

INTRODUCTION

L'évaluation de l'état de santé financière de collectivités locales ou d'entreprises constitue un problème de classification qui peut être avantageusement traité par des méthodes statistiques de classification supervisée telles que les réseaux de neurones. En effet, de nombreux travaux antérieurs ont montré que les réseaux de neurones sont, en général, de bons candidats pour la résolution de tels problèmes. Ainsi, à l'heure actuelle, les techniques neuronales ne cessent d'élargir leur champ d'application, pour deux raisons :

- d'une part, des résultats mathématiques solides ont été établis, notamment la propriété d'approximation parcimonieuse, qui expliquent pourquoi les réseaux de neurones constituent d'excellents modèles non linéaires,
- d'autre part, les techniques algorithmiques d'apprentissage ont fait de très grands progrès, grâce notamment à l'utilisation de méthodes d'optimisation non linéaires efficaces qui, associées à l'algorithme de rétropropagation pour l'évaluation du gradient, permettent des apprentissages rapides et précis.

Aujourd'hui, armés de ces propriétés, les réseaux de neurones permettent d'obtenir, lorsqu'ils sont convenablement mis en œuvre, des résultats supérieurs à ceux des méthodes classiques de modélisation et de classification non linéaire.

Bien entendu, dans la pratique, les techniques et les propriétés théoriques des réseaux de neurones n'éliminent pas complètement les nombreux pièges que l'utilisateur doit éviter :

- Le premier d'entre eux est inhérent à toutes les méthodes statistiques : en termes simples, il faut disposer de beaucoup de données afin de bâtir un modèle fiable. Les réseaux de neurones n'échappent pas à cette nécessité ; mais nous verrons qu'ils nécessitent moins de paramètres ajustables que les modèles classiques de régression non linéaire, ce qui est avantageux dès que l'on doit construire des modèles à plus de deux variables.
- Le choix des variables d'un modèle ou d'un classifieur est bien souvent déterminant pour sa qualité. Dans ce travail, nous proposons une méthode simple, et originale, de sélection des meilleures variables descriptives. Au cours de plusieurs présentations réalisées devant des auditoires d'horizons différents, elle a toujours suscité un grand intérêt car elle permet de justifier, de manière très intuitive, l'"inutilité" de certaines variables descriptives, ce qui n'est pas toujours le cas des autres méthodes de sélection de variables.
- Enfin, un autre piège réside dans la trop grande souplesse des réseaux de neurones ; en effet, les utilisateurs peuvent faire varier de nombreux paramètres tels que l'architecture du réseau, le nombre de neurones, etc. Ceci peut être considéré comme un avantage, mais conduit trop souvent à un mauvais dimensionnement du réseau et, par conséquent, à de mauvais résultats. Pour rendre l'emploi de ces

techniques plus sûr, nous proposons ici une méthode de sélection de l'architecture des réseaux de neurones.

Comme indiqué ci-dessus, l'axe principal de la recherche effectuée dans ce travail est la sélection de modèles, que nous avons appliquée aux problèmes de classification concernant l'évaluation de l'état de santé financière de collectivités locales ou d'entreprises ; ceux-ci constituent l'une des préoccupations de la filiale Informatique de la Caisse des Dépôts et Consignations. De façon plus générale, ce travail s'inscrit dans le cadre des études menées depuis plusieurs années au Laboratoire d'Électronique de l'ESPCI concernant la modélisation de processus à l'aide de réseaux de neurones.

Ce mémoire regroupe 6 chapitres, qui vont de la définition de la classification à la résolution pratique des problèmes posés par Informatique CDC.

Le chapitre 1 définit ce qu'est un problème de classification. Parmi la variété de problèmes de classification, nous analysons les caractéristiques de ceux qui peuvent avantageusement être abordés par des méthodes statistiques, notamment neuronales. Nous mettons également en évidence la nécessité de disposer de bonnes variables descriptives pour résoudre de tels problèmes de classification.

Ce travail étant consacré en grande partie à la classification, le chapitre 2 présente, d'une part, la formule de Bayes, et, d'autre part, la règle de décision de Bayes qui garantit le taux d'erreur de classification minimal. Puis, nous décrivons plusieurs méthodes de classification qui, sous certaines conditions, permettent d'obtenir des résultats proches de ceux du classifieur théorique de Bayes. Les performances de ces méthodes sont étudiées et comparées à l'aide d'un exemple de classification à une variable. Nous proposons une utilisation originale des réseaux de neurones pour l'estimation de densités de probabilités.

Le chapitre 3 présente quelques définitions relatives aux réseaux de neurones non bouclés (ou statiques) que nous utilisons dans ce travail. Puis, nous présentons les réseaux à une seule couche cachée. Ensuite, nous justifions l'utilisation de cette architecture en énonçant et commentant la propriété fondamentale de tels réseaux.

Le chapitre 4 est consacré à l'apprentissage des réseaux de neurones. Il décrit quelques méthodes classiques d'apprentissage. Nous étudions ensuite l'influence de l'ensemble d'apprentissage sur les minima locaux de la fonction de coût. Nous montrons que le nombre d'exemples d'apprentissage joue un rôle fondamental dans l'existence des minima locaux, et par conséquent, dans les performances du modèle neuronal.

Le chapitre 5 présente les méthodes de sélection de modèles qui ont pour objectif de choisir, parmi l'ensemble de modèles possibles, celui qui explique le mieux les phénomènes observés. Dans une première partie, nous présentons les bases des méthodes de sélection de modèles les plus fréquemment utilisées. Ensuite, nous présentons une procédure originale de sélection de variables. Enfin, nous appliquons cette méthode à la sélection de l'architecture d'un réseau de neurones à une couche cachée. Cette méthode originale nous permet ainsi de définir, de façon presque automatique, l'architecture finale du modèle neuronal.

Enfin, le chapitre 6 porte sur la résolution des problèmes de classification posés par Informatique CDC (analyse financière des collectivités locales et des entreprises) par les réseaux de neurones, en utilisant les méthodes de sélection de modèles présentées au cours des chapitres précédents. Dans le cas de l'analyse financière des entreprises, cette étude a débouché sur une application opérationnelle depuis 1995.

Quelques éléments du mémoire sont repris, en détail, dans les annexes A, B et C : l'annexe A présente en détail l'influence du nombre d'exemples d'apprentissage sur la fonction de coût ; le calcul de la répartition de la variable aléatoire pertinente pour la sélection de descripteurs est développé dans l'annexe B ; les notions de comptabilité nécessaires à la compréhension du travail d'analyse financière sont présentées dans l'annexe C. Un article publié dans "Industrial application of neural networks", consacré à l'évaluation financière d'entreprises, est reproduit dans l'annexe D. L'annexe E présente un article soumis à publication au congrès EUSIPCO-98 (European Signal Processing Conference) ; il concerne un dispositif embarqué de détection et de reconnaissance des défauts des rails du métro parisien. Enfin, des résultats numériques relatifs aux minima locaux des fonctions de coût sont décrits dans l'annexe F.