1. QU'EST-CE QUE LA CLASSIFICATION ?

Résumé

Nous présentons ce qu'est un problème de classification. Dans certains cas, il est possible de décrire complètement, de manière linguistique, la démarche de classification ; dans ce cas, un algorithme reproduisant cette démarche peut être construit, et le problème est résolu. Dans d'autres cas, il est impossible de décrire précisément la classification ; une solution consiste alors à demander à un professeur (ou superviseur, expert) de classer un échantillon d'individus. Des méthodes de résolution qui "apprennent par l'exemple" (ici un exemple est un individu déjà classé par le superviseur) sont capables de reproduire la classification de l'expert et, ensuite, de classer automatiquement de nouveaux exemples inconnus. Ces dernières méthodes sont donc essentiellement statistiques ; c'est à elles que nous nous intéresserons dans ce mémoire.

Nous posons également une question très importante en classification, et, plus généralement, dans tout problème de modélisation statistique : celle du choix des variables descriptives pertinentes (dont la connaissance est susceptible de contribuer utilement à la solution du problème posé) parmi un ensemble de variables descriptives possibles.

1.1 Introduction

Classifier des formes ou individus (par exemple des objets, des images, des phonèmes, ...) décrits par un ensemble de grandeurs caractéristiques (taille ou masse de l'objet, pixels de l'image numérisée, spectre acoustique du phonème, ...), c'est les ranger en un certain nombre de catégories ou classes définies à l'avance[1].

Citons quelques exemples de classification :

- Un exemple courant d'application de la classification est le tri automatique du courrier par un dispositif de lecture et d'interprétation du code postal ou de l'adresse manuscrite. Pour un dispositif d'interprétation du code postal, 10 *classes* sont possibles (les chiffres de 0 à 9) et les *variables descriptives* peuvent être les niveaux de gris des pixels, provenant d'une image numérisée du *chiffre* à identifier.
- Un établissement bancaire est fréquemment appelé à répondre à la demande de prêt d'un client, sur la base de quelques indicateurs décrivant sa capacité à rembourser. Dans ce cas, les *individus* à classer sont des personnes, et les *variables descriptives* sont, par exemple, le salaire, l'âge, la situation de famille, le nombre d'enfants... Nous pouvons imaginer plusieurs *classes* suivant le type de risque que peut admettre l'établissement.

---

[1] Dans ce contexte, les statisticiens utilisent fréquemment le terme de *discrimination*, et réservent le terme de classification au cas où les classes ne sont pas définies à l'avance ; nous conserverons néanmoins ici la dénomination consacrée par l'usage dans le domaine des réseaux de neurones.

- Pour un système de sécurité, le dispositif doit repérer au bon moment une situation inquiétante parmi la masse de situations normales et déclencher l'alerte. Dans le cas d'un dispositif de surveillance d'un réacteur chimique, les *individus* sont les états du processus au cours du temps et les *variables descriptives* sont, par exemple, la température, le débit, le pH ... Il y a deux *classes* possibles (situation normale et situation anormale).

Dans le présent travail, les objets à classer sont des entreprises ou des collectivités locales, les variables sont les données financières ou socio-économiques attachées aux objets et les classes sont les évaluations fournies par des analystes financiers. Bien que ce problème ne soit pas, à proprement parler, un problème de reconnaissance de formes, nous allons indiquer dans le paragraphe suivant, à titre d'illustration, comment la classification s'insère dans un système de reconnaissance de formes.

1.2 Chaîne de reconnaissance de formes

Un dispositif de reconnaissance automatique de formes est généralement conçu comme une chaîne de modules de traitement [voir par exemple Price 96]. Ainsi, un système de reconnaissance de formes comporte habituellement :

- un module d'acquisition : des capteurs mesurent des grandeurs caractéristiques de l'objet à classer. Cet ensemble de grandeurs constitue la première représentation de l'objet.
- un module de pré-traitement : il peut être judicieux de modifier les grandeurs brutes issues des capteurs par un algorithme afin de tenir compte des connaissances qui peuvent être disponibles *a priori* sur le problème. Par exemple, à partir de la réponse d'un capteur on peut appliquer un ensemble de filtres destinés à annuler les effets de taille ou de positionnement. Ainsi, on obtient une nouvelle représentation de l'objet, plus adéquate pour la classification envisagée.
- d'autres modules de traitement peuvent élaborer des représentations successives de l'objet ; ces différentes représentations ont généralement pour objectif de réduire la dimension de la représentation, c'est-à-dire de diminuer le nombre de descripteurs de l'objet, et d'élaborer des descripteurs de plus en plus pertinents pour la tâche de discrimination à accomplir.
- un module de classification : l'algorithme de classification considère la dernière représentation de l'objet et décide d'affecter celui-ci à une classe. Cet algorithme peut fournir soit une réponse binaire à valeurs discrètes (appartenance ou non à une classe) soit une réponse probabiliste à valeurs continues (l'image à 70% de chance de représenter le chiffre 5).

La figure 1.1 illustre une chaîne de classification comportant un seul module de pré-traitement. On distingue les trois modules et les représentations successives de l'objet. Naturellement, on peut imaginer un dispositif sans module de pré-traitement ; dans ce cas l'algorithme de classification travaille directement sur les grandeurs relevées par les capteurs.

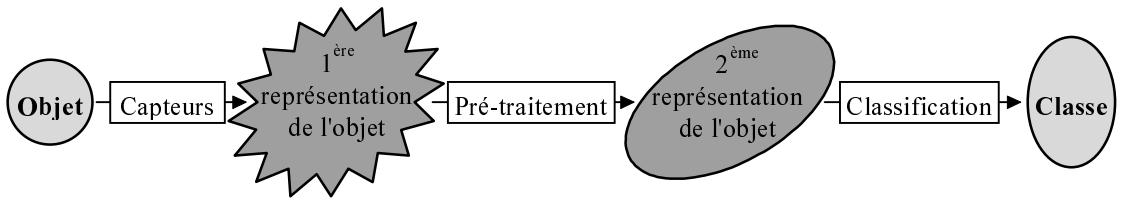


Figure 1.1 : Chaîne d'un dispositif de classification

La tâche de l'algorithme de classification est d'autant plus aisée que la représentation de l'objet est pertinente. Par exemple dans un problème de commande d'un processus chimique, on peut imaginer que la distinction entre les situations normales et les situations de danger est entièrement définie par la valeur de la pression. Si les modules d'acquisition ou de pré-traitement ne fournissent pas cette valeur à l'algorithme de classification, celui-ci ne pourra pas faire de miracle et distinguer les différentes situations.

1.3 Formalisation mathématique d'un problème de classification

Les exemples précédents font apparaître la classification comme une tâche qui consiste à ranger des formes ou ***individus*** décrits par un ensemble de ***variables descriptives*** en un certain nombre de catégories ou ***classes*** définies *a priori*.

Traduit en termes mathématiques, un problème de classification comporte les ingrédients suivants :

- une population de N individus I^i, (i variant de 1 à N),
- P variables descriptives X_d^i, qui permettent de décrire les individus ; elles sont aussi appelées plus simplement descripteurs (d variant de 1 à P),
- C classes C_k dans lesquelles on cherche à ranger les individus (k variant de 1 à C),

Résoudre un problème de classification, c'est trouver une application de l'ensemble des objets à classer, décrits par les variables descriptives choisies, dans l'ensemble des classes. L'algorithme ou la procédure qui réalise cette application est appelé *classifieur*.

Les variables descriptives considérées ici sont celles qui sont fournies à l'algorithme de classification. Comme indiqué plus haut, elles peuvent être le résultat d'un pré-traitement des variables initiales.

1.4 Un premier exemple

Nous trouvons un premier exemple de classification dans la vie de tous les jours : le rangement des pièces de monnaie. En effet, un commerçant doit, de temps à autre, rassembler les pièces identiques contenues dans sa caisse afin d'en faire des rouleaux qu'il remettra à la banque. Dans cet exemple, le fond de caisse du commerçant constitue la ***population*** concernée, chaque ***individu*** est une pièce de monnaie. Les ***classes*** sont au nombre de 9 :

Pièces de 5 cts	Pièces de 50 cts	Pièces de 5 F
Pièces de 10 cts	Pièces de 1 F	Pièces de 10 F
Pièces de 20 cts	Pièces de 2 F	Pièces de 20 F

On peut même imaginer une dixième classe pour les autres pièces (étrangères ou fausses !). Les variables descriptives sont nombreuses, on trouve par exemple :

Diamètre,
Épaisseur,
Poids,
Couleur(s),
Matériau (composition chimique),
Mots / chiffres / dessins en relief à la surface,
Surface de la tranche : tranche lisse ou dentelée et type de dents,
Bruit que fait la pièce en tombant,
Etc.

Ces descripteurs peuvent être considérés comme des grandeurs descriptives potentielles. Dans notre exemple, chacun d'entre eux est pertinent pour départager les pièces. Cependant, il n'est pas nécessaire de les utiliser tous. En réalité, les descripteurs dont peut se servir le commerçant sont la couleur et le diamètre (même s'il ne le mesure pas, mais l'évalue seulement). Dans ce cas, la relation *F* qui relie les variables descriptives à la classe est de la forme :

F (jaune, petit diamètre) = classe des 5 cts,
F (jaune, moyen) = classe des 10 cts,
...
F (jaune & blanc, grand) = classe des 20 F.

Notons toutefois qu'une personne aveugle n'utiliserait ni ces descripteurs ni cette fonction mais peut-être une fonction G telle que :

G (diamètre, épaisseur, surface de la tranche) = classe de pièce.

On voit donc que plusieurs règles de décision, toutes aussi pertinentes les unes que les autres, permettent de ranger cette population dans les classes désirées. Dans cet exemple, les fonctions peuvent être décrites explicitement (le commerçant ou l'aveugle peuvent expliquer comment ils procèdent) et elles conduisent au même résultat.

Lorsqu'une telle tâche doit être effectuée de manière répétitive, on est tenté de la confier à un automate (c'est d'ailleurs le cas dans les caisses automatiques de parkings, distributeurs de titres de transport, etc.). En effet, dès que les variables descriptives et la fonction peuvent être exprimées si simplement, une telle classification "mécanique" peut facilement être réalisée par un automate réalisant une suite d'opérations logiques (système expert). Celui-ci se fondera peut-être sur le gabarit, le poids ou la composition chimique des pièces, c'est-à-dire utilisera la fonction suivante :

H (poids, diamètre, composition chimique) = classe de pièce.

Malheureusement, les processus de classification ne sont pas toujours aussi simples et la règle de décision ne peut pas toujours être explicitée.

1.5 Deux ou trois étoiles dans le guide Michelin ? Les choses se compliquent...

D'autres tâches de classification, qui sont, elles, fondée sur l'intuition, sont susceptibles d'être automatisées. La notation des restaurants dans les guides touristiques est, par exemple, un problème de classification plus complexe. Il s'agit bien de classer n'importe quel restaurant dans l'une des quatre classes : aucune étoile, une étoile, deux étoiles ou trois étoiles. En essayant soi-même d'évaluer tous les restaurants où l'on a déjà mangé (en prenant quatre niveaux : *exceptionnel*, *satisfaisant*, *correct* et *à éviter*) puis d'expliquer sa propre classification, on constate plusieurs choses :

- il n'est pas toujours facile de faire la liste des éléments que l'on prend en considération (les variables descriptives),
- il est quasiment impossible de formaliser la règle de décision que l'on adopte, c'est-à-dire de décrire comment s'élabore notre jugement. Dans un cas, le sourire de la serveuse aura suffi à compenser la tiédeur du steak et la table sera classée "correcte" ; une valeur très positive de la variable "service" aura prédominé sur la piètre "qualité du repas". Dans une autre circonstance, un délicieux foie gras fera oublier qu'on l'a attendu trois quarts d'heure en contemplant des murs lépreux ; la variable "qualité du repas" l'a emporté sur les deux variables "service" et "cadre", etc.

Ainsi, la classification est souvent complexe dans les problèmes pour lesquels l'expert réagit en fonction de son intuition et ne peut pas toujours formaliser la fonction qu'il adopte. Pourtant, il peut être nécessaire de savoir reproduire la classification de l'expert. Par exemple, les chargés de clientèle d'une banque ne peuvent pas se contenter systématiquement d'une évaluation subjective et personnelle de la solvabilité d'un client qui leur demande un prêt. Or, donner un avis favorable ou défavorable à la demande du client revient à effectuer une classification des demandes en deux classes : celles que l'on accepte et celles que l'on refuse. L'image de marque de la banque et sa sécurité financière exigent que cette classification soit unifiée, dans toute la mesure du possible.

1.6 Vers une classification probabiliste

Dans les exemples précédents, la classe des individus est bien définie ; mais ce n'est pas toujours le cas. Considérons une autre tâche qui consiste à discriminer les femmes des hommes à partir du seul facteur *taille*.

Pour simplifier, supposons que l'on dispose des deux éléments suivants[2] :

- il y a autant de femmes que d'hommes dans la population considérée
- après la croissance, les femmes adultes mesurent en moyenne[3] 1.65 m avec un écart-type de 16 cm (moyenne = 1.75 m et écart-type = 15 cm pour les hommes).

---

2 Ces données n'ont bien évidemment aucune valeur significative.

Quelle est le sexe d'une personne mesurant 1.60 m ?

Comment répondre intelligemment à cette question ? Une première réflexion de bon sens conduit à dire que cette personne est une femme. Mais, tout le monde connaît des hommes de cette taille. La réponse est donc erronée. Une meilleure réponse consistera à dire, par exemple, que cet individu a une probabilité de 60% d'être une femme et la probabilité complémentaire d'être un homme (40%).

Nous n'avons plus à faire à une classification binaire (c'est une femme ou c'est un homme) mais à une classification probabiliste. De plus, face à un tel problème, une réponse probabiliste est une bonne solution ; en effet, la taille ne suffit pas à départager distinctement les deux classes, mais elle apporte une information interprétée en terme de probabilité.

Dans le chapitre suivant, nous verrons que la règle de décision de Bayes, qui est fondée sur la probabilité d'appartenance à chacune des classes, permet de minimiser le risque d'erreur de classification. En tout état de cause, la procédure et les méthodes de résolution des problèmes de classification présentées dans ce mémoire s'appliquent aux différents cas (classification binaire ou probabiliste).

Notons enfin que, lorsque les critères d'évaluation sont subjectifs (comme c'est le cas dans l'exemple de classification des restaurants, il est possible d'utiliser des techniques de classification *floue*, qui constituent une alternative aux techniques bayesiennes. Nous n'aborderons pas cette approche dans le présent mémoire.

1.7 Résolution des problèmes de classification

Lorsque l'expert ne peut pas expliciter son processus de classification, il faut se tourner vers des systèmes de classification qui "apprennent par l'exemple". A partir d'un lot d'individus déjà classés par l'expert, le système peut apprendre à classer comme l'expert. Après apprentissage, le système est capable de classer de nouveaux individus.

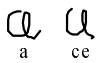
1.7.1 Un principe de résolution : l'élaboration d'un modèle statistique par *apprentissage*

Prenons l'exemple de la lecture qui est aussi un exercice de classification. En effet, elle consiste, pour un texte normal, à classer des signes en 26 classes que sont les lettres de l'alphabet. Si la classification sous-jacente à toute lecture ne pose pas beaucoup de problèmes lorsqu'il s'agit d'un document imprimé, on sait à quel point l'exercice peut devenir difficile avec certaines écritures manuscrites !

---

[3] Pour être plus précis, il faudrait donner la loi de distribution de la taille des hommes et des femmes comme par exemple la loi de Gauss, mais ce n'est pas le point important de ce paragraphe. Le chapitre suivant décrit la règle de décision tirée des courbes de distribution des individus des différentes classes.

Par exemple, les signes ci-dessous doivent-ils être lus "a" ou "ce" ?



Dans la pratique, le contexte permet d'élucider la plupart de ces difficultés de déchiffrage d'une écriture, c'est-à-dire de classification des signes qui la composent. Mais, lorsque le sens ne permet pas cette élucidation, il reste la possibilité de regarder comment sont écrits les autres "a" que l'on a reconnus de manière certaine.

Ce petit exemple illustre le principe de résolution des problèmes de classification à partir d'observations, que nous désignerons, conformément à l'usage dans le domaine des réseaux de neurones, sous le terme *d'exemples*. Pendant la phase d'apprentissage, on apprend à reconnaître la lettre "a" dans quelques cas non ambigus, et, par la suite, on peut identifier ce signe dans d'autres situations.

1.7.2 Procédure de résolution par apprentissage

La résolution des problèmes de classification par apprentissage se déroule donc en plusieurs étapes :

- première étape : faire classer un échantillon d'individus par un expert ; cet échantillon est désigné, dans le domaine des réseaux de neurones, sous le nom de *base d'apprentissage*,
- deuxième étape : concevoir et mettre en œuvre un algorithme (appelé *classifieur*) qui parvient à reproduire la classification de l'échantillon d'apprentissage,
- troisième étape : évaluer la qualité du classifieur en l'appliquant à un ensemble d'individus classés par l'expert, mais qui n'ont pas été utilisés au cours de la phase d'apprentissage (cet ensemble est la *base de test*),
- dernière étape : si le test est satisfaisant, appliquer la méthode de la deuxième étape à l'ensemble de la population à classer.

C'est cette procédure qui est appliquée lorsqu'on confie la résolution d'un problème de classification à une machine. Elle porte alors le nom de ***classification supervisée*** car elle requiert l'intervention d'un "superviseur" ou expert. Notons dès maintenant que la deuxième étape consiste bien à reproduire la classification de la base d'apprentissage à l'aide d'un algorithme numérique, et non pas à expliquer de manière linguistique la règle de décision mise en œuvre.

1.7.3 L'apprentissage ne résout pas tout ; il reste des précautions à prendre

En pratique, la procédure décrite précédemment peut se révéler très difficile à mettre en œuvre et conduire à un résultat inexploitable. Prenons l'exemple fictif d'une population d'individus {I} que l'on voudrait ranger dans trois classes A, B et C. Imaginons la situation suivante : un expert est sollicité pour effectuer cette classification et, après avoir classé les 15 premiers individus, il démissionne sans transmettre son savoir-faire. Afin de poursuivre ce

travail, l'entreprise demande à un de ses employés de remplacer l'expert. Le collègue, qui découvre cette étude, commence par recueillir 3 grandeurs (X, Y et Z) caractéristiques des 15 individus. Il lui faut maintenant retrouver la fonction de classification de l'expert.

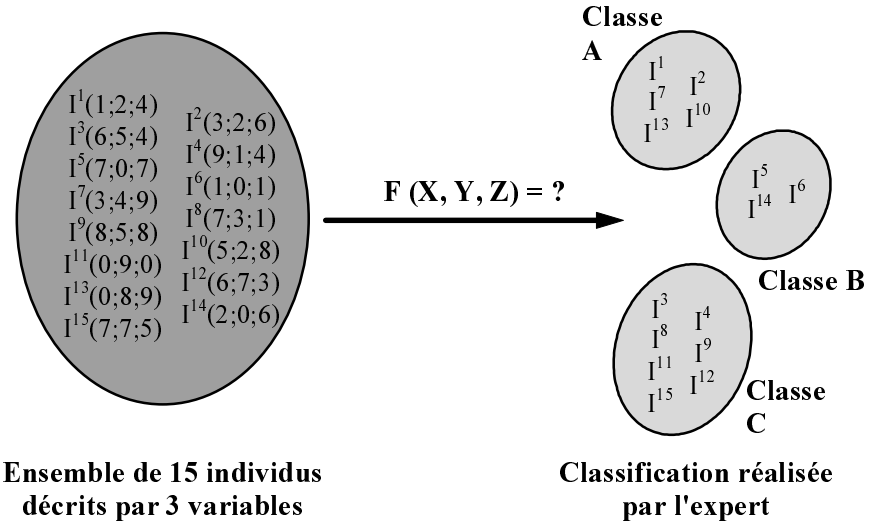


Figure 1.2 : Présentation d'un problème de classification

Ces 15 individus constituent donc la base d'apprentissage pour le collègue. Sa première tâche va consister à essayer de reproduire la classification de cet échantillon (deuxième étape de la procédure). En l'occurrence, il va tenter de trouver quelle réflexion l'expert a bien pu mener pour conclure que le premier individu I^1 décrit par (1;2;4) devait être rangé en classe A. Comme dans l'exemple des pièces, on se rend rapidement compte que **plusieurs règles de décision** peuvent réaliser la classification de l'échantillon[4]. La première à laquelle on peut penser est la fonction F_1 suivante :

Si X = 1, Y = 2 et Z = 4 alors F_1 (X, Y, Z) = A
Si X = 3, Y = 2 et Z = 6 alors F_1 (X, Y, Z) = A
...
Si X = 7, Y = 7 et Z = 5 alors F_1 (X, Y, Z) = C

F_1 énonce, individu par individu, la classification de l'expert qu'elle reproduit donc parfaitement. Mais, si l'individu suivant est décrit par (2;0;0), dans quelle classe se range-t-il ? Cette première fonction envisagée ne permet pas de répondre. On peut dire qu'elle est trop **spécialisée**.

Considérons une autre fonction, F_2, définie par :

Si X + Y < Z alors F_2 (X, Y, Z) = A
Si X + Y = Z alors F_2 (X, Y, Z) = B
Si X + Y > Z alors F_2 (X, Y, Z) = C

[4] En fait, il existe une infinité de fonctions capables de reproduire exactement la classification de n'importe quel nombre fini d'individus !

F_2 reproduit parfaitement la classification de l'expert à une exception près, l'individu I^{14} décrit par (2;0;6). F_2 le range en classe A (2+0 < 6) alors que l'expert l'a mis en classe B. Cependant, cette seconde fonction, plus **générale**, présente l'intérêt considérable de permettre de classer tout nouvel individu : c'est un modèle *prédictif* alors que la méthode précédente constituait seulement un modèle *descriptif*.

Alors que son collègue vient juste de trouver cette fonction F_2 qui semble satisfaisante, l'expert revient et lui apprend qu'en fait la fonction de classification ne portait que sur Y, qu'il ne fallait tenir compte ni de X ni de Z figurant dans le dossier pour d'autres utilisations. Sa règle est en effet :

Si	Y est pair	alors	F (X, Y, Z) = A
Si	Y = 0	alors	F (X, Y, Z) = B
Si	Y est impair	alors	F (X, Y, Z) = C

Cet exemple illustre, de façon caricaturale[5], les principales difficultés que l'on rencontre dans la résolution par apprentissage des problèmes de classification supervisée :

- **choix des variables descriptives :** dans l'exemple, si la seule variable Y avait figuré, la règle de classement fondée sur la parité aurait vraisemblablement sauté aux yeux ! Nous verrons dans ce travail que les méthodes statistiques de résolution sont, elles aussi, gênées par les variables non pertinentes vis-à-vis du problème posé.
- **optimisation de la fonction :** il faut toujours trouver un compromis entre une fonction très performante sur les individus de la base d'apprentissage et une fonction peut-être moins performante sur l'échantillon, mais qui présente de meilleures capacités de "généralisation".
- **taille de l'échantillon** : si la classe B avait comporté 50 individus, on aurait certainement vu que leur point commun était d'avoir une valeur de Y nulle. Autrement dit, la base d'apprentissage doit être suffisamment grande et représentative.

1.8 Conclusion

Dans ce chapitre, nous avons présenté, de manière empirique, ce qu'est un problème de classification. Nous avons vu que, pour certains de ces problèmes nous pouvons décrire explicitement, de manière linguistique, le mécanisme de classification (pièces de monnaie). Dans ce cas, un algorithme reproduisant ce processus peut être construit et le problème est résolu. Pour d'autres problèmes, il est malheureusement impossible de décrire précisément la

---

[5] Cet exemple présente en effet une pathologie que l'on ne rencontre heureusement pas dans la pratique : il suffit que la valeur de *Y* change de parité pour que l'objet change de classe ; autrement dit, la plus petite variation possible du descripteur pertinent entraîne un changement de classe. Dans l'espace du descripteur (l'ensemble des entiers) les deux classes se recouvrent complètement. Dans toute la suite, nous supposerons qu'il est possible de trouver des descripteurs tels que les classes ne se recouvrent que partiellement.

classification (évaluation des restaurants) ; il faut alors trouver des méthodes de résolution qui "apprennent par l'exemple", à partir d'un ensemble d'individus déjà classés par un superviseur. Le chapitre suivant présente les méthodes statistiques de résolution de ce type de problèmes.

Un point très important a été également soulevé : celui du choix des variables descriptives en fonction de leur contribution à la résolution du problème posé. Nous verrons qu'un mauvais choix de descripteurs peut, à lui seul, dégrader les résultats d'une bonne méthode de classification. Dans ce travail, nous proposerons une méthode originale de sélection de variables descriptives qui met bien en évidence l'inutilité de certaines d'entre elles. Cette méthode de sélection sert également à la définition de l'architecture de réseaux de neurones formels, ce qui permet de construire un classifieur performant.