# 6. APPLICATION À L'ANALYSE FINANCIÈRE

## Résumé

*Dans le cadre de la collaboration entre le laboratoire d'Électronique de l'ESPCI et Informatique CDC, la Caisse des Dépôts et Consignations (CDC) a souhaité développer de nouveaux Systèmes Informatiques d'Aide à la Décision (SIAD) utilisant les techniques neuronales.*

*Cette étude portant sur l'utilisation des réseaux de neurones comme classifieurs a débouché sur deux applications de natures voisines (seuls les objets à classer sont différents). En effet, les deux systèmes de classification réalisés ont pour objectif d'apporter une évaluation, soit d'entreprises, soit de collectivités locales (en terme de santé financière).*

*Nous présentons dans ce chapitre, ces deux applications dénommées respectivement :*

- *Analyse Financière des Entreprises (AFE)*
- *Analyse Financière des Collectivités Locales (AFCL)*

*Elles adoptent une démarche identique, qui commence par la sélection des meilleures variables descriptives du modèle, passe par la détermination du meilleur réseau de neurones, et s'achève par la mise en exploitation du classifieur.*

## 6.1 Présentation

Ce chapitre présente deux applications des réseaux de neurones centrées autour du thème de l'analyse financière soit d'entreprises soit de collectivités locales. Chronologiquement, l'étude de l'analyse financière des entreprises a commencé en premier ; elle a très rapidement fait apparaître la nécessité de sélectionner les meilleures variables descriptives d'un modèle. Nous avons donc orienté notre étude sur les méthodes de sélection de modèles, présentées d'une manière théorique dans le chapitre 5 (La sélection de modèles). Puis, à partir des meilleures variables descriptives, nous avons cherché à reproduire, de la meilleure façon possible, la classification de l'expert.

Nous présentons donc ces deux études, qui suivent la procédure, en quatre étapes, de résolution des problèmes de classification (voir paragraphe 1.7.2) :

- Ainsi, la première étape a consisté à construire un échantillon d'individus (les entreprises ou les collectivités locales).
- Puis nous avons demandé à des experts de la Caisse des Dépôts et Consignations d'évaluer ces individus en fonction de leurs propres connaissances.
- Ensuite, nous avons cherché à reproduire cette évaluation.
- Et, dans le cas de l'analyse financière des entreprises, cette étude a débouché sur une application opérationnelle depuis 1995.

Le travail sur la première étude (l'analyse financière des entreprises) est présenté en détail. La seconde (l'analyse financière des collectivités locales) suit le même scénario et sera donc moins détaillée.

## 6.2 Analyse Financière des Entreprises

Dans un premier temps, nous présentons brièvement la démarche de l'analyse financière des entreprises. Ensuite, nous définissons les ratios financiers qui sont les variables descriptives de notre problème.

Un important travail de sélection des meilleurs ratios a démarré au début de cette étude ; il a permis, en utilisant des critères qualitatifs, de diminuer le nombre de descripteurs du modèle. Nous indiquons les différentes performances des classifieurs élaborés à partir de l'ensemble des descripteurs initiaux et du sous-ensemble des descripteurs sélectionnés. Pour finir, nous présentons les résultats obtenus en utilisant la méthode "automatique" de sélection des descripteurs et de l'architecture du réseau de neurones.

### 6.2.1 Présentation de l'analyse financière

Le gestionnaire de portefeuille de la Caisse des Dépôts et Consignations a pour mission de gérer et de faire fructifier des fonds en les investissant en actions, suivant deux objectifs principaux :

- dégagement de plus-values par des décisions de vente de titres,
- constitution d'une réserve financière importante par l'achat de nouveaux titres.

Avant d'effectuer une opération d'achat ou de vente d'actions, il est nécessaire de procéder à l'analyse financière de l'entreprise concernée [Grémillet 73, Solnik 80 et Vernimmen 88]. Cette analyse consiste notamment à apprécier la solidité financière de la société et sa capacité à dégager des bénéfices ; elle est sanctionnée par l'attribution d'une note, image du risque encouru lors du placement.

Les décisions d'investissement ou de désinvestissement reposent sur une expertise du marché et sur l'évaluation des titres à acheter. L'un des rôles du gestionnaire de portefeuille est donc de repérer les "bons" et "mauvais" supports d'investissement à partir de données financières. La condition d'utilisation d'un système d'aide à la décision de type classifieur (tel que nous allons le concevoir) est de ne jamais commettre d'erreurs entre ces deux groupes d'entreprises (les "bons" et les "mauvais" supports d'investissement).

Devant l'énorme masse d'informations comptables, le gestionnaire sélectionne différentes rubriques des documents comptables (essentiellement le ***bilan*** et le ***compte de résultat***) et effectue des rapprochements entre celles-ci, afin de définir des ***ratios*** (ce sont les

variables descriptives du modèle) susceptibles de synthétiser et de mettre en évidence les caractéristiques économiques et financières de l'entreprise étudiée[1].

C'est cette évaluation, fondée sur les ratios, que l'on cherche à reproduire dans cette étude.

### 6.2.2 Définition des ratios utilisés

Le nombre et la définition des ratios à choisir pour évaluer l'état d'une société sont susceptibles de changer en fonction de l'analyste, du but recherché, des données disponibles, etc. Dans tous les cas, il est nécessaire d'en sélectionner un nombre restreint tout en s'assurant que ces ratios couvrent l'ensemble de la gestion de l'entreprise. Dans notre cas, l'approche est celle du gestionnaire de portefeuille d'actions, qui en utilise 15.

Pour effectuer son analyse, le gestionnaire utilise un domaine de validité et un critère associés à chaque ratio. Le domaine de validité permet de rejeter les ratios s'écartant trop des valeurs types (si un ratio n'appartient pas à son domaine il est considéré comme "mauvais"). Le critère permet de séparer les "bons" et les "mauvais" ratios.

Dans ce paragraphe, nous donnerons les définitions des ratios utilisés. L'annexe C en donne également une analyse succincte.

A chacun des 15 ratios sélectionnés par le gestionnaire, nous associons un numéro ; par souci de commodité, nous utilisons ici les numéros qui apparaissent dans la base de données provenant de la Centrale des Bilans[2]. Ces ratios peuvent être regroupés en quatre catégories :

- Les ratios de structure financière :

| Numéro | Définition |
|---|---|
| 10 | Dettes à long et moyen terme / Fonds propres |
| 15 | Capitaux permanents / Actif immobilisé |
| 25 | Dettes à long et moyen terme / Marge brute d'autofinancement |
| 35 | Total dettes / Total actif |

*Tableau 6.1 : Ratios de structure financière*

- Les ratios de rentabilité :

[1] L'annexe C (Éléments d'analyse financière) présente quelques notions comptables nécessaires à l'analyse financière des entreprises, les définitions des 15 ratios utilisés par le gestionnaire et enfin une interprétation de ces ratios.

[2] La centrale des Bilans est l'organisme qui a fourni les données.

| Numéro | Définition |
|---|---|
| 45 | Valeur ajoutée / Chiffre d'affaires |
| 50 | Excédent brut d'exploitation / Valeur ajoutée |
| 65 | Frais financiers / Chiffre d'affaires |
| 80 | Résultat net / Chiffre d'affaires |
| 85 | Résultat net / Fonds propres |

*Tableau 6.2 : Ratios de rentabilité*

- Les ratios de gestion :

| Numéro | Définition |
|---|---|
| 100 | Rotation des stocks (en mois) |
| 105 | Durée crédits clients (en mois) |
| 110 | Durée crédits fournisseurs (en mois) |
| 115 | Rotation du besoin en fonds de roulement (en mois) |

*Tableau 6.3 : Ratios de gestion*

- Les ratios financiers :

| Numéro | Définition |
|---|---|
| 145 | Investissements / Valeur ajoutée |
| 155 | Disponible après financement interne de la croissance / Valeur ajoutée |

*Tableau 6.4 : Ratios financiers*

A partir de ces données (les 15 ratios définis précédemment sur les 3 dernières années disponibles), le gestionnaire de portefeuille a analysé et classé un ensemble d'entreprises, et constitué ainsi l'échantillon d'exemples qui a servi à l'apprentissage des classifieurs.

### 6.2.3 Constitution de l'échantillon d'entreprises

Pour mettre en œuvre les méthodes statistiques de classification, il faut, dans un premier temps, constituer la base d'apprentissage (voir chapitre 1 : Qu'est-ce que la classification ?). Pour cela, nous avons demandé au gestionnaire de portefeuille d'évaluer, en fonction des ratios financiers, un échantillon d'entreprises.

#### *6.2.3.1 Échantillon de départ*

La base de données fournie par la Centrale des Bilans contient les ratios des entreprises cotées en bourse. Au début de cette étude, la base de données rassemblait les ratios de 624 entreprises. Sur ces 624 entreprises, 129 ont été écartées par manque de données. De plus, à la demande du gestionnaire, les entreprises financières et immobilières (62 entreprises) ont aussi été écartées, car l'analyse de leurs ratios diffère de celles des autres sociétés. Finalement, ce premier échantillon contient 433 entreprises.

Pour son analyse, le gestionnaire dispose donc :

- du nom de l'entreprise,
- de son secteur d'activité,
- et des 15 ratios financiers sur les trois dernières années (1990, 1991 et 1992).

A partir de ces données, il émet une évaluation que l'on code sur **3 classes** :

- Classe A (bon support d'investissement) : le gestionnaire investit, en priorité, sur ces entreprises.
- Classe B (entreprise neutre) : le gestionnaire n'investit pas sur ces entreprises mais il continuera à suivre leur évolution.
- Classe C (support d'investissement très risqué) : le gestionnaire ne s'intéresse plus à ces entreprises pour l'année qui vient.

Nous avons recueilli les notes des 433 entreprises, qui couvrent tous les secteurs d'activité à l'exception de trois d'entre eux (immobilier, services financiers et investissement). Ce premier échantillon constitue la base de données sur laquelle nous nous sommes appuyés dans la suite de ce travail.

#### *6.2.3.2 Pré-traitement*

A partir de échantillon des 433 entreprises notées, nous cherchons à éliminer les entreprises présentant des singularités.

C'est le rôle du pré-traitement qui intègre les connaissances *a priori* du problème pour obtenir un échantillon de travail le plus "sain" possible. Nous nous intéressons également à la répartition des ratios.

Pour cela on opère trois tris successifs :

- Le premier tri consiste à éliminer les entreprises dont un seul des ratios présente un écart à la moyenne (valeur moyenne des ratios du même type) trop important.

  Critère : Si "| Ratio - Moyenne | > 9×Écart-type" $\Rightarrow$ élimination de l'entreprise

  Résultat : 11 entreprises ont été supprimées.

- Le deuxième tri est lié aux définitions même des ratios. Ainsi, le ratio 25 (Dettes à long et moyen terme / Marge brute d'autofinancement) ne peut être négatif si le ratio 80 (Résultat net / Chiffre d'affaires) est positif.

  Critère : Si "Ratio 25 < 0.0 et Ratio 80 > 0.0" $\Rightarrow$ élimination de l'entreprise

  Résultat : 12 entreprises ont été supprimées.

- Le troisième tri s'intéresse au ratio 10 (Dettes à long et moyen terme / Fonds propres). D'après le gestionnaire, une entreprise qui possède un ratio 10 négatif est en situation de faillite. C'est une donnée *a priori* du problème ; on élimine ces entreprises de l'échantillon, et on les classe d'office dans la classe C.

Critère : Si "Ratio 10 < 0.0" $\Rightarrow$ élimination de l'entreprise

Résultat : 12 entreprises ont été supprimées.

Ces trois étapes conduisent à rejeter 35 exemples, et finalement l'échantillon se réduit à **398 entreprises**[3]. La répartition des classes est la suivante :

| Classe | Nombre d'entreprises |
|---|---|
| A | 172 |
| B | 172 |
| C | 54 |

*Tableau 6.5 : Répartition des classes (probabilités a priori)*

Comme nous ne disposons pas d'autres renseignements sur la répartition des classes, c'est ce dénombrement qui nous donne une estimation des probabilités[4] *a priori* d'appartenance à chacune des classes. On obtient :

$$\Pr_A = \frac{172}{398} \approx 0.43 \text{ pour la classe A,}$$

$$\Pr_B = \frac{172}{398} \approx 0.43 \text{ pour la classe B et}$$

$$\Pr_C = \frac{54}{398} \approx 0.14 \text{ pour la classe C.}$$

Après les différents traitements appliqués à la base de départ (recueil des données, rejet des entreprises non renseignées, évaluation du gestionnaire, élimination de certaines entreprises), nous obtenons une base d'apprentissage comportant **398 entreprises** appartenant à **3 classes**.

La première étape de la résolution d'un problème de classification de type "expertise" est achevée. Il faut maintenant passer à la reproduction de la notation de l'expert en mettant en œuvre les techniques développées dans les chapitres précédents.

### 6.2.4 Pertinence des ratios

Toutes les méthode d'analyse des données imposent un rapport entre le nombre d'individus et le nombre de descripteurs le plus élevé possible. De plus, nous avons vu qu'il était toujours préférable de garder les descripteurs pertinents et d'éliminer les autres.

---

[3] Tous les traitements présentés dans ce mémoire ont été réalisés sur cet échantillon comportant 398 entreprises.

[4] Les estimations des probabilités *a priori* sont utilisées par les méthodes indirectes de classification. Avec les méthodes directes, elles demeurent "cachées". Néanmoins, nous pouvons les modifier après l'apprentissage (voir chapitre 2 : Méthodes statistiques de classification) ; ce que nous ne ferons pas car nous ne disposons d'aucune connaissance *a priori* contradictoire.

Dans ce paragraphe, nous cherchons donc à réduire le nombre de descripteurs en distinguant les ratios essentiels à la notation de ceux qui ont moins d'importance.

#### *6.2.4.1 Analyse "qualitative" de la pertinence des ratios*

Dans ce but, un examen "qualitatif" et "visuel" des fiches de notation du gestionnaire nous a permis de mettre en évidence un sous-ensemble de ratios qui semblent avoir plus d'importance que les autres. Cette analyse des fiches de notation a fait ressortir 3 points :

- le gestionnaire semble donner une importance primordiale aux ratios de la dernière année, les autres années participent au jugement mais dans une moindre mesure ;
- tous les ratios de la dernière année ne semblent pas avoir le même "poids". Par exemple le ratio 80 (Résultat net / Chiffre d'affaires) semble être examiné avec le plus grand soin. De plus, une valeur négative implique, dans la majorité des cas, un mauvais classement de l'entreprise ;
- les autres ratios semblent moins influencer la classification.

Cette première analyse "visuelle" a donc conduit à la sélection d'un sous-ensemble de ratios de la dernière année parmi l'ensemble complet des 45 ratios initiaux. Nous avons couplé cette étude "visuelle" à une étude qualitative de l'évolution de la moyenne des ratios en fonction de la note.

En confrontant les résultats obtenus par ces deux méthodes, nous avons finalement sélectionné **7 ratios** de la dernière année comme étant les plus significatifs de la notation :

| **Numéro** | **Définition** |
|---|---|
| 10 | Dettes à long et moyen terme / Fonds propres |
| 25 | Dettes à long et moyen terme / Marge brute d'autofinancement |
| 35 | Total dettes / Total actif |
| 50 | Excédent brut d'exploitation / Valeur ajoutée |
| 80 | Résultat net / Chiffre d'affaires |
| 85 | Résultat net / Fonds propres |
| 155 | Disponible après financement interne de la croissance / Valeur ajoutée |

*Tableau 6.6 : Les 7 ratios sélectionnés*

Le tableau 6.6 présente les 7 ratios de la dernière année sélectionnés par la méthode d'analyse "qualitative" de la pertinence des descripteurs.

#### *6.2.4.2 Validation de la sélection des ratios importants*

Les remarques concernant la sélection des 7 ratios importants ont été présentées au gestionnaire, qui a dans l'ensemble, validé ce choix. Ses commentaires sont reportés ci-dessous.

- Si les ratios de la dernière année sont très bons ou très mauvais, le gestionnaire note la société sans examiner les ratios des deux premières années. En revanche, si les

ratios de la dernière année sont neutres ou seulement mauvais, il accorde beaucoup d'importance à l'évolution de la société dans le temps (principalement entre l'avant-dernière et la dernière année). Le fait de ne conserver que l'information de la dernière année ne peut donc pas entraîner de grosses erreurs de notation.

- Les ratios de gestion (ratios 100, 105, 110 et 115) sont moins importants que les autres. De plus leur analyse dépend beaucoup du secteur d'activité. Par exemple, la durée typique de rotation des stocks (ratio 100) peut varier de quelques jours (distribution) à quelques mois (aéronautique).
- Les ratios de structure financière (ratio 10, 15, 25 et 35) sont très importants. Mais ces ratios sont liés entre eux, on peut en supprimer sans perdre de l'information. Le secteur d'activité a une faible influence sur ces ratios.
- Pour les autres ratios (ratios financiers et de rentabilité), on peut exclure le ratio 45, très important pour l'analyse prévisionnelle, mais moins dans le cadre de cette étude. De même, les ratios 65 et 145 sont redondants si l'on conserve tous les autres.

Le choix des 7 ratios de la dernière année est donc assez bien validé. De plus, ces ratios (excepté le ratio 80 qui est différent pour le secteur de la distribution) sont peu sensibles au secteur d'activité ; ils sont donc susceptibles de conduire à un bon modèle statistique de classification.

Dans la suite du mémoire, nous allons comparer les résultats obtenus à partir des 7 ratios importants à ceux obtenus à partir des 45 ratios initiaux en utilisant les méthodes de classification présentées au chapitre 2 (Méthodes statistiques de classification).

### 6.2.5 Résultats

A partir des deux ensembles de descripteurs (les 45 ratios initiaux et les 7 sélectionnés), nous construisons différents classifieurs fondés sur les méthodes suivantes (voir chapitre 2) :

- $k$ plus proches voisins (ici $k = 1$) : en un point de l'espace de description, on affecte à l'individu inconnu celle de l'exemple (de l'ensemble d'apprentissage) le plus proche. Nous n'utilisons pas la méthode des noyaux de Parzen ; en effet, la grande dimension du problème conduit à de très mauvais résultats.
- analyse discriminante : on utilise ici l'analyse discriminante avec une règle d'affectation géométrique. Avec la règle d'affectation probabiliste (hypothèse de distribution gaussienne), les résultats sont moins bons.
- réseau de neurones (codage "grand-mère" ou 1/All) : c'est un réseau de neurones à une couche cachée et trois neurones de sortie (un pour chaque classe). On essaie toutes les configurations possibles (entre 0 et 20 neurones cachés) et on retient le meilleur réseau.
- réseau de neurones (séparation 2 à 2) : pour chacun des 3 sous-problèmes (séparation A/B, A/C et B/C) on utilise un réseau de neurones qui estime la

probabilité *a posteriori* d'appartenance à une classe sachant que l'individu appartient à l'une des deux. On combine ces probabilités 2 à 2 à l'aide de la règle donnée au paragraphe 2.7.2. Là encore, on essaie toutes les architectures.

Pour chaque méthode, et pour chaque ensemble de descripteurs, on réalise **100 partitions** différentes de la base des entreprises (80% en apprentissage et 20% en test), et on donne (tableau 6.6) la moyenne (et l'écart-type entre parenthèse) des taux d'exemples bien classés obtenus sur la **base de test**.

La tableau 6.7 fournit les résultats :

| **Méthode de classification** | **Variables descriptives** | | | |
|---|---|---|---|---|
| | **45 ratios** | | **7 ratios** | |
| | **Performance** | **Erreur A ↔ C** | **Performance** | **Erreur A ↔ C** |
| Plus proche voisin | 58,3% (4,4%) | 1,2% (1,1%) | 70,0% (4,2%) | 0,3% (0,5%) |
| Analyse discriminante | 67,5% (4,9%) | 0,9% (1,0%) | 68,9% (4,9%) | - |
| Réseau de neurones (1/All) | 79,1% (3,8%) | 0,2% (0,5%) | 82,7% (3,3%) | - |
| Réseaux de neurones (2 à 2) | 83,3% (3,7%) | 0,1% (0,4%) | **86,2% (3,3%)** | - |

*Tableau 6.7 : Résultats de l'analyse financière des entreprises*

Dans le tableau, on trouve :

- dans la première colonne, la méthode de classification,
- puis, pour le premier ensemble de variables descriptives considéré (ici les 45 ratios initiaux), la moyenne sur les 100 partitions de l'ensemble de test des taux d'entreprises bien classées (ainsi que l'écart-type, entre parenthèses),
- et la moyenne (et l'écart-type) du taux d'entreprises qui passent de la classe A vers la classe C (ou inversement).

Quelles sont les conclusions que l'on peut tirer de ce tableau de résultats ?

- La sélection des meilleures variables descriptives est un point clef. On constate, en effet, que les résultats sont toujours meilleurs avec les 7 ratios sélectionnés. Sans cette sélection, le classifieur n'aurait jamais pu faire l'objet d'une application opérationnelle car aucune des méthodes de classification ne permet de séparer, sans commettre d'erreur, les entreprises notées A de celles notées C.
- Les méthodes neuronales obtiennent de meilleurs résultats que les autres méthodes classiques. De plus, en décomposant ce problème (à 3 classes) en sous-problèmes (à 2 classes), nous avons atteints le meilleur taux d'entreprises bien classées.

La figure 6.1 présente l'architecture du réseau de neurones utilisé dans l'application opérationnelle à la Caisse des Dépôts et Consignations depuis 1995 :

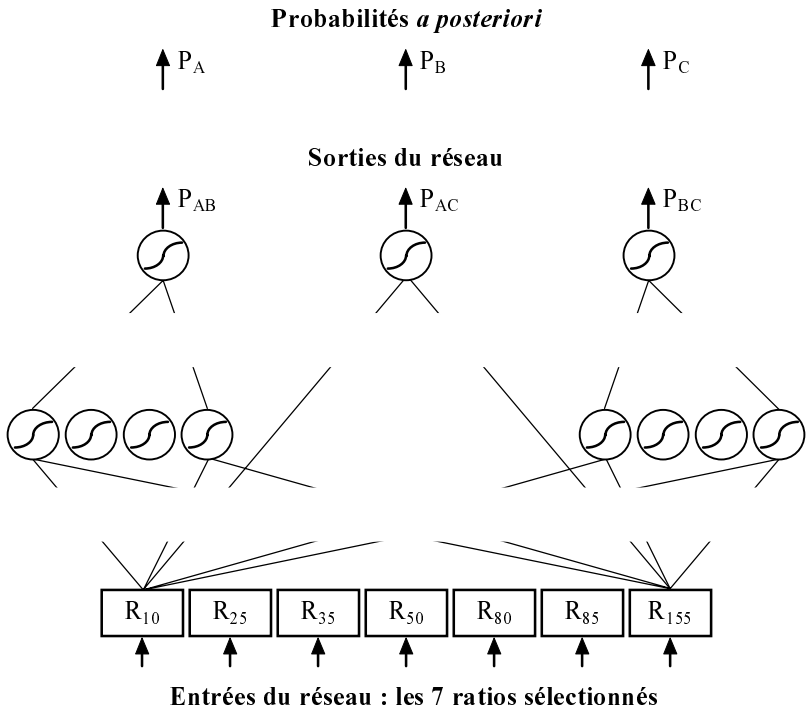


*Figure 6.1 : Réseau de neurones opérationnel*

Au cours de cette étude, nous avons développé une méthode automatique de sélection d'architecture (variables descriptives et neurones cachés) de réseaux de neurones. Il faut maintenant s'assurer qu'elle conduit aussi à des résultats satisfaisants sur un problème réel.

### 6.2.6 Méthode de sélection d'architecture de réseaux de neurones

Nous utilisons la méthode originale de sélection de modèles (présentée au chapitre 5) pour résoudre ce problème de classification. Nous décomposons le problème à 3 classes (A, B et C) de la même manière que celle illustrée par la figure 6.1 (séparation des classes 2 à 2).

Pour chaque sous-problème, l'algorithme d'orthogonalisation de Gram-Schmidt associé aux réalisations du descripteur aléatoire sélectionne un nombre de ratios. Puis, le même algorithme appliqué aux neurones de la couche cachée supprime les neurones inutiles.

On obtient finalement et automatiquement les résultats suivants :

| Méthode automatique de sélection de l'architecture d'un réseau de neurones | Performance | Erreur A ↔ C |
|---|---|---|
| | 86,7% (3,4%) | - |

*Tableau 6.8 : Résultats*

Nous retrouvons exactement (la différence n'est pas significative car très inférieure à l'écart-type) la même performance que celle obtenue avec le meilleur réseau de neurones à 7 entrées. En revanche, le gain de temps est très important puisqu'il ne faut plus tester toutes les configurations possibles. Grossièrement, on a gagné un facteur 20 en temps de calcul (on est passé de 2-3 jours à 3-4 heures).

### 6.2.7 En résumé

Sur cette étude expérimentale, nous avons montré que :

- la sélection des meilleures variables descriptives est un point très important pour un problème de classification (et plus généralement de modélisation).
- les réseaux de neurones, bien dimensionnés, obtiennent les meilleurs résultats. Les travaux antérieurs d'analyse de la solidité financière des entreprises sont fondés sur l'analyse discriminante [Altman 93]. Comme la limitation de l'analyse discriminante en classification est la linéarité de la méthode, nous avons montré que les réseaux de neurones, étant une méthode classification non linéaire, apportent une amélioration sensible en terme de taux d'exemples bien classés par rapport à l'analyse discriminante.

Pour trouver la bonne architecture du réseau de neurones, une méthode brutale consiste à essayer toutes les configurations possibles. Pour pallier cet inconvénient, nous avons utilisé la méthode originale de sélection de modèles qui a donné des résultats équivalents en nécessitant un temps de calcul très inférieur.

De plus, cette étude a fait l'objet d'un développement d'une application opérationnelle. Mensuellement, une liste d'entreprises classées par le réseau de neurones est fournie au gestionnaire de portefeuille ; les entreprises qui changent de classe par rapport à l'évaluation précédente sont spécialement signalées.

## 6.3 Analyse Financière des Collectivités Locales

Après cette première application sur l'analyse financière d'entreprises, nous nous sommes intéressés à l'analyse financière des collectivités locales. Ces deux applications sont très proches car elles consistent à analyser, puis noter, des objets (entreprises ou collectivités locales) en fonction de leur santé financière.

La démarche de cette deuxième étude est semblable à celle adoptée pour la résolution de la précédente. Sa description sera donc plus rapide. Nous n'aborderons que les points qui ont été traités différemment.

### 6.3.1 Présentation du problème

Le groupe Caisse des dépôts exerce des activités spécifiques en faveur du développement économique et social local. Ainsi, il accompagne la politique de la villes et des quartiers en difficulté en finançant des opérations de rénovation urbaine, d'insertion par l'activité économique, et la construction de logements. De plus, le groupe est aussi actionnaire et prestataire de services de près de 500 sociétés d'économie mixte (SEM), outils des collectivités locales (environ 36000) dans les domaines du logement social, de l'aménagement, des transports et de la gestion des services publics locaux.

Avant de prêter de l'argent aux Sociétés d'Économie Mixte (SEM), la Caisse des Dépôts et Consignations doit vérifier si le garant (c'est-à-dire la collectivité locale concernée) pourra

rembourser l'emprunt. Ainsi, l'attribution du prêt est déterminé par une analyse financière préalable [Bouinot 77, Kerviler 92 et Klopfer 93].

Bien que très proche de la précédente, cette application comporte toutefois quelques différences :

- Contrairement à l'analyse financière des entreprises où un seul expert est intervenu, les évaluations des experts de la CDC sont plus dispersées. En effet, plusieurs experts (de régions différentes) ayant été mis à contribution, il faut s'attendre à des disparités entre les personnes. Comme il n'a pas été possible de constituer un groupe commun de test (collectivités notées par tous les experts), nous n'avons pas pu mesurer leur différence de notation. En d'autres termes, le "bruit" de mesure est certainement beaucoup plus important que pour le problème de l'évaluation des entreprises.
- Les variables descriptives n'ont pas été fournies par les experts. Ils ont évalué les collectivités de l'échantillon en fonction de leur expérience personnelle (indicateurs financiers, intuition, etc). Notre travail a donc consisté à partir d'un ensemble de descripteurs le plus redondant possible (données comptables, fiscales, démographiques et socio-économiques), pour ensuite sélectionner les meilleurs descripteurs avec la méthode de sélection de modèles (voir chapitre 5 : La sélection de modèles) et enfin réaliser la classification.

En résumé, ce problème de notation des collectivités locales est moins bien "défini" que celui des entreprises. Ainsi, la résolution de ce problème de classification est, *a priori*, beaucoup plus difficile que le précédent. En tenant compte de ces remarques, l'objectif de cette étude n'est pas d'obtenir un taux d'exemples bien classés proche des 100 % mais plutôt de construire un classifieur qui ne commet pas d'erreur entre les classes extrêmes.

### 6.3.2 Constitution de l'échantillon de communes

Nous appliquons la procédure de résolution d'un problème de classification, en commençant par constituer un échantillon d'individus classés par un professeur (ou expert).

#### *6.3.2.1 Les notes des experts*

Les analystes de sept Directions Régionales de la Caisse des dépôts (Aquitaine, Bourgogne, Centre, Champagne, Franche-Comté, Pays de la Loire et Rhône-Alpes) ont donc évalué un échantillon de **583 communes**. Le classement est fondé sur l'analyse financière, et il est établi suivant une grille de **5 appréciations** possibles. L'échantillon est réparti de la façon suivante :

| Classe | Nombre d'entreprises |
|---|---|
| A - Très bonnes | 87 soit 14,9% |
| B - Bonnes | 229 soit 39,3% |
| C - Neutres | 169 soit 29,0% |
| D - Mauvaises | 56 soit 9,6% |
| E - Très mauvaises | 42 soit 7,2% |

*Tableau 6.9 : Répartition des 583 communes (probabilités* a priori*)*

### *6.3.2.2 Les descripteurs utilisés*

Comme indiqué plus haut, les experts n'ont pas défini l'ensemble des descripteurs du modèle. Pour résoudre ce problème, nous avons, dans un premier temps, recensé les données susceptibles d'être discriminantes ; elles proviennent de plusieurs "sources" :

- données comptables (fichier Comptabilité Publique, CP),
- données structurelles (fichier Dotation Globale de Fonctionnement, DGF),
- données fiscales (fichier Direction Générale des Impôts, DGI) et
- données de nature socio-économique (fichier Institut National de la Statistique et des Études Économiques, INSEE).

A partir de ces données, nous avons défini 58 descripteurs "potentiels" qui se rapprochent des ratios financiers. Ainsi, le tableau 6.10 présente 12 descripteurs (ratios) construits sur les données provenant du fichier INSEE et le tableau 6.11 présente 23 descripteurs (ratios) fondés sur les données CP-DGF-DGI pour les années 1990 et 1991. Au total, nous avons bien 58 descripteurs ($12 + 2\times23$) :

| Numéro | Descripteurs INSEE | Année |
|---|---|---|
| IN-01-90 | Évolution de la population (en %, de 1982 à 1990) | 1990 |
| IN-02-90 | Superficie forêt / Superficie commune (en %, 1990) | 1990 |
| IN-03-90 | (Population + résidences secondaires) / Sup. commune (1990) | 1990 |
| IN-04-90 | Nombre de chômeurs / Population (en %, 1990) | 1990 |
| IN-05-90 | Nombre d'étrangers / Population (en %, 1990) | 1990 |
| IN-06-90 | Revenus imposables / Nombre foyers fiscaux (1990) | 1990 |
| IN-07-90 | Évolution du ratio Rev. imp. / Foy. fisc. (en %, de 1988 à 1990) | 1990 |
| IN-08-90 | Nombre de résidences principales / Population (1990) | 1990 |
| IN-09-90 | Nombre de résidences secondaires / Population (1990) | 1990 |
| IN-10-90 | Nombre de maisons individuelles / Population (1990) | 1990 |
| IN-11-90 | Nombre de logements HLM / Population (1990) | 1990 |
| IN-12-90 | Nombre de logements vacants / Population (1990) | 1990 |

*Tableau 6.10 : Définition des 12 descripteurs INSEE*

| Numéro | Descripteurs CP-DGF-DGI | Années |
|---|---|---|
| CP-01-9x | Recette de fonctionnement / (Pop. + résidences secondaires) | 90-91 |
| CP-02-9x | Dépense de fonctionnement / (Pop. + résidence secondaires) | 90-91 |
| CP-03-9x | Recette d'investissement / (Pop. + résidence secondaires) | 90-91 |
| CP-04-9x | Dépense d'investissement / (Pop. + résidence secondaires) | 90-91 |
| CP-05-9x | Annuités / Recette de fonctionnement | 90-91 |
| CP-06-9x | Intérêt de la dette / Dépense de fonctionnement | 90-91 |
| CP-07-9x | Total des capitaux restants dus / Épargne brute | 90-91 |
| CP-08-9x | Total des capitaux restants dus / (Pop. + résidences secondaires) | 90-91 |
| CP-09-9x | Épargne disponible / Recette de fonctionnement | 90-91 |
| CP-10-9x | Épargne disponible / (Population + résidences secondaires) | 90-91 |
| CP-11-9x | Épargne brute / Recette de fonctionnement | 90-91 |
| CP-12-9x | Épargne brute / (Population + résidences secondaires) | 90-91 |
| CP-13-9x | Épargne de gestion / Recette de fonctionnement | 90-91 |
| CP-14-9x | Épargne de gestion / (Population + résidences secondaires) | 90-91 |
| CP-15-9x | Effort fiscal | 90-91 |
| CP-16-9x | Potentiel fiscal | 90-91 |
| CP-17-9x | Prélèvement fiscal | 90-91 |
| CP-18-9x | Taux moyen | 90-91 |
| CP-19-9x | Produit 4 taxes / (Population + résidences secondaires) | 90-91 |
| CP-20-9x | Dotation / Produit 4 taxes | 90-91 |
| CP-21-9x | Taxe professionnelle / Taxe d'habitation | 90-91 |
| CP-22-9x | Dépense de personnel / Dépense de fonctionnement | 90-91 |
| CP-23-9x | Taxes communale, départementale et régionale / Rec. de fct. | 90-91 |

*Tableau 6.11 : Définition des 46 descripteurs CP-DGF-DGI*

La principale difficulté rencontrée pour la définition de ces descripteurs est de s'assurer qu'ils couvrent l'ensemble des 36000 collectivités locales de France. En effet, dans l'optique d'une application opérationnelle, il faut pouvoir classer n'importe quelle collectivité. C'est pour cette raison que nous travaillons sur des données des années 1990 et 1991 ; pour les années plus récentes, toutes les communes ne sont pas renseignées.

### 6.3.3 Méthode de sélection d'architecture de réseaux de neurones

Nous avons bien entendu évalué les différentes méthodes de classification présentées dans le chapitre 2 (Méthodes statistiques de classification) ainsi que la méthode originale de sélection de descripteurs et d'architecture de réseaux de neurones. Les résultats sont tout à fait comparables à ceux décrits dans l'application précédente (voir § 6.2). En effet, les réseaux de neurones se révèlent plus performants que les autres méthodes, et la méthode automatique de définition d'architecture de réseaux de neurones retrouve ces résultats.

Nous ne présentons donc que les résultats obtenus en utilisant cette méthode de définition d'architecture de réseaux de neurones.

Ainsi, nous traitons séparément les 10 séparations des classes 2 à 2 (séparation A/B, A/C, A/D, A/E, B/C, B/D, B/E, C/D, C/E et D/E) : pour chacun de ces 10 sous-problèmes, la méthode détermine les descripteurs et l'architecture du réseau de neurones. La sortie de ce réseau constitue une estimation des probabilité 2 à 2, que nous combinons pour obtenir les probabilités *a posteriori* d'appartenance aux 5 classes.

### *6.3.3.1 Sélection des descripteurs*

Le tableau 6.12 présente les descripteurs sélectionnés pour chacun des 10 sous-problèmes de séparation des classes 2 à 2 :

| Descripteur | | Séparation des classes | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Numéro | Année | A/B | A/C | A/D | A/E | B/C | B/D | B/E | C/D | C/E | D/E |
| IN-02-90 | 1990 | | ⊗ | ⊗ | | | | | | | |
| IN-03-90 | 1990 | | | | | | | ⊗ | | | |
| IN-05-90 | 1990 | ⊗ | | | | | | | | | |
| IN-06-90 | 1990 | | ⊗ | | ⊗ | ⊗ | | ⊗ | | | ⊗ |
| IN-08-90 | 1990 | | | | | | | | ⊗ | | ⊗ |
| IN-09-90 | 1990 | ⊗ | ⊗ | ⊗ | | | | | | ⊗ | |
| IN-10-90 | 1990 | | | | | | ⊗ | | ⊗ | | ⊗ |
| CP-02-90 | 1990 | | | | | ⊗ | | | | | |
| CP-06-90 | 1990 | | | | ⊗ | | | | | ⊗ | |
| CP-07-90 | 1990 | | | | ⊗ | | ⊗ | ⊗ | ⊗ | | |
| CP-08-90 | 1990 | | | | | ⊗ | ⊗ | ⊗ | | | |
| CP-12-90 | 1990 | | ⊗ | | | | | | | | |
| CP-14-90 | 1990 | | | | | | | | ⊗ | | |
| CP-15-90 | 1990 | ⊗ | | | | | | | | | |
| CP-16-90 | 1990 | | | | | | | ⊗ | | | |
| CP-17-90 | 1990 | | | | | | ⊗ | | | | |
| CP-18-90 | 1990 | | | | ⊗ | | | ⊗ | | | |
| CP-20-90 | 1990 | | | | ⊗ | | | | | ⊗ | |
| CP-23-90 | 1990 | | | | | | | | | ⊗ | |
| CP-01-91 | 1991 | ⊗ | | | | | | | | | |
| CP-03-91 | 1991 | | | | ⊗ | | | | | | |
| CP-04-91 | 1991 | | | | | | | | | ⊗ | |
| CP-05-91 | 1991 | | ⊗ | | | | | | | | |
| CP-06-91 | 1991 | ⊗ | | | | | | | | | ⊗ |
| CP-07-91 | 1991 | | | | ⊗ | | ⊗ | ⊗ | ⊗ | | |
| CP-08-91 | 1991 | ⊗ | ⊗ | ⊗ | | | | | ⊗ | | |
| CP-09-91 | 1991 | | | ⊗ | | | | | | | |
| CP-10-91 | 1991 | ⊗ | | | ⊗ | | | | | ⊗ | |
| CP-11-91 | 1991 | | | | ⊗ | | | | | ⊗ | ⊗ |
| CP-12-91 | 1991 | | | | | | | ⊗ | | | |
| CP-16-91 | 1991 | | ⊗ | ⊗ | | ⊗ | | | | | ⊗ |
| CP-17-91 | 1991 | | | | | | | | | ⊗ | |
| CP-18-91 | 1991 | ⊗ | ⊗ | | | | | | | | |
| CP-19-91 | 1991 | | | | | | ⊗ | | | | |
| CP-20-91 | 1991 | | | | | | | ⊗ | | | ⊗ |
| CP-22-91 | 1991 | | | | | | | ⊗ | | | |
| CP-23-91 | 1991 | ⊗ | | | | | | | | ⊗ | |
| **Nombre** | | **9** | **8** | **5** | **9** | **4** | **6** | **10** | **6** | **9** | **7** |

*Tableau 6.12 : Sélection (⊗) des descripteurs pour les 10 séparations*

Après la sélection des descripteurs, il faut déterminer l'architecture de 10 réseaux de neurones.

#### 6.3.3.2 Sélection de l'architecture des réseaux de neurones

La procédure automatique de sélection de l'architecture de réseaux de neurones considère un réseau de neurones avec 20 neurones cachés, puis élimine les neurones dont la contribution est inférieure à celle d'un "neurone aléatoire".

Les architectures des réseaux de neurones (à une couche cachée) qui traitent chacun des 10 séparations des classes 2 à 2 sont présentées dans le tableau 6.13 :

| Réseau de neurones | Séparation des classes | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A/B | A/C | A/D | A/E | B/C | B/D | B/E | C/D | C/E | D/E |
| Entrées | 9 | 8 | 5 | 9 | 4 | 6 | 10 | 6 | 9 | 7 |
| Neurones cachées | 1 | 2 | 1 | 1 | 3 | 2 | 2 | 1 | 1 | 2 |

*Tableau 6.13 : Architecture des 10 réseaux de neurones*

### 6.3.4 Résultats

Nous rappelons que l'on réalise **100 partitions** différentes de l'échantillon de communes (80% en apprentissage et 20% en test), et on donne la moyenne (et l'écart-type) des taux d'exemples bien classés obtenus sur la **base de test**.

Avant de s'intéresser au taux d'exemples bien classés, il faut noter qu'avec le classifieur neuronal défini précédemment, aucune commune notée A (respectivement E) n'est classée E (resp. A).

Le tableau 6.14 présente les taux de classification obtenus avec ce classifieur :

| Performance du classifieur | | |
|---|---|---|
| **Erreur de classification** | **Moyenne** | **Écart-type** |
| Bien classées (écart = 0) | 54,9% | 4,3% |
| Écart = 1 (ex : B ↔ C) | 36,5% | 4,3% |
| Écart = 2 (ex : B ↔ D) | 7,4% | 2,5% |
| Écart = 3 (ex : B ↔ E) | 1,2% | 1,1% |
| Écart = 4 (ex : A ↔ E) | 0,0% | 0,0% |

*Tableau 6.14 : Résultats*

La première ligne donne la moyenne sur les 100 partitions (sur l'ensemble de test) des taux d'exemples bien classés. Les lignes suivantes présentent les moyennes des taux d'exemples classés avec un écart correspondant (un écart égal à 1 correspond, par exemple, aux communes classées B par l'expert et notées C, ou A, par le classifieur).

Les résultats bruts ne semblent pas très bons (54,9% d'exemples bien classés) ; toutefois, il faut tenir compte des remarques faites au paragraphe 6.3.1. En effet, la multiplication des experts a considérablement augmenté le "bruit" du problème. Ainsi, on ne peut s'attendre à ce qu'**un seul** classifieur reproduise parfaitement **plusieurs** notations. Il a donc été décidé de s'intéresser au taux d'exemples classés avec au plus une classe d'écart ; ce taux est égal à

91,4%. Avec ce résultat et 5 niveaux de classification, le classifieur neuronal a été jugé utilisable par les analystes de la Caisse des dépôts.

Les 36000 communes de France ont été soumises à ce classifieur ; elles sont représentées suivant un code de couleur sur la figure 6.2 :

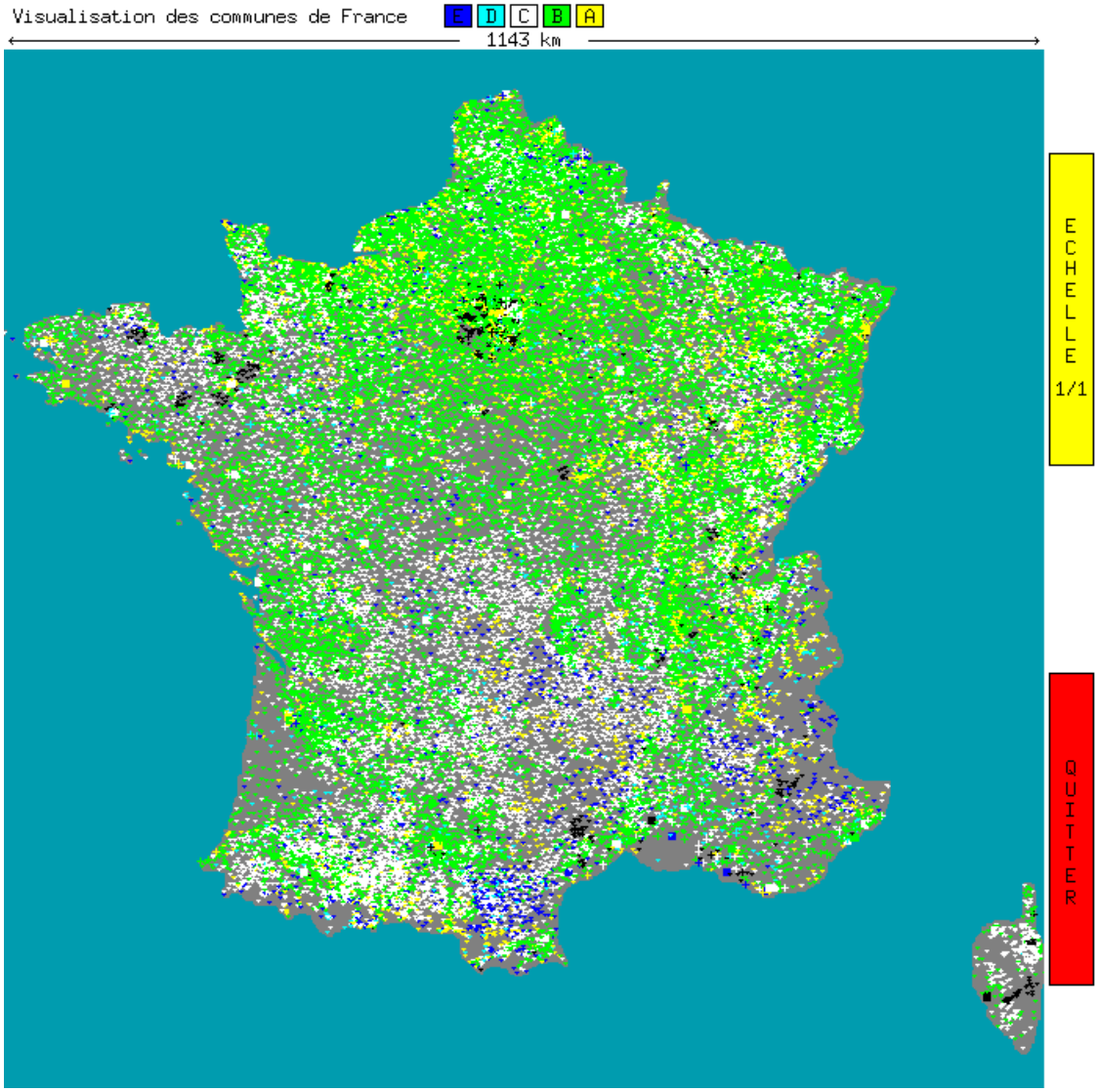


*Figure 6.2 : Carte des 36000 collectivités locales*

### 6.3.5 En résumé

Cette étude conduit aux mêmes conclusions que l'étude précédente. En effet, sur ce problème, les réseaux de neurones bien dimensionnés sont plus performants que les autres méthodes statistiques de classification.

Ce problème met en évidence la pertinence de la représentation des formes ; ici l'espace de description n'est probablement pas le plus adapté au problème posé. La difficulté intrinsèque de ce problème conduit à une limite théorique du taux d'exemples bien classés faible.

## 6.4 Conclusion

Après avoir passé en revue les principes des méthodes statistiques de classification, puis les avoir comparées sur des problèmes fictifs ; l'étude de deux applications concrètes permet de les confronter aux problèmes pratiques de la classification.

En effet, la première étape de la résolution d'un problème de classification - la constitution de l'échantillon d'exemples classés par un expert - est toujours négligée dans le cas d'un problème fictif puisque le nombre et la qualité des exemples sont ajustables. En revanche, face à un problème réel, le "prix" des exemples devient vite exorbitant (intervention d'un expert, mesure et enregistrement des caractéristiques) et la qualité des exemples peut être médiocre (intervention de plusieurs experts de sensibilité différente, mesure très bruitée des caractéristiques).

Si l'étude de cas concrets n'apporte pas d'information supplémentaire sur les qualités des méthodes statistiques de classification, elle permet de mieux cerner les caractéristiques liées à l'exploitation de ces méthodes.

L'étude, en parallèle, de problèmes fictifs et réels, conduit ainsi à une évaluation plus objective des différentes méthodes de classification. Après ce travail, nous pouvons dire que les réseaux de neurones bien dimensionnés apportent toujours une solution performante au problème posé.