PNN : une nouvelle bibliothèque R pour la modélisation d'un réseau de neurones probabilistes de Specht (Rencontres R, Lyon, le 27-28/06/2013)

P.-O. Chasset [a]

[a] Chercheur indépendant
Nancy, France
pierre-olivier@chasset.net

Mots clefs : Réseau de neurones artificiels, Probabilité.

Dans le domaine de l'apprentissage automatique, l'algorithme proposé par Specht [1] présente un intérêt important. Celui-ci n'a pourtant pas encore fait l'objet d'une implémentation dans le langage de programmation statistique R. La nouvelle bibliothèque logicielle *PNN* comble cette lacune.

Considérons un ensemble d'observations représentées par des variables quantitatives réelles. Nous réalisons une classification de ces observations en plusieurs groupes. Connaissant cet ensemble d'observations et le groupe associé à chacune d'elles, nous voulons prédire le groupe d'appartenance d'une nouvelle observation. Hastie *et al.* [2] exposent plusieurs méthodes pour résoudre ce problème d'apprentissage automatique, ou plus précisément un problème d'apprentissage supervisé, car un ensemble d'observations dont le groupe est connu a été sélectionné par un superviseur. Le réseau de neurones artificiels constitue une des méthodes. Fondée sur une analogie avec le réseau que forment les neurones du cerveau, cette méthode s'est montrée particulièrement adéquate pour toute une série de problèmes dont l'aide à la décision, la reconnaissance de régularités et la classification. De la même manière que le cerveau adapte sa structure en fonction des apprentissages, le réseau de neurones artificiels nécessite une phase préalable d'apprentissage visant à adapter ses paramètres en fonction des observations sélectionnées par le superviseur. La technique d'adaptation des paramètres la plus commune est la rétropropagation du gradient. Bien que les réseaux de neurones artificiels constituent une méthode statistique d'excellence, cette technique d'adaptation souffre cependant de la nécessité d'un nombre important d'observations dont le groupe est connu et, surtout, d'un temps de calcul très important. Specht [1] résout le problème en proposant un modèle de réseau de neurones appelé « *Probabilistic neural network* » ou *PNN*, permettant un apprentissage instantané et fonctionnant même avec un faible nombre d'observations.

Le réseau de neurones de Specht [1] est conçu selon quatre couches de neurones. Seuls les neurones de deux couches adjacentes sont interconnectés. L'information transite dans un seul sens, d'une couche n à une couche $n+1$. Chaque neurone d'une couche est dédié à une même tâche. La première couche associe à chaque neurone une variable de l'observation nouvelle. Ses informations sont distribuées à tous les neurones de la seconde couche. Dans cette couche, il existe un neurone par observation apprise. Chaque neurone calcule une distance euclidienne entre l'observation nouvelle et l'observation apprise, pondérée par un paramètre de lissage. Ce paramètre permet de contrôler la finesse de généralisation de la méthode. Il évolue sur l'ensemble des réels positifs non nuls et tend vers 0 lorsque le nombre d'observations d'apprentissage tend vers l'infini. Sur cette distance est appliquée ensuite une fonction d'activation exponentielle. Ces informations sont ensuite transférées à un neurone spécifique à un groupe d'observations, de la troisième couche, qui les somme. Une quatrième et dernière couche de neurones reçoit toutes les informations spécifiques à chaque groupe et opère la prédiction de la classe.

La description de ce fonctionnement s'apparente à un réseau de neurones artificiels classique. Elle en diffère cependant sur deux éléments. Premièrement, au lieu d'avoir un nombre réduit de neurones dans la deuxième couche, la méthode utilise un neurone pour chaque observation sélectionnée pour l'apprentissage, conservant ainsi la totalité de l'information initiale. Deuxièmement, la transformation utilisée à la fin de la seconde étape est une fonction d'activation exponentielle, au lieu d'une fonction sigmoïde couramment utilisée. La particularité supplémentaire de ce réseau de neurones réside dans son fondement probabiliste. Pour une nouvelle observation donnée, le réseau de neurones, au lieu de prédire uniquement son groupe d'appartenance, estime également ses probabilités d'appartenance à chaque groupe.

Ces particularités nous permettent d'accéder à un certain nombre d'avantages. Ainsi, en effectuant la prédiction directement avec les observations sélectionnées pour l'apprentissage, l'avantage de la méthode réside dans sa capacité d'apprentissage immédiate à partir d'un faible nombre d'observations. De plus, la méthode possède une faible complexité : un seul paramètre de lissage est à calibrer. Enfin, cette méthode permet la prise en compte de la connaissance acquise préalablement en ajustant les résultats par des probabilités *a priori*.

En revanche, dans le cas d'un nombre important d'observations d'apprentissage, l'inconvénient de la méthode est son temps de calcul pour réaliser une prédiction. Il sera plus long que les autres méthodes du fait de la nécessité pour chaque prédiction d'effectuer un calcul sur l'ensemble des observations ayant servies à l'apprentissage.

Au regard des avantages et du faible nombre d'inconvénients procurés par la méthode, nous avons réalisé une implémentation de celle-ci sous le logiciel statistique R. L'installation de cette bibliothèque exporte quatre fonctions : *learn* effectue l'apprentissage à partir d'une ou plusieurs observations avec une classe connue, *smooth* détermine le paramètre optimal de lissage, *perf* calcule la performance de la méthode, et *guess* permet d'estimer la classe d'une observation, ainsi que les probabilités d'appartenance à chaque classe. Chaque fonction a fait l'objet d'un contrôle qualité : une suite de tests de fonctionnalité vérifie le bon comportement des fonctions au cours du développement et lors de l'installation. L'usage de cette bibliothèque dans sa version 1.0.1 est facilité par la mise à disposition d'un jeu de données *norms*, d'un guide utilisateur avec des exemples [3] et d'un *post* [4] relatant les différentes ressources, dont quatre tutoriels sur l'installation, l'utilisation, l'optimisation et l'évaluation de la performance d'un réseau de neurones probabilistes de Specht.

Le nouveau programme *PNN* écrit dans le langage statistique R vient ainsi compléter l'imposante bibliothèque communautaire dans le domaine de l'apprentissage supervisé avec l'implémentation d'un réseau de neurones de Specht [1]. Cette nouvelle bibliothèque logicielle, utilisable sans connaissance particulière d'optimisation ou de calibrage, incorpore toutes les méthodes nécessaires permettant une prédiction immédiate de la classe d'une nouvelle observation même à partir d'un faible nombre d'observations d'apprentissage.

Références

[1] Specht, D. F. (1990) Probabilistic neural networks. *Neural networks*, 3(1):109–118.
[2] Hastie ,T., Tibshirani, R., Friedman, J. (2008) The Elements of Statistical Learning. Data-mining, Inference, and Prediction. Springer series in statistics. Springer, Berlin, 2˝edition.
[3] http://cran.r-project.org/web/packages/pnn/pnn.pdf
[4] Chasset P.-O. (2013). *PNN: Probabilistic neural network for the R statistical language*. Software, http://flow.chasset.net/r-pnn/