile:4/sable-limon:1	171.562	oui:6	248.834	178.178	19.0906	15.7281	14.3344	123.169	2332.17	5.87813	73.1625	73.8062	69.6469	378.428	306.837	60.4594	52.15	53.1187	222.3	2596.33	7.33125	88.0719	95.9562	84.8375
3	5	[124,125]	argile-calcaire:1/sable-argile:1	180	oui:2	217.75	173.1	17.7	16.35	16.35	123.85	2387.6	5.9	78.75	80.2	77	348.35	254.75	61.25	43.4	47.35	203.9	2640.75	7.6	91.75	99.7	91.25
2	3	[143,144]	argile-limon:1/argile-limon-sable:1/sable-argile-limor	80	oui:3	216.35	169.55	15.95	15.175	14.275	111.075	2282.27	5.65	73.725	74.225	70.525	358.125	306.8	51.375	51.7	50.225	199.3	2560.83	7.125	91.35	96.675	84.975
2	4	[136,140]	argile-limon-sable:1/sable-argile:2	75	oui:3	256.025	179.7	18.625	15.825	14.225	126.525	2350.65	6.05	75.725	75.675	70.925	397.875	317.9	51.9	56.325	53.9	212.5	2601.3	7.375	89.525	96.575	86.075
1	4	[146,149]	argile-calcaire:1/sable-argile:2	135	oui:3	219.2	171.575	16.1	15.225	14.875	115.425	2261.8	5.525	73.425	73.325	69.875	363.325	311.2	51.475	56.1	53.225	204.8	2531.62	6.95	89.35	96.225	84.525
4	4	[127,127]	argile-calcaire:1	150	oui:1	187.3	157.8	12.3	14.6	11.7	100	2293.1	5.6	73.4	77.5	72.8	326.1	268.6	54.3	39.3	42.1	189.2	2587.6	7.4	96.1	96.3	89.3

Tableau 18. Définition des classes décrivant les différents profils de parcelles basée sur notre propre classification météo (Source Vivadour)

Date Semis	Précocite	Rendement	TypeSol	RU	Irrigation	PluieHiv_Q20	PluieSem_Q20	PluieFlo_Q20	PluieFlo2_Q20	PluieFlo3_Q20	PluieRempQ20	STM6_Q20	TSemQ20	ETPFlo1_Q20	ETPFlo2_Q20	ETPFlo3_Q20	PluieHiv_Q80	PluieSem_Q80	PluieFlo_Q80	PluieFlo2_Q80	PluieFlo3_Q80	PluieRempQ80	STM6_Q80	TSemQ80	ETPFlo1_Q80	ETPFlo2_Q80	ETPFlo3_Q80
5	1	[115,126]	limon:2	208.5	non:2	244.75	181.35	19.5	16.35	15.6	134	2388	5.95	78.15	79.9	75.15	376	270.2	59.15	43.55	56.15	225.45	2634	7.55	91.7	98.95	89.45
5	2	[107;141]	limon:5	206	non:4/oui:1	244.75	181.35	19.5	16.35	15.6	134	2388	5.95	78.15	79.9	75.15	376	270.2	59.15	43.55	56.15	225.45	2634	7.55	91.7	98.95	89.45
5	2	[120;156]	argile:1/sable-argile:3	197	oui:4	244.75	181.35	19.5	16.35	15.6	134	2388	5.95	78.15	79.9	75.15	376	270.2	59.15	43.55	56.15	225.45	2634	7.55	91.7	98.95	89.45
5	3	[90;128]	argile:2/argile-calcaire:1/limon-argile:3	238.75	non:6	310.238	222.684	21.3656	20.825	20.3969	163.725	2262.53	5.70937	67.5094	70.4281	65.3969	505.969	369.072	60.4281	56.2812	60.2281	263.428	2520.47	7.14063	83.3125	88.8187	78.6969
4	1	[112;123]	limon-argile:2	180	oui:2	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
5	1	[125;145]	limon-argile:2	250	oui:2	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
6	2	[115;155]	sable-argile:3	177.5	oui:3	271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
4	2	[115;137]	sable-argile:3/limon-argile:1	161.25	oui:4	278.9	187.7	19.875	17.675	13.425	138.65	2397.62	6.275	74.1375	76.1125	70.5	410.387	293.438	55.075	47.775	55.0875	236.487	2629	7.5875	87.8625	95.8125	86.3375
6	3	[111;128]	sable-argile:2/limon-argile:1	110	oui:3	271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
6	3	[100;115]	limon-argile:2	250	non:2	306.2	245.05	26.15	21	23.95	174.4	2188.5	5.45	63.85	67.25	62.55	521.5	400.25	63.75	64.2	66.65	272.55	2459.5	6.95	80.3	85.1	74.95
4	3	[116;119]	sable-argile:3	120	oui:3	252.675	178.825	18.05	16.475	12.925	129.275	2377.75	6.1	74.625	76.75	71.5	380.75	283.625	53.75	49.025	54.25	224.125	2611.5	7.45	88.725	96.775	87.6
4	3	[126;126]	limon:1	180	non:1	271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
4	2	[121;121]	limon:1	180	non:1	271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
4	3	[112;112]	limon-argile:1	250	non:1	344.4	267.4	32.6	23.8	26.5	211.6	2251	5.8	59.4	66.7	61.1	599	428	67.2	71.5	77.3	320	2549	7.4	78.1	81	71.1
3	3	[118;145]	calcaire:1/argile-calcaire:1/argile-limon:1/sable-argile:5	159.344	oui:8	201.977	163.561	13.0219	14.8266	13.4203	115.592	2312.11	5.56563	74.3859	77.6031	73.3031	343.572	277.602	51.8563	53.2109	53.1875	205.111	2561.72	7.0625	90.2891	97.6328	89.0281
3	1	[134;180]	limon-argile:4/limon-sable:1	185.688	oui:5	264.256	194.444	19.675	17.3813	16.3125	137.05	2325.44	5.85625	74.1	75.4063	70.5688	407.844	305.469	57.9687	48.5312	55.7625	230.137	2565.88	7.3	87.9187	95.25	85.525
3	3	[139;139]	limon-argile:1	250	oui:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
3	2	[120;151]	limon:4/limon-argile:1	226.5	non:5	248.712	180.087	18.775	16.4125	14.2625	131.638	2382.88	6.025	76.3875	78.325	73.325	378.375	276.913	56.45	46.2875	55.2	224.787	2622.75	7.5	90.2125	97.8625	88.525
3	2	[90;136]	sable-argile:3	150	oui:3	233.55	172.95	16.1	15.85	13.05	123.95	2352.5	5.9	74.45	77	72.1	366.7	281.65	52.9	50.75	53.8	216.85	2592	7.3	89.25	97.05	88.1
4	2	[126;139]	sable-argile:9/sable-limon:1	157.48	oui:10	263.433	182.13	19.1469	16.8266	12.8547	132.27	2391.95	6.2125	74.7234	76.6094	71.1625	388.653	284.736	54.2281	48.0547	54.5031	228.217	2622.47	7.53437	88.4297	96.6203	87.3187
2	2	[126;161]	calcaire:2/sable-argile:5	62.6719	oui:7	258.506	181.081	18.8562	16.6781	13.1328	131.548	2387.39	6.15469	75.0156	76.9594	71.6359	385.134	283.045	54.5484	47.9187	54.5828	226.75	2620.31	7.51094	88.8234	96.8844	87.6016
2	2	[119;164]	argile:1/limon:1/limon-argile:4	220	non:1/oui:5	269.684	192.391	19	18.4594	18.1187	149.15	2366.91	5.83437	75.875	78.6594	73.9156	433.278	298.628	61.2156	45.0969	57.6937	244.156	2626.69	7.5	90.6531	96.525	86.7375
5	3	[133;169]	calcaire:1/argile:1	104.5	oui:2	300.2	196.7	19.5	19.4	15.3	150.8	2381.5	6.2	72.15	74.95	69.3	457.15	316.95	56.5	49.2	56.25	251.75	2623	7.55	86.85	93.75	84.05
1	3	[52;52]	argile-calcaire:1	124	non:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
2	3	[54;88]	argile-calcaire:1/limon-argile:1	145.5	non:2	217.7	178	19	15.6	18.4	133.4	2373	5.6	81.5	83.3	79.4	357.2	254.8	63.7	39.8	57.6	219.5	2637	7.5	95.2	101.4	91.8
1	2	[124;161]	calcaire:1/argile:2/argile-calcaire:1/argile-limon:2	171.094	oui:6	187.269	157.987	13.6562	13.3562	12.7938	108.575	2415.06	6.19375	73.6	80.8188	71.9625	325.462	257.156	57	45.5313	52.9875	204.75	2686.12	7.99375	95.4188	103.819	92.225
3	3	[117;134]	limon:2	179.5	non:2	273.15	193.35	19	18.65	18.1	150.2	2366.5	5.85	75.5	78.35	73.55	438.35	301.55	61.05	45.45	57.7	245.8	2626	7.5	90.35	96.2	86.4
1	3	[153;153]	argile-calcaire:1	160	non:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
2	1	[106;192]	argile:1/limon:1/limon-argile:1	218.5	oui:3	245.425	185.675	19	17.125	18.25	141.8	2369.75	5.725	78.5	80.825	76.475	397.775	278.175	62.375	42.625	57.65	232.65	2631.5	7.5	92.775	98.8	89.1
1	2	[138;138]	limon:1	194	oui:1	344.4	267.4	32.6	23.8	26.5	211.6	2251	5.8	59.4	66.7	61.1	599	428	67.2	71.5	77.3	320	2549	7.4	78.1	81	71.1
2	2	[132;159]	argile-calcaire:1/limon:1	246	oui:2	300.2	196.7	19.5	19.4	15.3	150.8	2381.5	6.2	72.15	74.95	69.3	457.15	316.95	56.5	49.2	56.25	251.75	2623	7.55	86.85	93.75	84.05
5	1	[175;175]	limon-sable-argile:1	227	oui:1	344.4	267.4	32.6	23.8	26.5	211.6	2251	5.8	59.4	66.7	61.1	599	428	67.2	71.5	77.3	320	2549	7.4	78.1	81	71.1
3	2	[150;150]	calcaire:1	273	oui:1	217.7	178	19	15.6	18.4	133.4	2373	5.6	81.5	83.3	79.4	357.2	254.8	63.7	39.8	57.6	219.5	2637	7.5	95.2	101.4	91.8
2	1	[182;182]	calcaire:1	273	non:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
4	4	[127;148]	argile-calcaire:3	193.5	oui:3	233.55	172.95	16.1	15.85	13.05	123.95	2352.5	5.9	74.45	77	72.1	366.7	281.65	52.9	50.75	53.8	216.85	2592	7.3	89.25	97.05	88.1
6	3	[146;146]	limon:1	303	oui:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
5	2	[148;148]	limon-argile:1	303	non:1	328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
2	2	[139;148]	argile-calcaire:4	125.625	oui:4	181.59	155	13.7	12.7	12.1	104.6	2460	6.5	72.7	81.8	70.8	317.8	249.9	58.5	42.9	52.6	204.3	2739	8.4	97.3	106.3	93.4
2	3	[143;153]	calcaire:1/argile-limon-sable:1	180	oui:2	233.55	172.95	16.1	15.85	13.05	123.95	2352.5	5.9	74.45	77	72.1	366.7	281.65	52.9	50.75	53.8	216.85	2592	7.3	89.25	97.05	88.1
1	3	[138;156]	sable-argile:4	143.75	oui:4	214.42	167.075	14.15	15.225	13.175	118.625	2327.25	5.7	74.275	77.25	72.7	352.65	279.675	52.05	52.475	53.35	209.575	2572.5	7.15	89.775	97.325	88.6
3	4	[126;131]	argile-calcaire:2/argile-limon:1/sable-argile:1	165	oui:4	214.42	167.075	14.15	15.225	13.175	118.625	2327.25	5.7	74.275	77.25	72.7	352.65	279.675	52.05	52.475	53.35	209.575	2572.5	7.15	89.775	97.325	88.6
2	4	[128;149]	sable-argile:5/sable-limon:1	178.125	oui:6	271.801	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
2	5	[124;132]	argile-calcaire:1/sable-argile:4/sable-limon:1	171.562	oui:6	231.159	172.216	15.8562	15.7719	13.0656	123.284	2349.34	5.875	74.4281	77.0312	72.175	364.944	281.403	52.7937	50.9656	53.7438	215.941	2589.56	7.28125	89.3156	97.0844	88.1625
3	5	[124;125]	argile-calcaire:1/sable-argile:1	180	oui:2	244.75	181.35	19.5	16.35	15.6	134	2388	5.95	78.15	79.9	75.15	376	270.2	59.15	43.55	56.15	225.45	2634	7.55	91.7	98.95	89.45
1	3	[143;144]	calcaire:1/argile-limon:1/argile-limon-sable:1	80	oui:3	195.3	161.2	12.2	14.6	13.3	113.3	2302	5.5	74.1	77.5	73.3	338.6	277.7	51.2	54.2	52.9	202.3	2553	7	90.3	97.6	89.1
2	4	[136;140]	argile-limon-sable:1/sable-argile:2	75	oui:3	233.55	172.95	16.1	15.85	13.05	123.95	2352.5	5.9	74.45	77	72.1	366.7	281.65	52.9	50.75	53.8	216.85	2592	7.3	89.25	97.05	88.1
1	4	[146;149]	argile-calcaire:1/sable-argile:2	135	oui:3	195.3	161.2	12.2	14.6	13.3	113.3	2302	5.5	74.1	77.5	73.3	338.6	277.7	51.2	54.2	52.9	202.3	2553	7	90.3	97.6	89.1
4	4	[127;127]	argile-calcaire:1	150	oui:1	195.3	161.2	12.2	14.6	13.3	113.3	2302	5.5	74.1	77.5	73.3	338.6	277.7	51.2	54.2	52.9	202.3	2553	7	90.3	97.6	89.1

Tableau 19. Définition des classes avec 142 individus en classification supervisée contrainte

Parcelle	Classe créée	Classe Reconnue
1	1	
2	2	
3	3	
4	3	
5	4	
6	4	
7	5	
8	6	
9	7	
10	5	
11	8	
12	9	
13	2	
14		12
15	7	
16	3	
17	10	
18		10
19	11	
20	9	
21	11	
22	12	
23	13	
24	8	
25	8	
26	14	
27	15	
28	15	
29	16	
30	17	
31	11	
32		11
33	18	
34	18	
35	19	
36		7
37	20	
38	21	
39	22	
40	6	
41	23	
42		11
43	2	
44	19	
45	7	
46	4	
47	10	
48	23	
49	21	
50	24	
51	15	
52	16	
53	25	
54	25	
55	26	
56	27	
57	28	
58	16	
59	22	
60	16	
61	26	
62	29	
63	22	
64	18	
65		28
66	27	
67	16	
68	22	
69		5
70	29	
71	18	
72	29	
73	3	
74	30	
75	31	
76	18	
77	32	
78	2	
79	1	
80	33	
81	34	
82	21	
83	35	
84	31	
85	4	
86	22	
87	36	
88	37	
89	26	
90	26	
91	38	
92	38	
93	26	
94	39	
95	40	
96	39	
97	20	
98	40	
99	20	
100	20	
101	41	
102	41	
103	42	
104	42	
105	41	
106	42	
107	42	
108	41	
109	43	
110	43	
111	44	
112	44	
113	38	
114	26	
115	20	
116	15	
117	45	
118	15	
119	45	
120	46	
121	47	
122	47	
123	46	
124	48	
125	42	
126	46	
127	47	
128	43	
129	43	
130	43	
131		40
132	21	
133	38	
134	21	
135	20	
136	15	
137	35	
138	45	
139	20	
140	15	
141	40	
142	20	
143	20	
144	20	
145	40	
146	35	
147	42	
148	19	
149	21	
150	21	

Tableau 20. Répartition des 150 individus dans les 48 classes, avec 142 individus en apprentissage et 8 individus en reconnaissance

Les parcelles sélectionnées pour n'intervenir qu'en reconnaissance correspondent aux numéros 14, 18, 32, 36, 42, 65, 69, 131. En comparant les tableaux 17 et 20, nous observons une similarité de 100% quant à la reconnaissance de ces 8 parcelles, prouvant ainsi la robustesse du système. Evidemment, le résultat aurait été bien plus faible si nous avions choisi des parcelles définissant seules des classes, comme c'est le cas, par exemple, pour la 17, puisqu'alors la classe - en l'occurrence la 10 - n'aurait pas pu être créée, et la parcelle 18 n'aurait pas pu y être affectée lors de la reconnaissance. Mais cette difficulté sera levée avec le temps, lorsque le système se verra enrichi de données d'apprentissage complémentaires.

Pour évaluer la qualité de la partition, nous avons considéré chaque classe dans son sous-contexte, c'est-à-dire comparativement aux classes marquées du même couple "date de semis"/"précocité qu'elle ; l'indice ainsi calculé vaut 0.88. La figure 33 montre comment est calculé cet indice, sachant que $C_V(C_k)$ représente le plus proche voisin de C_k dans leur sous-contexte.

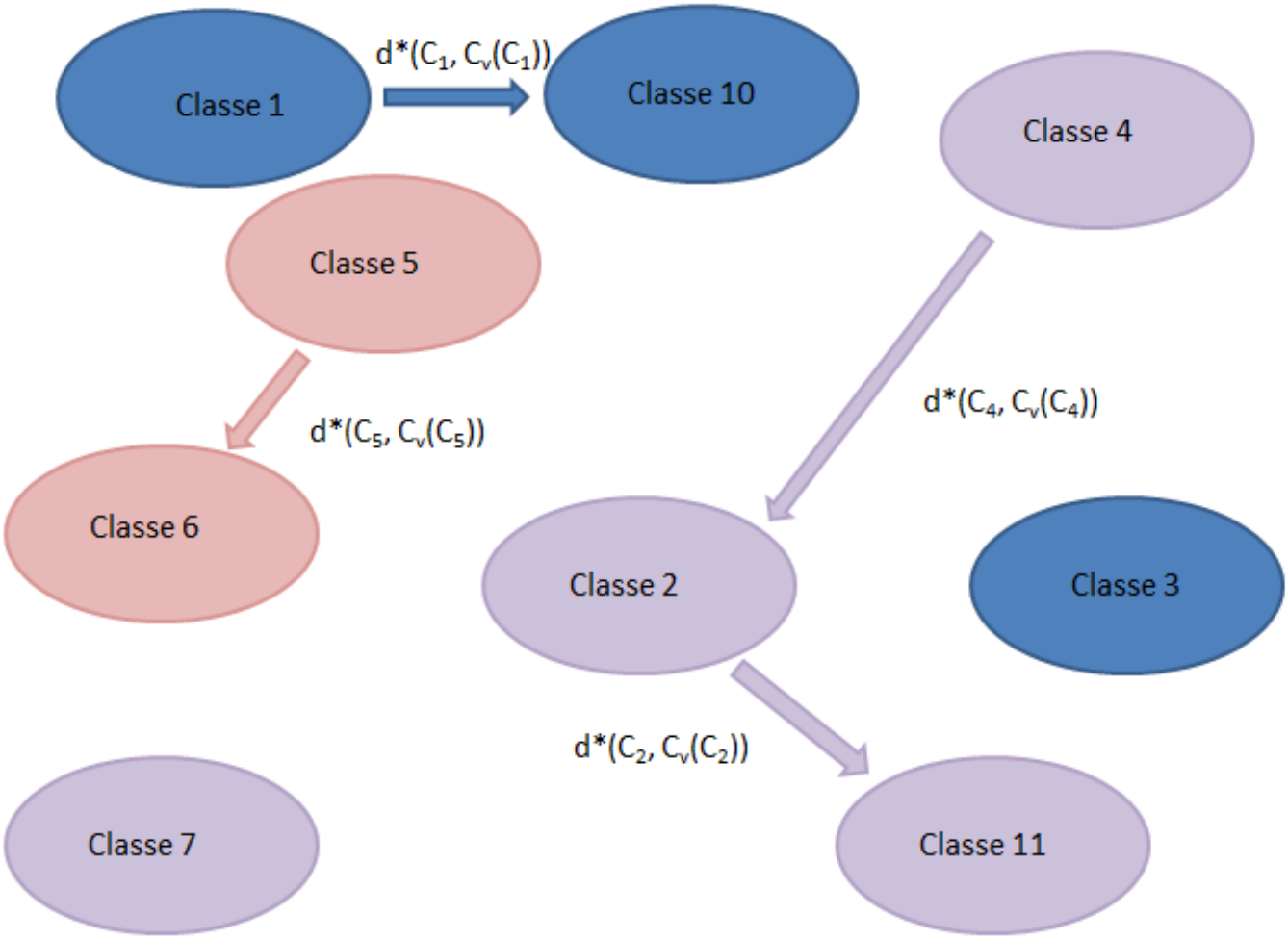


Figure 33. Calcul de l'indice de qualité d'une partition dans le cas d'une classification supervisée contrainte

Dans ce schéma, les classes 1, 3, et 10 appartiennent au même sous-contexte. Il en est de même pour les classes 5 et 6 d'une part, et les classes 2, 4, 7, 11 d'autre part. La classe $C_V(C_k)$ ne peut être le plus proche voisin que d'une classe du même sous-contexte qu'elle, même si des classes appartenant à un même sous-contexte lui ressemblent davantage.

Finalement, pour éprouver l'intérêt de notre architecture multicouche par rapport à une classification simple, nous avons comparé les résultats obtenus lors du traitement des parcelles dans les deux cas.

Nous avons choisi arbitrairement deux classes ; le seul critère de sélection concernait leur ressemblance et, corrélativement, la présence, dans les deux classes, de parcelles similaires. Nous avons représenté, sur la figure 34, les valeurs des MADs des 11 descripteurs météorologiques à la classe sélectionnée à la suite de la classification multicouche et à la classe issue de la classification simple - rappelons que dans la classification multicouche, la valeur de chaque attribut constituant le descripteur météorologique a été remplacée par la valeur de la classe météorologique la plus proche. Pour cette analyse, la classification multicouche a été réalisée à partir de notre propre classification météo.

En outre, à la suite des deux classifications, le CV a été calculé ; il s'élève à 0.69 pour la classification employant l'architecture multicouche et 0.2 pour la classification simple. Les résultats se montrent donc très encourageants puisqu'ici, la représentation des données par le type composite a permis de multiplier par 3 la qualité de la compacité des classes.

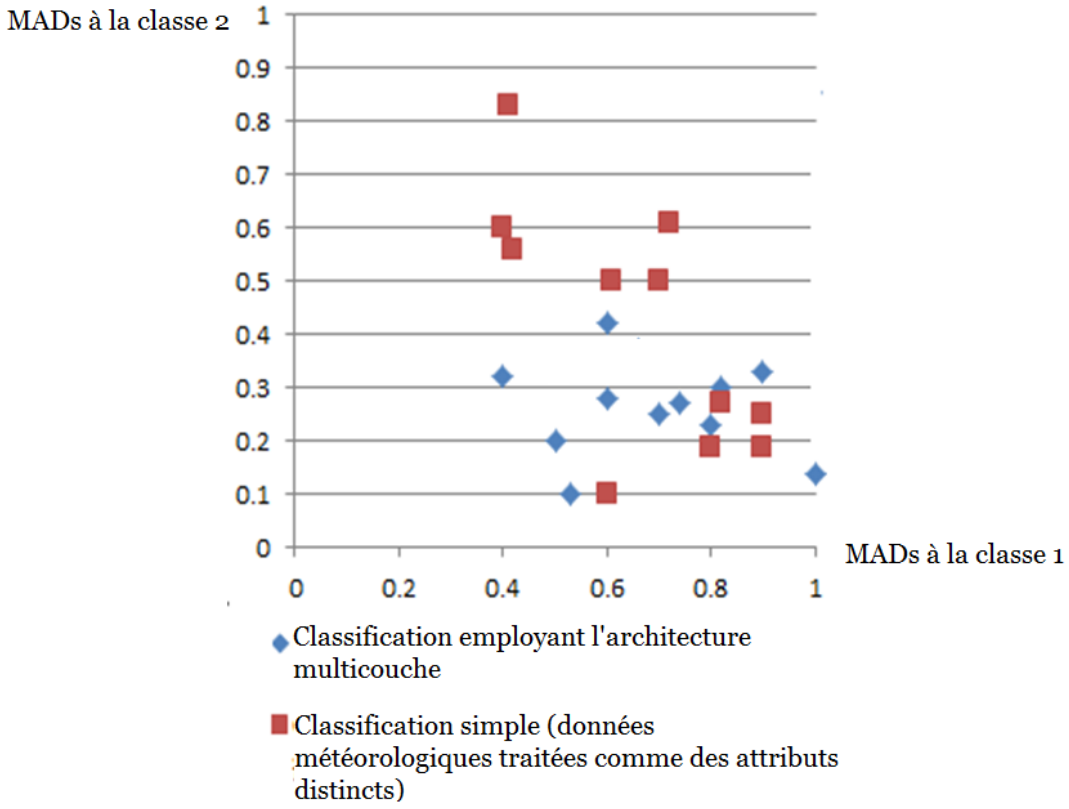


Figure 34. Comparaison des MADs

Pour améliorer la partition et s'assurer de sa représentativité, il faudrait encore davantage de données d'entraînement, car il y a encore trop de classes définies à partir d'une unique parcelle.

5.2 Classification des îlots d'exploitations

Comme nous l'avons indiqué précédemment, le choix des attributs des individus est une phase primordiale dans le travail de classification. Nous avons exposé dans la section 1.3.3 ceux qui nous semblaient le plus à même de répondre à la problématique : quels sont les besoins en eau d'irrigation des différents îlots afin de permettre au gestionnaire de la ressource hydrique du bassin versant de décider au mieux de la répartition de l'eau. La seconde étape concerne les choix des moyens de représenter ces données ; nous détaillons maintenant quels ont été les nôtres.

Contrairement à la partie relative au classement des parcelles, les données étaient toutes parfaitement complètes et nous n'avons pas rencontré de difficultés quant à leur rassemblement. Vous trouverez dans

le tableau A.10 de l'annexe IV.3 une partie du tableau dans lequel étaient répertoriés tous les individus ; le tableau complet contient 2916 individus.

5.2.1 Représentation des données

5.2.1.1 Orientation

Le type composite nous a été à nouveau utile pour représenter la caractéristique "Orientation". Pour rappel, la figure 35 présente une rose des vents.

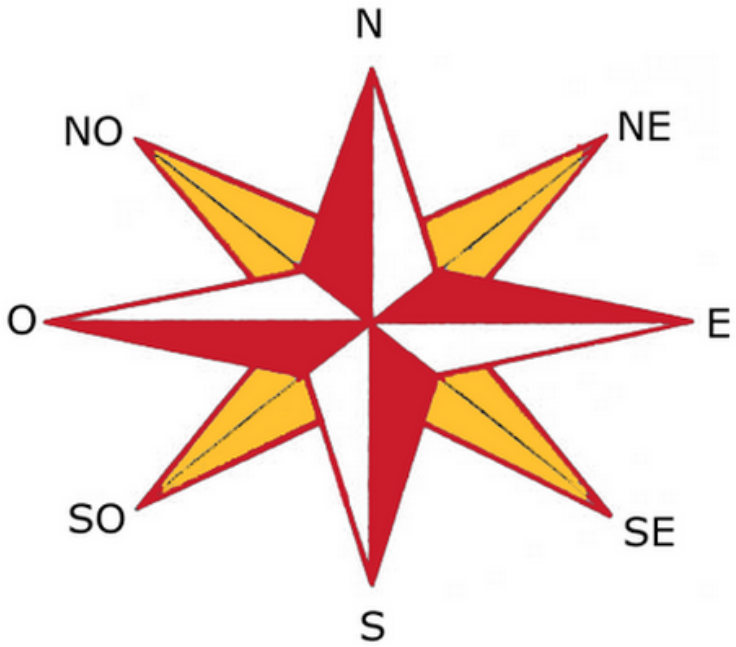


Figure 35. Rose des vents

Ainsi, il faut que, lors de la classification, chaque point cardinal soit traité en fonction de sa position précise sur la rose des vents. Le type choisi doit donc tenir compte des contraintes suivantes :

- La continuité entre les points situés sur la rose des vents (N, puis NE, puis E, puis SE, ...),
- Le caractère cyclique de la rose (le retour à N finalement).

Ainsi la caractéristique "Orientation", représentée par des points cardinaux, n'est pas représentable par une valeur qualitative simple, car alors chaque modalité serait envisagée comme strictement différente des autres ; "N" et "NE", par exemple, ne seraient pas considérés comme plus proches que "N" et "SE". Une valeur qualitative prenant en compte la proximité inter-modalités n'est pas non plus adaptée car elle évaluerait a priori "N" comme aussi proche de "E" que de "S". Enfin, représenter chacun des points par une valeur qualitative (par exemple : 1 pour "N", 2 pour "NE",...) prendrait en considération la graduation entre les différents points mais occulterait le caractère cyclique de la rose des vents, car alors nous aurions 8 pour "NO" et toujours 1 pour "N".

Pour pallier cette complexité, nous avons choisi de représenter cette caractéristique par une variable composite dont les deux sous-attributs constitutifs sont l'abscisse et l'ordonnée des projetés orthogonaux du point cardinal. Les différentes modalités et leur expression sont présentées dans le tableau 21.

	ABS	ORD
N	0	1
NE	√2/2	√2/2
E	1	0
SE	- √2/2	√2/2
S	0	-1
SO	- √2/2	- √2/2
O	1	0
NO	√2/2	- √2/2

Tableau 21. Correspondance entre les points cardinaux et les deux attributs de sa représentation par le type composite

De cette manière, nous avons donc bien une variable - la variable "Orientation" - constituée de deux attributs liés ensemble pour la représenter dans toute sa complexité : les notions de continuité et de cyclicité sont fidèlement retranscrites.

5.2.1.2 Autres caractéristiques

Les caractéristiques "Distance au cours d'eau", "Altitude entre l'îlot et la retenue collinaire la plus proche", "Distance entre l'îlot et le réseau en concession d'état", et "Altitude de la parcelle" étant exprimées en mètres, leur représentation par une valeur quantitative était immédiate ; il en a été de même pour la caractéristique "Pente" exprimée en degrés. Dans le tableau A.10 présenté annexe IV.3, les colonnes correspondant à chacun de ces attributs sont respectivement : "PROXIMIT(0,1)_", "DIFALT_COL", "PROXIMITE_1", "DELTAALT", "PENTEMOY". L'attribut "Type du point de prélèvement", quant à lui, a été représenté par une variable de type qualitatif simple pouvant admettre trois modalités : "barrage", "DCE", et "graviere", car il n'y a aucun lien notable entre aucune de ces valeurs. Il est indiqué dans le tableau A.10 par la colonne "type_inra",

5.2.2 Résultats

Pour ce lot, les données d'apprentissage non supervisé et de reconnaissance sont les mêmes. La partition obtenue au terme de la phase d'apprentissage est présentée dans le tableau 22. A partir des 2916 individus de départ, nous avons défini 5 profils d'îlots, avec un indice de qualité de partition évalué à 0.59, les poids des attributs étant indiqués dans le tableau 23. L'indice de partition n'a pas été maximisé, parce que les experts attendaient un nombre final de classes compris entre 5 et 10 pour avoir une vision suffisamment globale de la zone concernée ; aussi a-t-il fallu définir un indice d'exigence relativement faible, en l'occurrence 0.6, pour atteindre cet objectif. Un indice d'exigence monté à 0.75 permettait d'obtenir 8 classes mais le taux d'individus reconnus chutait : pour les 5 classes présentées dans le tableau 22, une opération de reconnaissance effectuée sur les mêmes parcelles que celles employées pour l'apprentissage (les 2916 individus) portait le taux d'individus affectés à la même classe lors de la reconnaissance et lors de l'apprentissage à 51%. Celui-ci tombait à 14% lorsque nous réalisions la même opération en conservant les 8 classes décrites dans le tableau 24.

Distance cours d'eau	Altitude îlot	Pente	Distance réseau	type de retenue	Altitude (îlot	Abscisse	Ordonnée
1693.7	18.7953	3.31568	9218.83	DCe:8/barrage:1693	-0.865361	0.00	0.97
2598.92	15.7861	2.60542	9819.35	DCe:14/barrage:532	277.915	-0.18	0.88
17088.8	50.3528	0.320846	14019.8	barrage:85	-110.81	0.59	0.3
2848.21	41.6283	3.80817	11324.6	barrage:576	5.12719	-0.77	-0.37
882.761	19.5	2.87888	9133.74	DCe:6	11.6625	0.04	0.71

Tableau 22. Description des 5 classes définies par apprentissage non supervisé

Descriptors	normalisés
Distance cours d'eau	0.18
Altitude îlot	0.01
Pente	0.02
Orientation	0.18
Distance réseau	0.19
type de retenue	0.28
Altitude (îlot/collinaire)	0.14

Tableau 23. Poids normalisés des attributs

Distance cou	Altitude îlot	Pente	Distance réseau	type de rete	Altitude (îlot/collinaire)	Abscisse	Ordonnée
2446.56	51.4266	14.0682	1238.51	barrage:945	-72.5165	-0.22	0.64
430.591	9.99741	4.42008	4070.11	barrage:925	-18.0536	0.16	-0.08
17088.8	50.3528	0.320846	14019.8	barrage:278	-110.81	0.59	0.3
863.381	17.4284	3.78463	28127.2	DCe:9/barra	-5.50851	0.17	-0.71
3179.06	30.6956	3.92346	9500.49	barrage:213	2.05126	0.01	0.99
1816.9	25.5521	4.05392	9056.17	barrage:116	-16.4487	-0.77	0.02
3545.19	54.19	3.93479	12719.5	barrage:130	14.4225	-0.66	-0.42
628.648	17.1572	2.89757	9454.97	DCe:19	10.0455	0.06	0.75
2810.39	9.5	2.6766	9774.04	barrage:4	282.625	-0.26	0.89

Tableau 24. Description des 8 classes définies par apprentissage non supervisé

Nous avons par la suite sélectionné aléatoirement 400 îlots parmi ces 2916 individus, de manière à réaliser un apprentissage sur les 2516 individus restants, et la reconnaissance sur ces 400 zones, avec un indice d'exigence fixé à 0.6 puisque c'est la valeur ayant été arrêtée au terme des tests précédents. L'apprentissage a permis de dessiner des classes très proches de celles définies lors de l'apprentissage opéré sur les 2916 individus ; ces résultats sont présentés dans le tableau 25. Le taux de ressemblance dans l'affectation des individus aux classes dans les deux cas est de 97%, et le taux de ressemblance dans la reconnaissance des 400 individus sélectionnés est de 81%.

Distance cou	Altitude îlot	Pente	Distance rés	type de retenue	Altitude (îlot	Abscisse	Ordonnée
1693.7	18.7953	3.31568	9218.83	DCe:8/barrage:924	-0.865361	0	0.97
2599.33	15.8013	2.60613	9826.26	DCe:14/barrage:900	277.943	-0.18	0.88
17088.8	50.3528	0.320846	14019.8	barrage:85	-110.81	0.59	0.3
2848.21	41.6283	3.80817	11324.6	barrage:576	5.12719	-0.77	-0.37
882.761	19.5	2.87888	9133.74	DCe:6	11.6625	0.04	0.71

Tableau 25. Description des 5 classes définies par apprentissage non supervisé sur 2516 individus sélectionnés aléatoirement

A partir de cette partition, les experts seront en mesure d'attribuer à chacun des profils une estimation de la quantité d'eau requise et d'organiser au mieux leur distribution.

5.3 Conclusion

L'introduction dans la méthode LAMDA du type composite a permis de représenter les données météorologiques et l'orientation sans perdre d'information et en tenant compte de leur unité. Le type sol a pu être pris en compte par l'intermédiaire du type liste, sans lequel deux sols dont la composition était ressemblante mais pas précisément similaire auraient été considérés comme totalement différents.

La capacité de LAMDA à traiter des données lacunaires lors de la reconnaissance nous a permis de sélectionner des attributs dont l'importance pour la question de l'évaluation des besoins en eau d'une parcelle et de son rendement ne faisait aucun doute pour les experts agronomes, alors même qu'il n'est pas garanti que les agriculteurs détiennent systématiquement cette information. Sans cela, il aurait fallu réduire conséquemment le nombre de descripteurs, ce qui aurait sensiblement affecté la classification, ainsi que la fiabilité et l'intérêt technique du système : plus l'outil dispose d'informations sur lesquelles appuyer son analyse, plus il est en mesure d'opérer une classification réaliste et informative - quitte ensuite, à éliminer des caractéristiques si elles témoignent d'un poids insignifiant dans l'algorithme.

Le système a été validé, en l'état, sur la zone test. Par la suite, il pourra être étendu et concerner une partie plus importante du territoire ; cette opération devra néanmoins être précédée d'une phase d'apprentissage sans laquelle il ne saurait être fiable. En effet, si les parcelles au sein d'un même département témoignent d'une relative similarité, il est évident qu'en s'en éloignant, les disparités s'accentueraient, affectant sensiblement le diagnostic.

En outre, en ce qui concerne la classification des parcelles, le système nécessite encore des données d'apprentissage pour s'affiner et pouvoir prétendre à une certaine exhaustivité. Actuellement, nous n'avons pas pu intégrer à notre système des parcelles à la précocité "très précoce" par exemple. Néanmoins, il est opérationnel et la partition actuellement fixée couvre un nombre de cas représentatifs de la zone test suffisant pour pouvoir être mis en oeuvre en situation réelle et fournir un conseil consistant à une grande partie des producteurs de maïs. Au fil des années, la partition se précisera et élargira l'ensemble de ses connaissances, en mettant à jour ses informations après que les agriculteurs auront fait remonter les informations relatives à leur parcelle et les stratégies employées au cours de la campagne.

La classification des îlots d'exploitations, quant à elle, bénéficiait d'un grand nombre de données d'entraînement et est tout à fait fonctionnelle et exploitable, sans intervention supplémentaire. Un rétrocontrôle ne sera pas nécessaire puisque les informations dépendent essentiellement de données satellitaires ayant été accessibles dès le début du projet. Evidemment là encore, si la zone couverte par le système se voit finalement étendue, un apprentissage portant sur les terrains supplémentaires sera nécessaire.

Conclusion

La présente thèse s'inscrit dans un projet agronomique et propose, dans un premier lot, une stratégie de diagnostic pour fournir aux cultivateurs de maïs un conseil adapté à leur parcelle, intégrant dans l'analyse des informations qu'ils sont tous en mesure de recueillir aisément. La partie de notre implication relative au deuxième lot réside dans la conception d'un système d'aide à la décision pour le gestionnaire de la ressource hydrique dans le bassin versant, afin de lui apporter les informations nécessaires à une distribution optimale de l'eau d'irrigation aux différents îlots d'exploitations de sa zone. Dans les deux lots, l'objectif du projet MAISEO était d'économiser au maximum l'eau d'irrigation dévolue aux cultures de maïs tout en assurant, pour le moins, la conservation du rendement actuel.

Dans notre travail, nous réinvestissons la méthode LAMDA pour l'adapter à ces deux situations précises. Notre implication dans ces deux lots intervient à deux moments : celui de l'apprentissage et celui du diagnostic ; aussi notre recherche s'inscrit-elle tant dans le champ de la reconnaissance que dans celui de la classification non supervisée.

L'intérêt de cette méthode réside dans sa capacité à traiter rapidement un nombre élevé de données multidimensionnelles, et de différents types. L'utilisation de la logique floue revêt en outre l'avantage de prendre en compte les incertitudes inhérentes au domaine agricole et la difficulté de classer les différentes parcelles en classes strictes et drastiquement séparées. Elle est en outre très maniable et offre à l'utilisateur un rétro-contrôle nécessaire à une bonne maîtrise de la partition des classes : il est possible, dans une certaine mesure, de gérer la quantité de classes crées, la flexibilité de la classification, de connaître le poids particulier de chaque attribut dans cette situation. Travaillant avec des experts en agronomie, il était important de pouvoir leur fournir toutes ces informations afin d'étayer leur analyse de la partition obtenue et leur caractérisation des différents profils ainsi définis.

Néanmoins, malgré la grande adaptabilité de LAMDA et sa capacité à traiter des types de données hétérogènes, certaines caractéristiques ne pouvaient être traitées dans cette situation sans perte d'information ; dans le monde réel, tout n'est pas représentable par un mot unique, un nombre, ou un intervalle. Pour pallier cette limite, nous nous sommes intéressés à de nouvelles manières génériques de représenter les données, par l'extension de types existants ou la création d'un nouveau, afin de couvrir le maximum de cas réels.

La principale contribution de notre travail concerne les types de données traités par LAMDA. Les caractéristiques des parcelles de maïs, pour être traitées de manière optimale, nécessitait l'analyse spécifique de données composites de manière à prendre en compte sans perte les informations les constituant. Cette étude nous a amenés à apporter deux modifications aux types de données intégrées par LAMDA. La première concerne les données de type qualitatif, puisqu'il s'agissait de s'intéresser non plus aux modalités dans leur globalité mais dans leur structure interne, de telle sorte que deux modalités ne sont plus considérées immédiatement comme complètement différentes dès que les mots les représentant ne sont pas strictement identiques, mais peuvent être rapprochées s'ils présentent des similarités évidentes. La prise en compte de la proximité entre les différentes modalités, permettant de rompre avec

l'évaluation binaire de la similarité entre variables qualitatives, autorise un traitement plus souple de ce type de données, une meilleure adaptation à la complexité langagière. Elle nous a en outre permis d'introduire la gestion des listes dans LAMDA, par le traitement de la modalité non plus dans sa globalité mais dans les unités syntaxiques minimales qui la composent. La taille de ces listes n'est pas contrainte et, pour une même caractéristique, chaque individu peut bénéficier de sa propre liste, à la taille qui sied. L'algorithme de traitement de ce type de données se base alors sur la structure syntaxique de chaque modalité, mais également sur le nombre d'éléments présents dans chaque liste.

La deuxième modification apportée relève de la création d'un type de données à part entière, s'inspirant de l'architecture orientée-objet : le type composite. Ce nouveau type implique une architecture multi-couches et tolère ainsi la gestion de données multi-dimensionnelles sans perte d'informations par le classifieur. Il se base ainsi sur un traitement en profondeur des données, chaque caractéristique de type composite devenant à son tour un individu à part entière, devant être classé dans une couche inférieure. Cette méthode permet une bonne visibilité des processus en jeu et fournit à l'utilisateur une partition spécifique à chaque couche. Le type composite résulte d'une volonté de faire correspondre dans la classification une donnée à la perception que l'utilisateur en a dans la réalité : s'il perçoit un objet dans sa globalité, il nous a semblé intéressant de permettre au classifieur de bénéficier de la même perception afin d'adapter son traitement à ce que l'humain peut en attendre. Ainsi, il peut par exemple être employé pour représenter des caractéristiques manifestement liées entre elles pour former une entité signifiante, que chacune d'elles prennent sens auprès des autres ou non. Par exemple, nous avons défini dans le lot 1 la caractéristique "météo" comme étant l'agrégation des mesures de précipitations, d'évapotranspiration, et de températures - car il s'agissait des informations particulières qui nous importaient dans le contexte de la culture du maïs - mais nous avons éludé d'autres phénomènes atmosphériques qui, dans la réalité, sont tout aussi constitutifs de l'entité "météo". Notre caractéristique "météo" dépend de l'ensemble de ces caractéristiques parce que nous l'avons définie ainsi, mais ce n'est qu'une adaptation de ce que nous appelons "Météo" dans la vie quotidienne et chacune d'elles garde sa signifiance intrinsèque, même si elles sont toutes liées entre elles dans un même contexte météorologique. A l'inverse, dans la caractéristique "Orientation", les variables "Abscisse" et "Ordonnée" ne prennent sens dans ce contexte qu'ensemble.

Nous nous sommes également intéressés à une première approche de la gestion de données lacunaires, mais notre réflexion n'a porté pour le moment que sur l'aspect reconnaissance et ne permet qu'une estimation de la fiabilité de la partition ainsi obtenue. Il serait intéressant d'approfondir cet aspect, par exemple en proposant des extrapolations de données manquantes, de manière à améliorer cette fiabilité.

Nous avons montré l'intérêt de notre apport par l'évaluation de la qualité des partitions obtenues après exploitation de données test et de données réelles ; les modifications apportées aux types de données initiaux assurent une représentation des données plus immédiate, un traitement des informations des données en entrée plus réaliste, et une séparation plus claire des classes : les distances inter-classes sont augmentées et les distances intra-classes sont réduites.
Ces tests ont été réalisés dans le cadre du projet MAISEO et, les résultats ayant été validés par les experts agronomes, l'outil de classification a pu être intégré au système d'aide aux agriculteurs et aux gestionnaires de la distribution d'eau d'irrigation du bassin versant.

Pour les besoins du lot 1, il nous a en outre fallu adapter l'algorithme de LAMDA à une situation particulière ; il ne s'agissait plus simplement de définir des classes en fonction de la proximité des différents individus, prenant en compte l'ensemble de leurs caractéristiques sans connaissance a priori de leurs différents poids : nous devions contraindre la classification non supervisée, de manière à créer au sein du contexte global des sous-contextes décrits respectivement par un couple "date de semis"/"précocité du grain" spécifique.

L'objectif des traitements opérés sur les types de données étant de faire mieux correspondre la classifi-

cation avec la réalité, en transmettant au classifieur l'ensemble des informations et des caractéristiques auxquelles nous avons accès en tant qu'humains, dans leurs inter-relations et leurs contraintes, il serait intéressant de réfléchir à une méthode d'évaluation de la classification selon des critères sémantiques. En effet, l'indice de qualité de la classification que nous sommes actuellement en mesure de calculer évalue la distance entre les classes et leur compacité, mais sans analyse de leur structure interne, de leur réelle signification.

Enfin, nous pensons réfléchir à l'intérêt que pourrait présenter l'architecture multicouche pour des travaux de prédiction, en particulier pour étudier l'évolution temporelle du comportement d'un système décrit par des variables multidimensionnelles.

Annexe

A.1 Annexe I : Le projet MAISEO

A.1.1 Contexte du projet

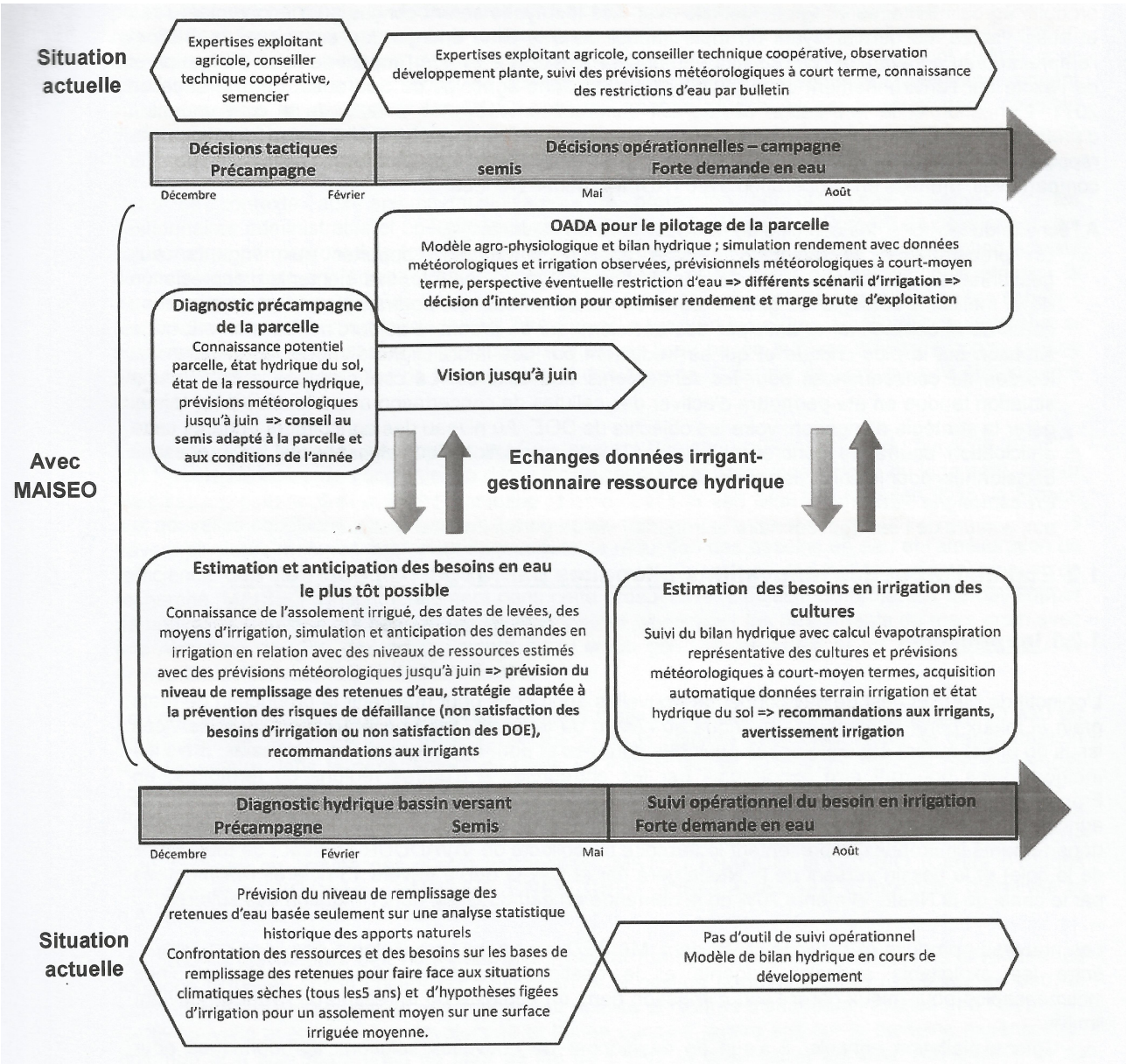


Figure A.1. Déroulement et motivations du projet MAISEO

Cultures	Besoins en eau en mm (10 m3/hectare)
Canne à sucre	1250
Bananes	1200
Dattes	1100
Pamplemousses	825
Riz	770
Coton	750
Betterave à sucre	650
Soja	637
Arachide	600
Maïs	575
Blé	550
Patates douce	537
Pommes de terre	487
Sorgho	475
Oignons	475
Tomates	450
Tabac	400
Haricots	375

Tableau A.1. Les cultures qui consomment le plus d'eau

A.1.2 Les données météorologiques

Contrainte climatique	Paramètre météorolo-gique	Période	Description agronomique	Ordre d'importance sur le dév. cultural (1 = très important ; 5 faible)	Poids à attribuer dans l'analyse
Disponibilité en eau	Cumul précipitation	1/11/n-1→ 14/03/n	Recharge hivernale	5	1
		15/03→14/06	Semis	4	1
		15/06→04/07	Floraison 1	3	2
		05/07→25/07	Floraison 2	1	3
		26/07→14/08	Floraison 3	2	2
		15/08→01/10	remplissage	5	1
Accumulation de chaleur	Somme de température	15/03→31/10	Cycle cultural		2
froid	Température minimale quotidienne moyenne	15/03→30/04	Début de cycle		2
Besoin en eau	Cumul d'évapo-transpiration potentielle	15/06→04/07	Floraison		2 (retenu après tests)
		05/07→25/07			
		26/07→14/08			

Tableau A.2. Paramètres météorologiques d'intérêt pour la culture du maïs

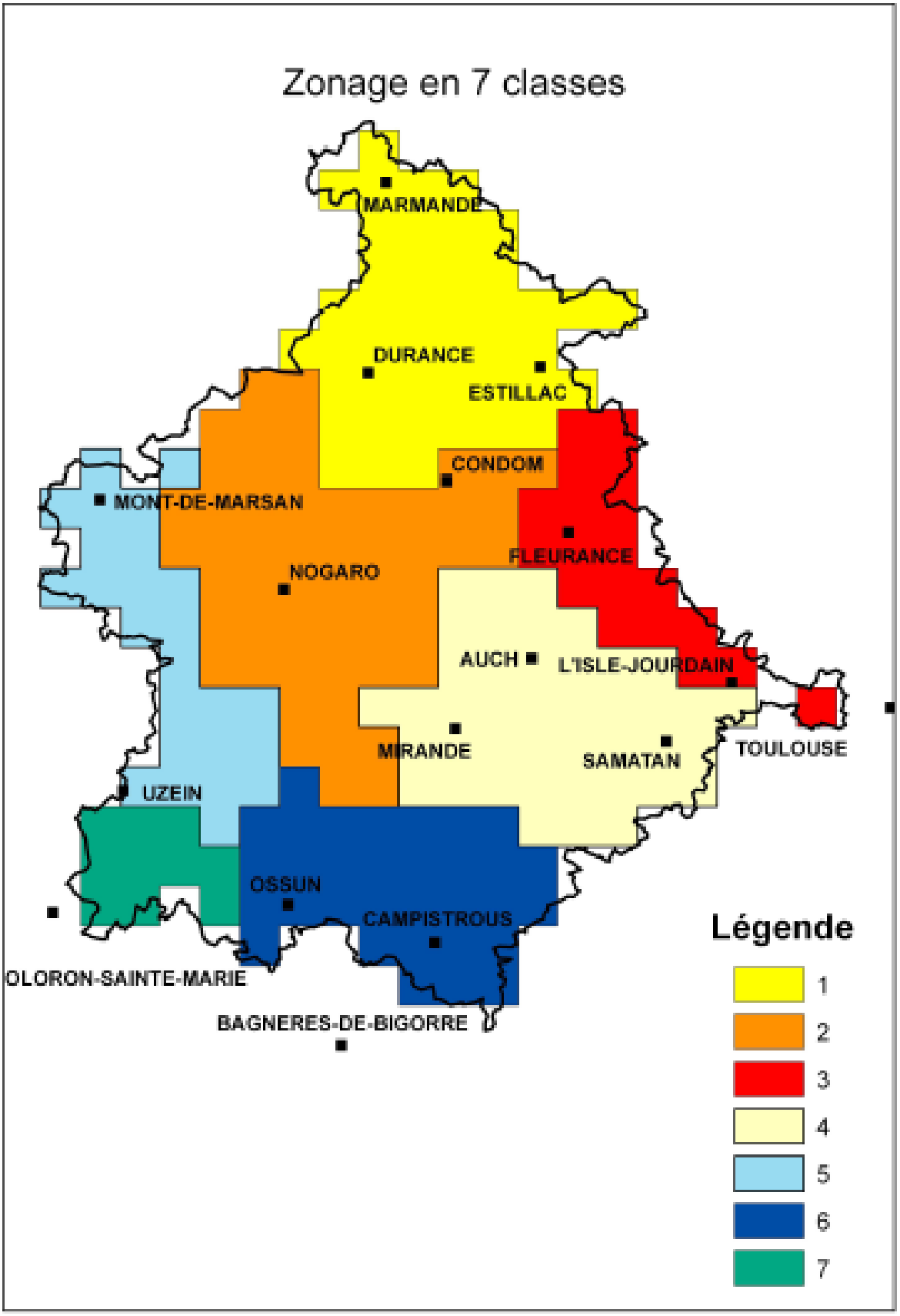


Tableau A.3. Zonage MAISEO sur les 226 mailles SAFRAN du domaine (Source : Météo France - Vivadour)

	Q20 groupe 1	Q20 groupe 2	Q20 groupe 3	Q20 groupe 4	Q20 groupe 5	Q20 groupe 6	Q20 groupe 7
Pluie HIVER (mm)	217,7	271,8	181,6	195,3	328,6	268,0	344,4
Pluie SEMIS (mm)	178,0	184,7	155,0	161,2	208,7	222,7	267,4
Pluie FLORAISON 1 (mm)	19,0	20,0	13,7	12,2	19,0	19,7	32,6
Pluie FLORAISON 2 (mm)	15,6	17,1	12,7	14,6	21,7	18,2	23,8
Pluie FLORAISON 3 (mm)	18,4	12,8	12,1	13,3	17,8	21,4	26,5
Pluie REMPLISSAGE (mm)	133,4	134,6	104,6	113,3	167,0	137,2	211,6
STM6 CYCLE (°C)	2373	2403	2460	2302	2360	2126	2251
TNmoy SEMIS (°C)	5,6	6,3	6,5	5,5	6,1	5,1	5,8
ETP FLORAISON 1 (mm)	81,5	74,8	72,7	74,1	69,5	68,3	59,4
ETP FLORAISON 2 (mm)	83,3	76,5	81,8	77,5	73,4	67,8	66,7
ETP FLORAISON 3 (mm)	79,4	70,9	70,8	73,3	67,7	64,0	61,1

Tableau A.4. Médiane spatiale des Q20 (Source : Météo France)

	Q80 groupe 1	Q80 groupe 2	Q80 groupe 3	Q80 groupe 4	Q80 groupe 5	Q80 groupe 6	Q80 groupe 7
Pluie HIVER (mm)	357,2	394,8	317,8	338,6	519,5	444,0	599,0
Pluie SEMIS (mm)	254,8	285,6	249,9	277,7	348,3	372,5	428,0
Pluie FLORAISON 1 (mm)	63,7	54,6	58,5	51,2	58,4	60,3	87,2
Pluie FLORAISON 2 (mm)	39,8	47,3	42,9	54,2	51,1	56,9	71,5
Pluie FLORAISON 3 (mm)	57,6	54,7	52,6	52,9	57,8	56,0	77,3
Pluie REMPLISSAGE (mm)	219,5	231,4	204,3	202,3	272,1	225,1	320,0
STM6 CYCLE (°C)	2637	2631	2739	2553	2615	2370	2549
TNmoy SEMIS (°C)	7,5	7,6	8,4	7,0	7,5	6,5	7,4
ETP FLORAISON 1 (mm)	95,2	88,2	97,3	90,3	85,5	82,5	78,1
ETP FLORAISON 2 (mm)	101,4	96,5	106,3	97,6	91,0	89,2	81,0
ETP FLORAISON 3 (mm)	91,8	87,1	93,4	89,1	81,0	78,8	71,1

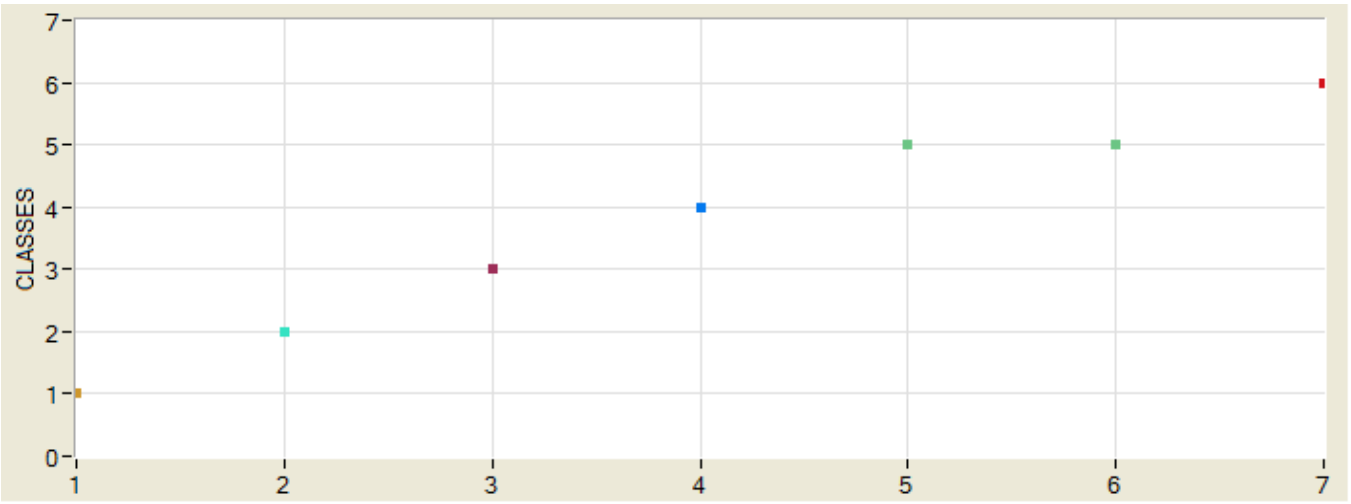
Tableau A.5. Médiane spatiale des Q80 (Source : Météo France)

A.2 Annexe II : Les méthodes de classification

	ANN	Arbres de décision	SVM	K-voisins	K-Moyennes	Cartes de Konohen	LAMDA
Type de données	Numérique	numérique et symbolique	Numérique	Numérique	Numérique	Numérique	Numérique et Symbolique
Paramétrique	Non	Non	Oui	Non	Non	Non	Non
Classification stricte	Oui	Oui	Oui	Oui	Oui, mais flou possible	Oui	Oui, mais flou possible
Mode supervisé	Oui	Oui	Oui	Oui	Non	Non	Oui
Mode non supervisé	Oui	Non	Non	Non	Oui	Oui	Oui
Nombre de classes connu a priori	Oui	Oui	Oui	Non	Non	Non	Oui ou Non
Classes évolutives	Non	Non	Non	Non	Non	Oui	Oui

Tableau A.6. Tableau comparatif des différentes méthodes de classification

A.3 Annexe III : Prise en considération de la proximité des modalités d'une même variable qualitative

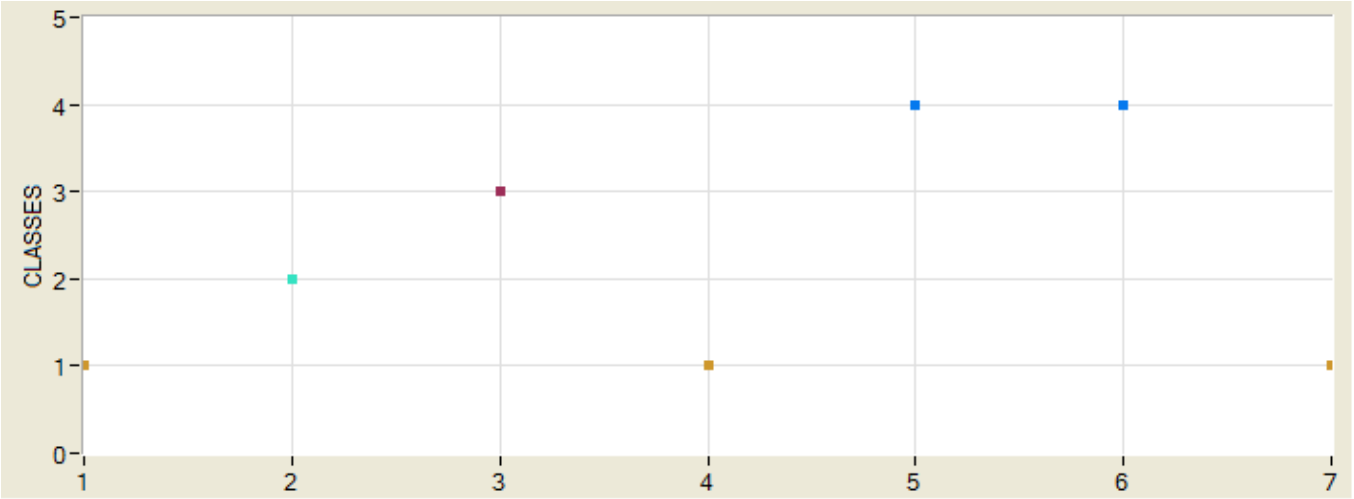


a. Répartition des individus dans les classes

	Forme	Couleur
Classe 1	triangle: 1	Rouge : 1
Classe 2	triangle-isocele: 1	Bleu : 1
Classe 3	cercle : 1	Rouge : 1
Classe 4	triangle-rectangle: 1	Rouge : 1
Classe 5	rectangle: 2	Bleu : 2
Classe 6	triangle-isocele: 1	Rouge : 1

b. Description des classes

Figure A.2. Résultats de la classification des individus décrits par leur forme et leur couleur sans prise en considération de la proximité des modalités

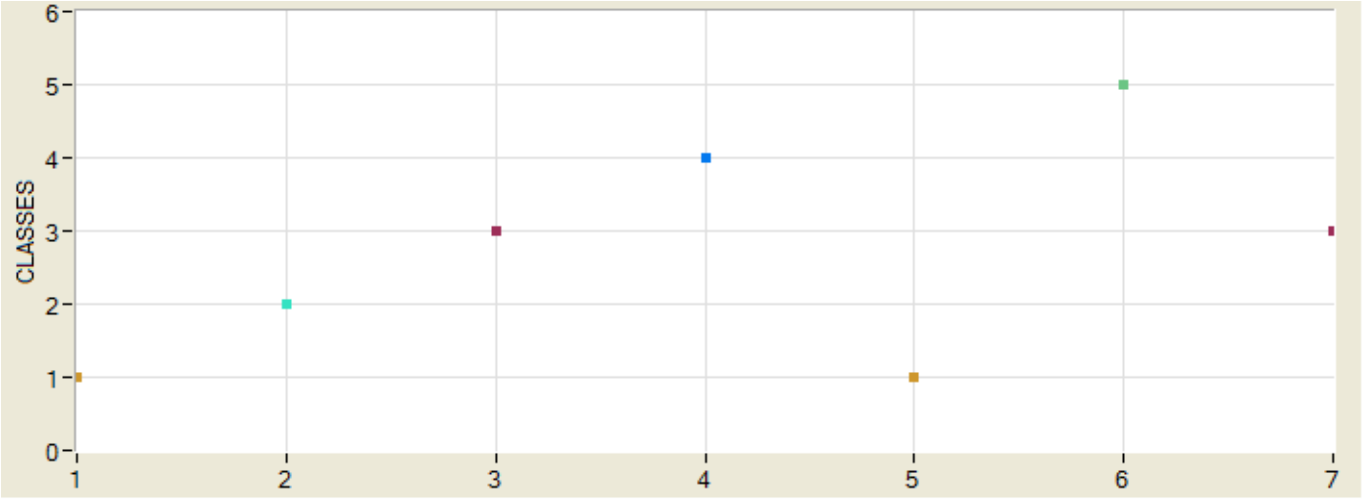


a. Répartition des individus dans les classes

	Forme	Couleur
Classe 1	triangle: 1 / triangle-rectangle: 1 / triangle-isocele: 1	Rouge : 3
Classe 2	triangle-isocele: 1	Bleu : 1
Classe 3	cercle : 1	Rouge : 1
Classe 4	rectangle: 2	Bleu : 2

b. Description des classes

Figure A.3. Résultats de la classification des individus décrits par leur forme et leur couleur avec prise en considération de la proximité des modalités

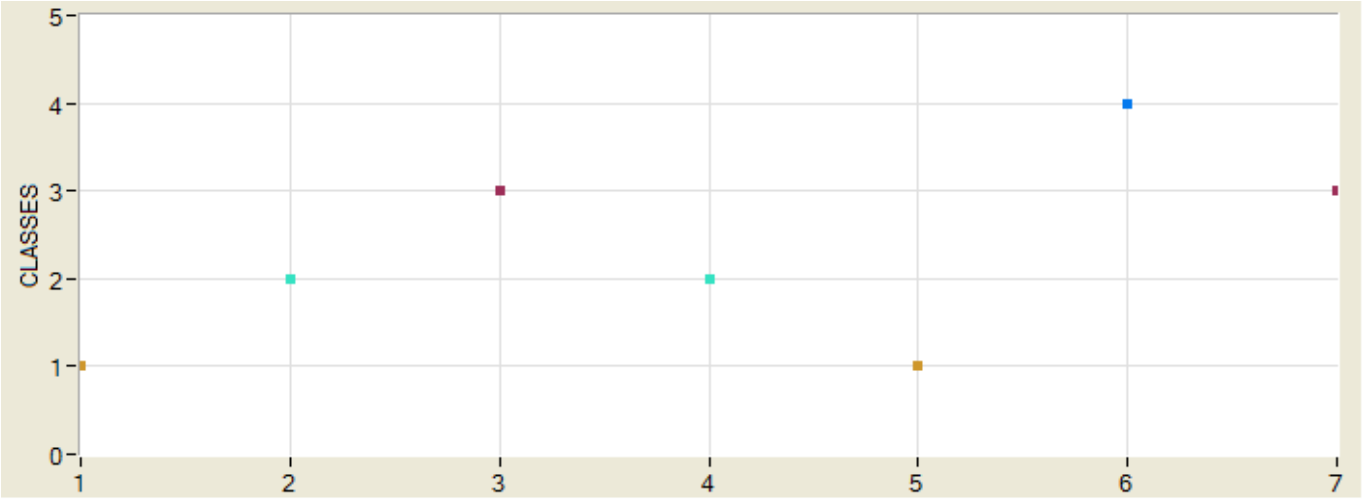


a. Répartition des individus dans les classes

	Poids	Sports	Sexe
Classe 1	79.5	Rugby-Tennis: 2	Homme : 2
Classe 2	72	Marathon-PingPong: 1	Homme : 1
Classe 3	67	Tennis : 2	Femme : 2
Classe 4	88	Marathon-PingPong-Rugby: 1	Homme : 1
Classe 5	91	Marathon-Tennis: 1	Homme : 1

b. Description des classes

Figure A.4. Résultats de la classification des individus décrits par leur poids, le sport qu'ils pratiquent, et leur sexe sans prise en considération de la proximité des modalités



a. Répartition des individus dans les classes

	Poids	Sports	Sexe
Classe 1	79.5	Rugby-Tennis: 2	Homme : 2
Classe 2	80	Marathon-PingPong: 1 / Marathon-PingPong-Rugby: 1	Homme : 2
Classe 3	67	Tennis : 2	Femme : 2
Classe 4	91	Marathon-Tennis: 1	Homme : 1

b. Description des classes

Figure A.5. Résultats de la classification des individus décrits par leur poids, le sport qu'ils pratiquent, et leur sexe avec prise en considération de la proximité des modalités

A.4 Annexe IV : Données relatives au projet MAISEO

A.4.1 Annexe IV.1 : Conception du système de classification

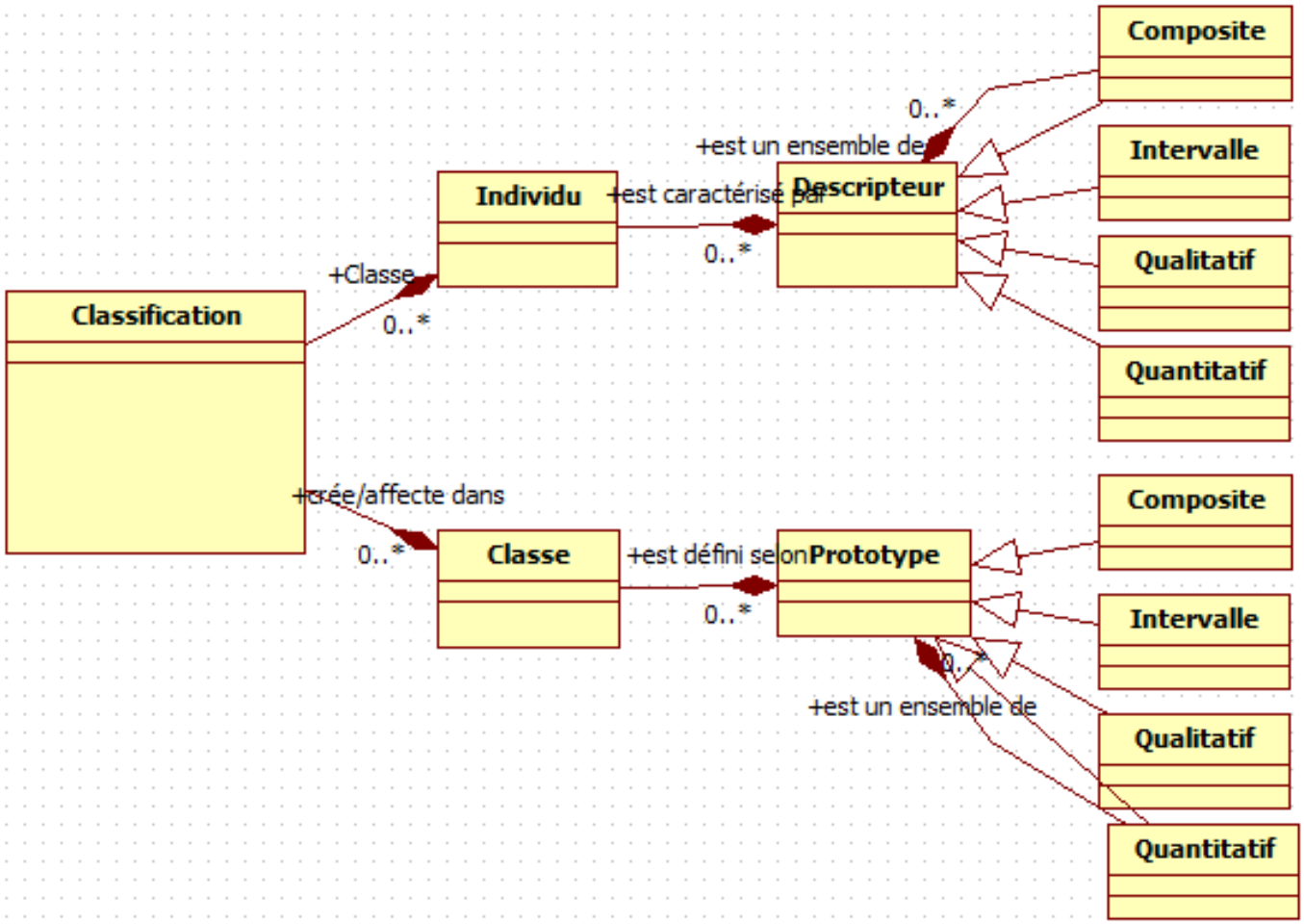


Diagramme 1. Diagramme de classes du système

A.4.2 Annexe IV.2 : Partie concernant le conseil aux exploitants

	Rendement	Profondeur RU	Type Sol	Irrigation	Précocité grain	Date Semis
Classe 1	121.7087	160.86	Argile-Limon:16/ Limon: 4	non: 20	Semi-tardif	10-mai
Classe 2	123.7222	180	Argile-Sable:8 /Argile: 2	oui:10	Tardif	05-mai
Classe 3	106.125	241.25	Argile-Sable:2	non:2	Semi-précoce	15-mai
Classe 4	130.65	128	Argile-limon:9	non:9	Très tardif	05-avr
Classe 5	124.1	180	Argile:2	non:2	Tardif	20-avr
Classe 6	132.1667	180	Limon:2/ Argile-Limon: 5	oui:7	Semi-tardif	20-avr

Météo										
Pluie hiver	Pluie Sem	Pluie Flo1	Pluie Flo2	Pluie Flo3	Pluie Rempli.	Temp CC	Temp Min	ETP Flo1	ETP Flo2	ETP Flo3
277.24	205.9	22.46	19.86	18.69	159.01	2324.1	5.9	70.7	74.6	68.57
305.8	244.2	25.3	21.09	23.93	170.32	2187.03	5.5	69.48	67.3	63.2
276.67	204.92	22.17	19.79	18.65	158.59	2325.01	6	70.77	74.7	68.59
302.4	229.5	24.2	21.01	21.43	168.5	2228.25	5.8	69.67	69.5	64.3
281.12	214.3	23.1	20.02	18.99	161.7	2298.8	5.9	70.4	71.5	66.42
304.8	239.9	25.2	21.01	23.81	170.01	2189.21	5.6	69.15	67.6	63.5

Tableau A.7. Définition des classes décrivant les différents profils de parcelles en classification non supervisée non contrainte.

Type de sol	Aspect	Avantages	Inconvénients
Argileux	Compact, collant lorsqu'il est humide, très dur et fendillé lorsqu'il est sec.	Retenant bien l'humidité et les minéraux. Ce type de sol peut être productif s'il est correctement enrichi en éléments nutritifs.	Difficile à travailler et s'engorge vite lors de fortes pluies. Compact, il empêche une bonne circulation de l'eau et de l'air, un enracinement profond. Ce type de sol se réchauffe lentement au printemps, occasionnant un retard de la végétation.
Limoneux	Doux au toucher, poudreux lorsqu'il sèche.	Très fertile, il est facile à travailler, propice au bon développement des plantes.	Fragile, il a tendance à former une croûte sous l'effet de la pluie et des arrosages.
Sableux	Granuleux au toucher, terre sans cohésion.	Très perméable à l'eau et à l'air, ce type de sol est facile à travailler. Il se draine naturellement grâce à sa texture poreuse. Il ne s'engorge jamais et se réchauffe facilement.	Très filtrant, il retient peu l'eau et peu les éléments nutritifs. Dépourvu de matière organique, il est facilement lessivé lors de l'arrosage ou des pluies. Il doit donc être fréquemment amendé pour rester fertile.
Calcaire	Blanchâtre d'aspect crayeux, terre souvent légère.	Perméable à l'eau, il se réchauffe rapidement	Peut bloquer certains éléments fertilisants qui deviennent alors non disponibles pour les plantes. Ce type de sol doit être fréquemment amendé. Sec en été, il est facilement boueux en cas de pluie.

Tableau A.8. Caractéristiques des types de sol les plus fréquents

Rendement	TypeSol	RU	Irrigation	DateSemis	Precocite
126	limon	190	non	5	1
135	limon	190	non	5	2
120	sable-argile	200	oui	5	2
120	sable-argile	200	oui	5	2
103.5	limon-argile	250	non	5	3
98	limon-argile	250	non	5	3
123	limon-argile	180	oui	4	1
125	limon-argile	250	oui	5	1
115	sable-argile	110	oui	6	2
112	limon-argile	180	oui	4	1
115	limon-argile	250	oui	4	2
111	sable-argile	110	oui	6	3
120	limon	190	non	5	2
107	limon	190	non	5	2
126	sable-argile	200	oui	6	2
126	sable-argile	200	oui	5	2
100	limon-argile	250	non	6	3
90	limon-argile	250	non	5	3
116	sable-argile	150	oui	4	3
128	limon-argile	110	oui	6	3
116.7	sable-argile	110	oui	4	3

PluieHiv_Q20	PluieSem_Q20	PluieFlo_Q20	PluieFlo2_Q20	PluieFlo3_Q20	PluieRempQ20	STM6_Q20	TSemQ20	ETPFlo1_Q20	ETPFlo2_Q20	ETPFlo3_Q20	PluieHiv_Q80	PluieSem_Q80	PluieFlo_Q80	PluieFlo2_Q80	PluieFlo3_Q80	PluieRempQ80	STM6_Q80	TSemQ80	ETPFlo1_Q80	ETPFlo2_Q80	ETPFlo3_Q80
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
344.4	267.4	32.6	23.8	26.5	211.6	2251	5.8	59.4	66.7	61.1	599	428	67.2	71.5	77.3	320	2549	7.4	78.1	81	71.1
268	222.7	19.7	18.2	21.4	137.2	2126	5.1	68.3	67.8	64	444	372.5	60.3	56.9	56	225.1	2370	6.5	82.5	89.2	78.8
328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
328.6	208.7	19	21.7	17.8	167	2360	6.1	69.5	73.4	67.7	519.5	348.3	58.4	51.1	57.8	272.1	2615	7.5	85.5	91	81
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
344.4	267.4	32.6	23.8	26.5	211.6	2251	5.8	59.4	66.7	61.1	599	428	67.2	71.5	77.3	320	2549	7.4	78.1	81	71.1
268	222.7	19.7	18.2	21.4	137.2	2126	5.1	68.3	67.8	64	444	372.5	60.3	56.9	56	225.1	2370	6.5	82.5	89.2	78.8
195.3	161.2	12.2	14.6	13.3	113.3	2302	5.5	74.1	77.5	73.3	338.6	277.7	51.2	54.2	52.9	202.3	2553	7	90.3	97.6	89.1
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1
271.8	184.7	20	17.1	12.8	134.6	2403	6.3	74.8	76.5	70.9	394.8	285.6	54.6	47.3	54.7	231.4	2631	7.6	88.2	96.5	87.1

Tableau A.9. Données d'apprentissage

A.4.3 Annexe IV.3 : Partie concernant le conseil au gestionnaire du bassin versant

PROXIMIT(0,1)_	DELTAALT	PENTEMOY	ORIENTATION	PROXIMIT_1	type_inra	DIFALT_COL
1082.062036	30	2.089456	SE	2322.121659	barrage	-13.9
106.040256	2	1.415835	N	4220.562116	barrage	-61.9
2486.665802	118	7.40021	S	3061.627134	barrage	76.1
97.251357	2	1.400913	SE	2368.468334	barrage	-44.9
903.927213	17	1.413654	NE	4423.875212	barrage	-40.9
404.192957	9	1.64018	NO	3230.376297	barrage	-42.9
740.666547	14	1.753856	NO	4820.765604	barrage	-49.9
193.954805	4	1.630815	N	3545.809433	barrage	-48.9
1142.02593	27	2.688593	S	3812.030073	barrage	-25.9
713.779884	15	1.427164	S	2123.686252	barrage	-26.9
1160.45009	33	3.29111	NE	4869.769531	barrage	-30.9
1256.177022	32	3.77097	S	3980.773508	barrage	-20.9
87.091514	-1	2.599955	SO	3786.367505	barrage	-59.9
939.592486	20	1.475514	NO	4663.447687	barrage	-41.9
665.874619	15	1.046384	N	2185.410669	barrage	-28.9
307.498931	7	1.315153	O	3840.366035	barrage	-45.9
310.463811	7	1.26048	SE	3986.138363	barrage	-46.9
760.495588	19	1.643868	S	3276.547498	barrage	-32.9
819.948689	21	1.784723	E	4873.398844	barrage	-42.9
1887.311594	48	4.141391	O	2408.376602	barrage	6.1
1075.265801	23	4.03608	N	2749.359141	barrage	-23.9
314.046444	7	1.31146	N	3901.346278	barrage	-45.9
919.882735	19	1.687819	SE	3861.362387	barrage	-33.9
1695.054933	38	3.714537	S	2618.525328	barrage	4
185.988988	1	3.5267	S	31262.88156	barrage	-5.11
146.600889	0	1.178466	SE	31911.06118	barrage	-6.11
618.318297	22	3.859167	N	32174.43134	barrage	15.89

Tableau A.10. Extrait du tableau répertoriant les informations disponibles pour la classification des individus (Source CACG)

B Développement d'un outil d'aide au diagnostic pour la production de maïs permettant la réduction de la consommation en eaux d'irrigation et en traitements phytosanitaires

B.1 Résumé

La présente thèse concerne la conception d'un outil d'aide à la décision, et s'intéresse tout particulièrement aux aspects relatifs à l'apprentissage et au diagnostic. Le premier objectif est de le rendre capable de choisir le meilleur scénario pour le couple « date de semis »/ « type de semis » en fonction des caractéristiques des parcelles à diagnostiquer dans un objectif d'aide à la décision pour les agriculteurs. Le second concerne le diagnostic hydrique établi avant le début des irrigations en juin, au cours duquel il s'agit d'évaluer les besoins en eau d'irrigation des différents îlots de parcelles afin d'anticiper la demande et de répartir au mieux les quantités d'eau pourvue aux irrigants. L'apprentissage utilise la méthode de classification LAMDA qui est basée sur la logique floue. Afin de permettre la prise en compte optimale de tous les facteurs pouvant intervenir dans le rendement d'une parcelle, un nouveau type a été intégré à la méthode LAMDA, et les outils d'apprentissage ont été modélisés, implémentés, et testés de sorte à correspondre aux besoins spécifiques des deux parties du projet. Le nouveau type a été conçu pour être générique et permettre à la méthode LAMDA un traitement multicouche des données d'apprentissage. Il autorise ainsi la gestion de données multidimensionnelles issues de contextes différents. Son efficacité a été évaluée sur les cas pratiques du projet MAISEO mais a vocation à pouvoir être appliqué à tous les autres domaines de recherche dans lesquels la classification multivariée peut être employée.

B.2 Mots Clés

– Aide à la décision,
– Apprentissage,
– Diagnostic,
– Logique Floue,
– Développement logiciel.

C Development of a diagnosis support tool for the corn production to reduce the irrigation water consuption and the quantity of phytosanitary treatments used

C.1 Summary

My thesis deals with the conception of a decision support tool, and particularly focuses on the aspects relative to machine learning and diagnosis. The first goal is to make it capable of selecting the best scenario for the couple « date of sowing »/ « type of seed » taking in account the plots to diagnose caracteristics with the aim of guiding the corn farmers practices. The second is goal is about the water diagnosis, set up in june, before the irrigation is started. This diagnosis aims at evaluating the water needs of large plots areas, with the purpose to anticipate demand and better allocate the water quantity provided to the farmers. The LAMDA method has been chosen to realize the machine learning ; this is a fuzzy logic based classification method. A new type of data has been integrated to the method to ensure that the main factors that influence a plot yield are fully taken into account. In this process, the machine learning tools have been modelled, implemented, and tested in the order to correspond to the specific needs of both parts of the project. The new type of data has been designed to be generic and allows a multilayer clustering to diagnose the complex systems. Multidimensional data coming from various contexts are so able to be manage by the LAMDA method. The efficiency of this technique has been assessed on the practical cases of the MAISEO project, but is intented to be apply to every research field in which the multivariate classification is used.

C.2 Key Words

– Decision support,
– Machine learning,
– Diagnosis,
– Fuzzy logic,
– Software development.

D Remerciements

Je tiens tout d'abord à témoigner toute ma reconnaissance à ma directrice de thèse, Marie-Véronique Le Lann, qui a su, au long de ces trois années, manifester une aide à la fois rassurante et discrète, disponible et circonspecte, attentive et solide, avisée et communicative. Je la remercie de m'avoir accordé sa confiance, de m'avoir témoigné un soutien indéfectible malgré les tracas que les problèmes administratifs pouvaient lui causer, et d'avoir accepté mon rythme de travail. Je la remercie parce que si j'ai pu accomplir sur ma thèse dans de si bonnes conditions, et si je finis de l'écrire aujourd'hui, c'est surtout grâce à elle.
Je tiens également à remercier le directeur du LAAS-CNRS, monsieur Jean Arlat, pour m'avoir reçue dans son laboratoire et avoir chaque année apposé sa signature dans la place qui lui était réservée sur ma fiche de réinscription.
Je remercie également madame Louise Trave-Massuyes, responsable du groupe DISCO, pour son accueil dans l'équipe. Merci par ailleurs à tout les membres du groupe DISCO pour leurs conseils et leur sourire, et en particulier à Thomas Monrousseau, qui a su décorer notre bureau avec autant de goût.
Je tiens également à manifester ma reconnaissance à madame Nathalie Perrot et monsieur Rafaël Gouriveau, qui m'ont fait l'honneur d'accepter d'être les rapporteurs de ma thèse ; j'espère de tout coeur qu'ils trouveront de l'agrément à cette tâche.

Au cours de ma thèse, j'ai également pu effectuer des enseignements ; en vertu du plaisir que j'y ai trouvé, je tiens à remercier monsieur Daniel Marre pour la confiance qu'il m'a accordée tout au long de ces trois années, son enthousiasme, et sa conversation.

Je remercie également ma mère, Nadine Roux, sans qui je n'aurais pas pu faire grand chose, il faut bien l'avouer, et qui a toujours été une oreille attentive et m'a toujours manifesté un soutien inébranlable. Merci à Alain Roux, mon grand frère, pour son écoute et ses discussions tellement intéressantes. Merci à Isabelle Roux-Gregson, ma grande soeur, pour sa bonne humeur rayonnante. Merci à Steve et Charlie, qui font maintenant partie de la famille. Merci à mes deux grands-mère, qui ne comprennent tout de même pas très bien ce que je fais, mais qui sont là, et c'est bien le plus important. Merci à Bribri, Toto, Amélie, leur affection et leur réactivité.

Je remercie Sébastien Chane-Sam, mon époux depuis le 26 septembre 2015, qui a su me témoigner un amour soutenu et perpétuel, et supporter mes weekends de labeur, ainsi que ma mauvaise humeur les soirs où "oh lala, j'ai codé toute la journée, je n'en peux plus !", et qui a bien voulu endurer les huit saisons de 24h chrono et l'infatigable mais épuisant Jack Bauer lorsque je voulais juste "poser mon cerveau quelque part et ne plus avoir à réfléchir, par pitiéé".

Merci à Jean-Luc et Valérie Chane-Sam pour m'avoir déchargée d'une partie de l'organisation du mariage, et m'avoir ainsi fait gagner un temps précieux pour ma thèse.

Je remercie également tous mes anciens camarades, mais toujours amis, grâce à qui j'ai pu accomplir mes études dans la bonne humeur, et qui se sont toujours montrés disponibles : Maxime Shubin, tou-

jours prêt à m'accompagner petit déjeuner au Capitole en pyjama, Erwan Abolivier, qui sait si bien nous faire rire avec un simple stylo, Matthieu Dubet pour le vif intérêt qu'il a toujours porté à ma thèse (et à ma vie en général, d'ailleurs, hein), Guillaume Verdier qui sait répondre à toutes nos questions. Toujours.

Merci enfin à tous mes amis : Célia pour l'ami-versaire de nos 20 ans que nous allons bientôt fêter, Julia pour toutes nos pérégrinations à l'autre bout du monde, Olivia pour notre amitié bisounours, Gwen pour toutes ces années qui passent et qui n'étiolent rien - pas même les bulles de notre boisson favorite en soirée -, Jérémy qui m'a fait aimer les cours de maths à coups de mots croisés et m'accompagne toujours, de loin, au fil de ces années, Loïc pour ses messages si drôles le goût de bonbons colorés qu'il a donné à mes cours de C, Marion pour sa présence et son écoute, qu'on soit voisines ou à-trois-stations-de-métro-dont-un-changement-quand-même, Héléna pour nos longues discussions et nos churros au chocolat en rentrant du Mirail, Laura pour sa gaité si contagieuse, Alice et Marie, pour tous nos goûthés de pipelettes gourmandes, Gaëlle, pour tout ce qu'elle s'est impliquée dans son rôle de demoiselle, Rémy pour avoir été la moitié du 222, Valou pour organiser les meilleures murders parties de Toulouse et être toujours prêt à se déguiser en personnage de Disney quand vient le marché de Noël, Adrien pour avoir été un super maître de stage, Denis pour ses conseils en c++ et son goût immodéré pour les raclettes, Chérina pour nos pauses du midi au chinois à volonté le plus proche (ou du moins, le plus proche et proposant une fondue au chocolat pour le dessert !), Emy pour sa gaité imperturbable, Raph pour son amitié toujours aussi présente malgré la distance, Marie-Line et la source d'inspiration intarissable que sont ses mille et une recettes, Julian pour son calme et son amour du chocolat (et parce qu'il me permet de parler deux fois de chocolat dans le même paragraphe !).

Merci à Laetitia et sa maman, mes deux marchandes de fruits et légumes préférées, qui ont pourvu à mes besoins en vitamines tout au long de cette thèse et ont, ainsi, participé activement à son avancement - tout en manifestant à son égard un intérêt constant et enthousiaste.

Merci à Alexandre Astier qui m'a régulièrement fait rire durant ces trois ans et a sorti son DVD de l'Exoconférence pile au moment de mon anniversaire ; je ne devrais pas avoir à l'dire, ça. Merci aux livreurs de sushis qui se sont dévoués et ont affronté les intempéries pour assurer mon ravitaillement dans les moments les plus critiques.

Merci à tous, vous qui m'avez soutenue et ne m'avez jamais rien reproché, lorsque je ne pouvais pas sortir parce que "oh je suis vraiment désolée, mais je suis complètement overbookée en ce moment". Parce que c'est aussi grâce à vous que je finis ma thèse aujourd'hui, et que j'ai pu parallèlement poursuivre mes études de psycho et de philo - parce que vous m'épaulez alors même que vous ne suivez plus très bien ce que je fais parfois. Merci de vous être autant souciés de la progression de ma thèse, et de m'avoir témoigné cet intérêt par l'éternelle - et pas toujours apaisante - question : "et sinon ta thèse, ç'avance ?". Merci, parce que c'est un peu par vous tous que tout ça a pris un sens.

Bibliographie

Construil, l. and rivest, rl. *Information Proceding Letters*, pages 5(1) :15–17, 1976.

D.H. ACKLEY, G. E. HINTON et T.J. SEJNOWSKI : A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

AGUADO et AGUILAR-MARTIN : A mixed qualitative-quantitative self-learning classification technique applied to diagnosis. *QR*, 99:124–128, 1999.

J. AGUILAR-MARTIN et R. López de MÁNTARAS : The process of classification and learning the meaning of linguistic descriptors of concepts. *Approximate Reasoning in Decision Analysis*, 1982:165–175, 1982.

D.W. AHA : Incremental, instance-based learning of independent and graded concept descriptions. *In Proceedings of the sixth international workshop on Machine learning*, pages 387–391. Morgan Kaufmann Publishers Inc., 1989.

V. BADEAU et E. ULRICH : Renecofor - etude critique de faisabilité sur : la comparabilité des données météorologiques « renecofor » avec celles de météo france, l'estimation de la réserve utile en eau du sol et le calcul des volumes d'eau drainée en vue du calcul de bilans minéraux sur les placettes du sous-réseau cataenat. pages 108–166, 2008.

G.H. BALL et D.J. HALL : Isodata, a novel method of data analysis and pattern classification. Rapport technique, DTIC Document, 1965.

Y BENGIO, IJ GOODFELLOW et A COURVILLE : Deep learning. URL `http://www.iro.umontreal.ca/ bengioy/dlbook`. Book in preparation for MIT Press, 2015.

JC. BEZDEK : A convergence theorem for the fuzzy isodata clustering algorithms. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):1–8, 1980.

JC. BEZDEK : *Pattern recognition with fuzzy objective function algorithms*. Springer Science and Business Media, 2013.

JC. BEZDEK, RJ. HATHAWAY, MJ. SABIN et WT. TUCKER : Convergence theory for fuzzy c-means : counterexamples and repairs. *Systems, Man and Cybernetics, IEEE Transactions on*, 17(5):873–877, 1987.

RJ. BOUCHET : Évaportranspiration réelle et potentielle : signification climatique. Rapport technique, Station centrale de Bioclimatologie (Versailles) - Institut national de la Recherche agronomique (France).

A. BOUTHIERT : Mesurer : Connaître la réserve utile de ses sols pour mieux évaluer ses besoins en eau. *PERSPECTIVES AGRICOLES*, (399), 2013.

L. BREIMAN : Bagging predictors. *Machine learning*, 24:123–140, 1996.

L. BREIMAN, J. FRIEDMAN, CJ. STONE et RA. OLSHEN : *Classification and regression trees*. CRC press, 1984.

BRUCKERT : D´esignation et classement des sols agricoles d'apres des criteres de situation et d'organisation : application aux terres franc-comtoises du domaine climatique tempere semicontinental. *Agronomie, EDP Sciences,*, (9(4)):353–361, 1989.

H. BUHRMAN et R. DE WOLF : Complexity measures and decision tree complexity : A survey. *Theoritical Computer Science*, pages 288(1) :21–43, 2002.

A. COSSETTE : *Travaux de linguistique quantitative*, volume 53. H. Champion, 1994.

AA. COURNOT : Essai sur les fondements de nos connaissances, t. 1. *Paris1851*, 1851.

TM. COVER et PE. HART : Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

M. de ARIZA, GALINDO et J. AGUILAR-MARTIN : Clasificación de la personalidad y sus trastornos, con la herramienta lamda de inteligencia artificial en una muestra de personas de origen hispano que viven en toulouse-francia. *Revista de estudios sociales*, 18:99–110, 2004.

C. DESBOURDES, F. PIRAUX et B. DE SOLAN : Hétérogénéités du sol, comment les déceler ? *PERSPECTIVES AGRICOLES*, 338, 2007.

P. DESROCHES : Syclare : Systeme de classification avec apprentissage et reconnaissance de formes. *Manuel d'utilisation. Rapport de recherche, Centre d'estudis avançats de Blanes, Espagne*, 1987.

D. DUBOIS et H. PRADE : The three semantics of fuzzy sets. *Fuzzy sets and systems*, 90(2):141–150, 1997.

RO. DUDA, PE. HART et DG. STORK : *Pattern classification*. John Wiley & Sons, 2012.

J.C. DUNN : A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

Simonne et AL. : Principles and practices of irrigation management for vegetables. *University of Florida - IFAS Extension*, 3:17–27, 2011.

E. FIX, Jr. HODGES et L. JOSEPH : Discriminatory analysis-nonparametric discrimination : consistency properties. Rapport technique, DTIC Document, 1951.

J. FUHRER : Estimation des besoins en irrigation pour l'agriculture suisse. *Station de recherche Agroscope Reckenholz-Tänikon (ART)*, 2010.

J. FUHRER et K. JASPER : Besoins en irrigation en suisse. rapport final du projet bi-ch. Rapport technique, Station de recherche Agroscope Reckenholz-Tännikon (ART), 2009.

Ph. GUYOT et M. BORNAND : Cartes départementales des terres agricoles, intégration des données sols et des données économiques. *Science du sol*, 25/1:1–16, 1987.

MA. HALL : *Correlation-based feature selection for machine learning*. Thèse de doctorat, The University of Waikato, 1999.

Bezdek HATHAWAY, Richard : Recent convergence results for the fuzzy c-means clustering algorithms. *Journal of Classification*, 5(2):237–247, 1988.

Tucker HATHAWAY, Bezdek : An improved convergence theorem for the fuzzy c-means clustering algorithms. *The Analysis of Fuzzy Information*, 3(8), 1987.

i L. HEDJAZ, J. AGUILAR-MARTIN, MV. LE LANN et Tatiana KEMPOWSKY : Towards a unified principle for reasoning about heterogeneous data : a fuzzy logic framework. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20(02):281–302, 2012.

L. HEDJAZI : *Outil d'aide au diagnostic du cancer à partir d'extraction d'informations issues de bases de données et d'analyses par biopuces*. Thèse de doctorat, Université Paul Sabatier-Toulouse III, 2011.

L. HEDJAZI, T. KEMPOWSKY-HAMON, L. DESPÈNES, MV. LE LANN, S. ELGUE et J. AGUILAR-MARTIN : Sensor placement and fault detection using an efficient fuzzy feature selection approach. *In Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 6827–6832. IEEE, 2010.

Teh HINTON, Osindero : A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

K. HORNIK, M. STINCHCOMBE et H. WHITE : Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Q. HU, Z. XIE et D. YU : Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. *Pattern recognition*, 40(12):3509–3521, 2007.

R. HUNGER : Arrosage efficace – mot d'ordre actuel. *Technique agricole*, juin/juillet 2010.

P. JACCARD : *Nouvelles recherches sur la distribution florale*. 1908.

L. KAUFMAN et PJ. ROUSSEEUW : *Finding groups in data : an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

T. KEMPOWSKY, J. AGUILAR-MARTIN, A. SUBIAS et MV. LE LANN : Classification tool based on interactivity between expertise and self-learning techniques. *In 5th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, Safeprocess*, 2003.

K.i KIRA et LA. RENDELL : A practical approach to feature selection. *In Proceedings of the ninth international workshop on Machine learning*, pages 249–256, 1992.

S. KIRKPATRICK, CD. GELATT, MP. VECCHI *et al.* : Optimization by simulated annealing. *science*, 220 (4598):671–680, 1983.

T. KOHONEN : *Self-organizing maps*, volume 30. Springer Science and Business Media, 2001.

Teuvo KOHONEN : Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

BR. KOWALSKI, S. WOLD, PR. KRISHNAIAH et LN. KANAL : Classification, pattern recognition and reduction of dimensionality. *Pattern Recognition in Chemistry,(Edited by PR Krishnaiah and L, N, Kanal). North-Holland, Amsterdam*, 1982.

GW LEIBNIZ : Dissertio de arte combinatoria. *Leipzig (1666)*.

Pitts MCCULLOCH, Warren : A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

RS. MICHALSKI et RE. STEPP : Automated construction of classifications : Conceptual clustering versus numerical taxonomy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (4):396–410, 1983.

T. MOHRI et H. TANAKA : An optimal weighting criterion of case indexing for both numeric and symbolic attributes. *In AAAI-94 Workshop Program : Case-Based Reasoning, Working Notes*, pages 123–127, 1994.

MORGENTHALER : *Génétique statistique - Statistique et probabilités appliquées*. Springer Science and Business Media, 2008.

CVI NARVAEZ : *Diagnostic par techniques d'apprentissage floues : concept d'une méthode de validation et d'optimisation des partitions*. Thèse de doctorat, INSA de Toulouse, 2007.

S. PEPIN et G. BOURGEOIS : Outils agrométéorologiques pour la planifiation de l'irrigation des cultures. *Colloques Agroclimatologie*, 1992.

N. PIERA, Ph. DESROCHES et J. AGUILAR-MARTIN : Lamda : An incremental conceptual clustering method. *Rapport technique LAAS-CNRS*, 89420, 1989.

Ricco RAKOTOMALALA : Pratique de la régression logistique. *Régression Logistique Binaire et Polytomique, Université Lumière Lyon*, 2, 2011.

L. RENOULT, B. DE SOLAN et N. BOUSQUET : Les systèmes d'information géographique : Cartographier son exploitation. *PERSPECTIVES AGRICOLES*, 322, 2006.

F. ROSENBLATT : Principles of neurodynamics. 1962.

E. ROUX, L. TRAVE-MASSUYES et MV. LE LANN : Applied multi-layer clustering to the diagnosis of complex agro-systemss. *Proceedings of the 26th International Workshop on Principles of Diagnosisl*, pages 19–25, 2015.

D. RUMELHART, GE. HINTON et RJ. WILLIAMS : Parallel distributed processing. 1986.

EH. RUSPINI : A new approach to clustering. *Information and control*, 15(1):22–32, 1969.

VAPNIK et NAUMOVICH : *Statistical learning theory*, volume 1. Wiley New York, 1998.

J. WAISSMAN, J. AGUILAR-MARTIN, B. DAHHOU et G. ROUX : Généralisation du degré d 'adéquation marginale de la méthode de classification lamda. *6èmes rencontres de la Société Francophone de Classification., Montpellier France*, 1998.

LA. ZADEH : Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy sets and systems*, 90(2):111–127, 1997.

H. ZHANG : The optimality of naive bayes. *AA*, 1(2):3, 2004.