



Etude technique réalisée par FOSS

Calibrations neuronales et spectroscopie proche infrarouge : étude de cas sur la teneur en protéines du blé complet

Par Lars Nørgaard*, PhD, Chef d'équipe développement chimiométrique, FOSS
et professeur affilié, Université de Copenhague
Martin Lagerholm, chimiométricien, FOSS
Mark Westerhaus, scientifique chercheur, FOSS

*auteur correspondant lno@foss.dk

Calibrations neuronales et spectroscopie proche infrarouge : étude de cas sur la teneur en protéines du blé complet

Les réseaux artificiels neuronales sont des méthodes d'étalonnage bien établies qui présentent de réels avantages pour la modélisation des bases de données complexes. Le principe de base des calibrations neuronales (ANN) sera présenté et la puissance de l'ANN sera illustrée par une étude de cas sur la prédiction de la teneur en protéines du blé complet par proche infrarouge (NIR).

Par Lars Nørgaard*, Martin Lagerholm et Mark Westerhaus, FOSS

*auteur correspondant lno@foss.dk

Introduction

Le concept des réseaux artificiels neuronales (ANN) a permis de développer des solutions et des connaissances pour une multitude de problèmes complexes depuis son introduction en tant que puissant outil d'analyse de données dans les années 1980-1990 [1]. L'efficacité des méthodes ANN est incontestable mais les méthodes sont relativement complexes en termes de mise en œuvre, de configuration des méthodes, de formation et d'estimation des paramètres par rapport aux méthodes de régression linéaires comme la régression sur composantes principales ou la régression par les moindres carrés partiels. Les méthodes susmentionnées sont couramment utilisées pour le développement d'applications qui associent les données spectroscopiques aux analyses de référence pertinentes alors que l'utilisation des méthodes ANN a été plus limitée. Nous décrivons ici le principe des ANN et appliquerons cette méthode au développement de modèles d'étalonnage mondiaux sur de grandes bases de données. Le cas étudié est la prédiction PIR Infratec mondiale de la concentration en protéines du blé complet [2].

La mise en œuvre sur ordinateur

L'inspiration du développement des réseaux de neurones artificiels trouve son origine dans les neurosciences. Les neurones du cerveau sont connectés par le biais de réseaux complexes. C'est ce concept qui a inspiré le réseau de neurones artificiels par analogie avec le réseau de neurones biologiques humain. Le cerveau humain possède des capacités de reconnaissance de modèles extraordinaire pour ce qui concerne le monde qui nous entoure. Sur la base des données sensorielles transmises au cerveau, des décisions sont prises et des modèles sont distingués et reconnus avec un degré de confiance élevé. Par ailleurs, il est très difficile pour le cerveau humain d'extraire des informations qualitatives et surtout quantitatives concernant la teneur en protéines du blé complet à partir des spectres NIR ! Sur la Figure 1, les spectres NIR de 100 échantillons de blé complet sélectionnés aléatoirement sont illustrés. Les spectres sont colorés selon la teneur en protéines et il est évident qu'il est difficile pour la plupart des cerveaux humains d'extraire des informations quantitatives et précises. Dans ces cas-là, les réseaux de neurones artificiels feront un excellent travail.

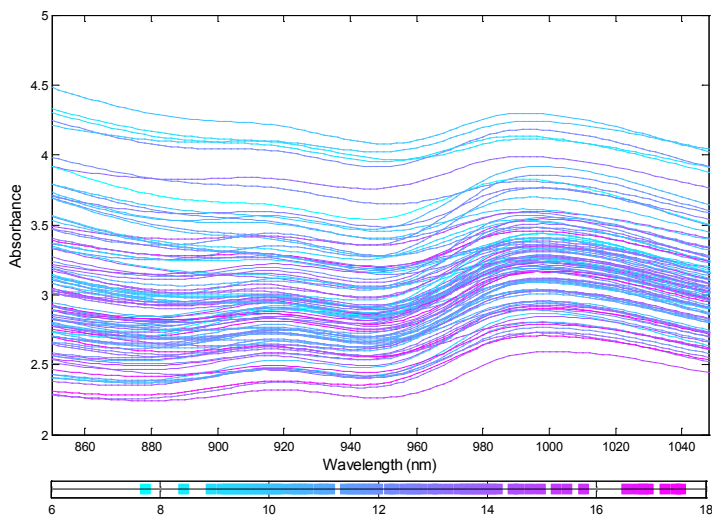
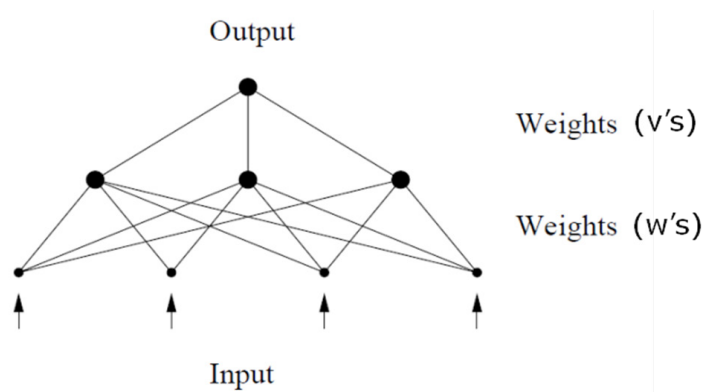


Figure 1. Cent spectres PIR sélectionnés aléatoirement d'échantillons de blé complet colorés selon la teneur en protéines allant de 7,74 % à 17,54 %.

Conception du réseau

Un réseau de neurones peut avoir une multitude de conceptions ; par exemple, nous allons nous concentrer sur le réseau à deux couches à propagation avant qui inclut a) des entrées, b) une couche cachée et c) une couche de sortie - voir Figure 2 (les entrées ne comptent pas comme couches dans la terminologie ANN). Les neurones d'entrée - première couche de neurones ou d'unités - représentent simplement les valeurs spectrales enregistrées ; dans le cas des spectres de transmission NIR des grains de blé complet, nous disposons d'enregistrements pour tous les 2 nm de 850 nm à 1048 nm correspondant à 100 données d'entrée. Différentes méthodes de pré-traitement peuvent être appliquées avant l'incorporation des spectres dans le réseau ; nous n'allons pas les décrire ici, mais soulignons qu'un pré-traitement intelligent est essentiel pour obtenir des modèles solides et performants. Chez FOSS, nous utilisons une étape de pré-traitement mathématique propriétaire avant le développement des étalonnages par réseau de neurones artificiels.



Output : Sortie ; Input : Entrée ; Weights (v's) : Pondérations (v's) ;
Weights (w's) : Pondérations (w's)

Figure 2. ANN à propagation avant simplifié où chaque cercle représente un neurone ; dans cet exemple, trois neurones cachés sont utilisés et seulement quatre entrées (représentant les variables spectrales) avec une sortie (protéines). Toutes les connexions sont dirigées de l'entrée vers la sortie [3].

Les points de données d'entrée, par ex. les mesures spectrales ou pré-traitées, sont combinés par poids exactement comme dans un modèle de régression linéaire :

$$\begin{aligned}x_1 * w_{11} + x_1 * w_{12} + \dots + x_{100} * w_{1100} &= \text{ihn1} \\x_1 * w_{21} + x_1 * w_{22} + \dots + x_{100} * w_{2100} &= \text{ihn2} \\x_1 * w_{31} + x_1 * w_{32} + \dots + x_{100} * w_{3100} &= \text{ihn3}\end{aligned}$$

où ihn1 signifie l'entrée dans le neurone caché numéro 1 et l'indice supérieur w_{2100} signifie le 2ème neurone caché tandis que 100 comme indice inférieur signifie le chiffre de pondération 100. La non-linéarité est généralement traitée par la couche cachée entre les données d'entrée (spectres) et la couche de sortie (protéines). Supposons que nous avons trois neurones cachés dans notre configuration. Nous allons donc calculer ihn1 , ihn2 et ihn3 . Chacun d'entre eux sera transformé par le biais d'une fonction non-linéaire qui est la clé de la modélisation non-linéaire d'un modèle ANN ; en sélectionnant une fonction non-linéaire appropriée, le degré de non-linéarité nécessaire sera régulé par cette fonction de transfert. Sur la Figure 3, un exemple de fonction de transfert non-linéaire sigmoïde est illustré ; une multitude de fonctions de transfert non-linéaires peuvent être utilisées à des fins différentes afin de fournir des solutions optimales dans des applications particulières.

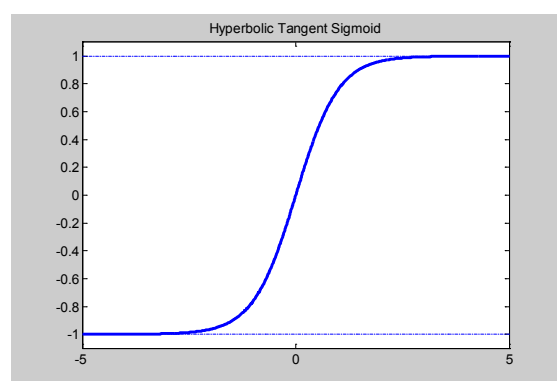


Figure 3. Une fonction de transfert non-linéaire - la fonction sigmoïde à tangente hyperbolique.

La sortie des neurones cachés, ohn1 à ohn3 , est ensuite pondérée linéairement afin de fournir une « couche entrée vers sortie » conjointe

$$\text{ohn1} * v_1 + \text{ohn2} * v_2 + \text{ohn3} * v_3 = \text{« couche entrée vers sortie »}$$

Cette somme pondérée peut passer par une fonction linéaire ou une fonction non-linéaire afin d'augmenter encore davantage la flexibilité de la modélisation non-linéaire. En principe, des biais sont également inclus dans les équations ci-dessus mais, pour des raisons de simplicité, nous les avons omis dans la description.

Quantification des protéines de blé complet avec l'ANN

Nous allons maintenant développer un modèle de prédiction ANN pour les protéines des grains de blé complet. Les données sont issues de la base de données de blé et d'orge de FOSS.

Base de données et jeux de données

Il est important de disposer d'une vaste gamme de valeurs de protéines et de teneur en eau. Pour l'obtenir, des données ont été recueillies pendant 25 ans et aujourd'hui, la base de données de FOSS compte plus de 50 000 échantillons. En outre, 25 ans de larges variations saisonnières, géographiques, de type de grain, d'instrument, de dispersion spectroscopique et de température sont inclus. Il est extrêmement important d'avoir autant de variations pertinentes que possible représentées dans votre base de données d'étalonnage afin d'obtenir des prédictions solides et précises et d'assurer la stabilité dans le temps.

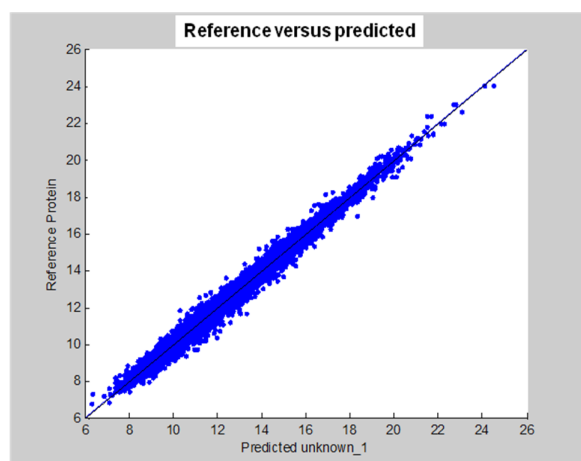
Dans le cas présent, nous nous concentrerons sur les protéines et la plage couverte est de 7 à 24 % de protéines. Le jeu de données d'étalonnage possède les propriétés suivantes [4]

- Plus de 40 000 échantillons de blé
- Les échantillons sont recueillis dans le monde entier depuis 25 ans
- La variation de température de 0 à 40 °C a été ajoutée de façon contrôlée
- De nombreuses variétés différentes sont incluses
- Variations d'instruments avec des Infratec de toutes les générations

Afin d'estimer la performance du modèle d'étalonnage ANN mondial, les prédictions d'un ensemble de validation mondial indépendant sont évaluées. L'ensemble de validation se compose de 11 908 échantillons.

Développement de modèles et prédiction

Un ANN est formé sur plus de 40 000 échantillons de la base de données afin d'obtenir le paramétrage optimal de toutes les pondérations du réseau. Comme décrit, il s'agit d'une tâche complexe mais qui est facilement prise en charge par l'ordinateur. Lorsque l'ANN est formé, les pondérations sont fixes et nous avons un modèle ANN dans lequel nous connaissons tous les paramètres. Il est désormais possible d'effectuer la prédiction des protéines pour un nouvel échantillon. Le spectre est enregistré (50 secondes) et incorporé dans l'architecture ANN avec les pondérations formées. Par pré-traitement, multiplication, addition et transferts de fonction non-linéaire de ces données d'entrée, nous obtenons une estimation de la teneur en protéines de l'échantillon analysé. Sur la Figure 4, le modèle est évalué sur l'ensemble d'essai indépendant couvrant toutes les variations possibles. Le RMSEP est de 0,27 % et le biais est de 0,011 %. Le modèle est très solide en termes de saisons, régions, températures, instruments et types de grains. Les modèles linéaires ne sont pas en mesure de traiter ces données non-linéaires et donneront des performances nettement moins bonnes [2].



Reference versus predicted	Référence par rapport à la prédiction
Reference Protein	Protéines de référence
Predicted unknown_1	Prédiction inconnue_1

Figure 4. La prédiction de l'ensemble d'essai indépendant comparée aux valeurs de référence telles qu'obtenues par chimie par voie humide.

Comme évoqué dans un précédent Gros plan sur la chimiométrie, la combinaison d'un Infratec, d'un étalonnage ANN et d'une base de données complète surpasse la précision de la méthode d'analyse de laboratoire correspondante [5].

Formation de modèles ANN

Pour les modèles linéaires, la tâche qui consiste à estimer les coefficients de régression, par exemple au moyen d'une estimation par les moindres carrés, est assez simple et minimise les valeurs résiduelles entre les valeurs de référence du laboratoire et les valeurs prédites. Dans un modèle linéaire, nous estimons 100 coefficients de régression plus un décalage - au total 101 coefficients. Dans la modélisation, la partie d'estimation est plus complexe ; dans le cas présenté, nous avons besoin de 100×3 neurones (w 's) + 3×1 (v 's) (à l'exception des biais) en plus de la complexité d'estimation des pondérations par rapport à plusieurs couches et fonctions non-linéaires. C'est le prix à payer pour avoir un modèle capable de prendre en charge la non-linéarité, et l'estimation des coefficients peut être effectuée par des méthodes d'optimisation issues d'une analyse numérique, par exemple par la méthode du gradient ou des méthodes modernes et personnalisées plus avancées.

Lors de la formation d'un réseau, l'objectif est de parvenir à ce que le réseau réponde à un spectre d'entrée donné de façon intelligente. Il y a deux états : a) un état de formation ou d'apprentissage où l'ANN apprend comment répondre et b) un état de prédiction où le réseau est appliqué à un spectre qu'il n'a jamais vu auparavant. Dans cet état de prédiction, aucune correction n'est apportée aux pondérations. La formation et l'apprentissage ont lieu en comparant la sortie des protéines du réseau à la valeur des protéines de référence, puis en modifiant les pondérations pour se rapprocher de la valeur de référence. Par conséquent, le réseau apprend par l'exemple à la manière du cerveau. L'apprentissage se produit simplement en modifiant les pondérations de façon systématique.

Outro

Les réseaux de neurones artificiels sont très efficaces pour l'extraction des informations quantitatives issues des grandes bases de données spectroscopiques où la non-linéarité est inhérente en raison de variations biologiques, environnementales et instrumentales complexes. La base d'un modèle ANN performant est une base de données couvrant toute les variations appropriées, ainsi que les modèles ANN et les outils de modélisation efficaces et optimisés. La combinaison de l'ANN et d'une base de données bien équipée offre une meilleure solidité, une stabilité améliorée, et surtout, une plus grande précision.

Références

- [1] CM Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995
- [2] NB Büchmann, H Josefsson, IA Cowe, Performance of European Artificial Neural Network (ANN) Calibrations for Moisture and Protein in Cereals Using the Danish Near- Infrared Transmission (NIT) Network, Cereal Chemistry, 78(5), 572–577, 2001.
- [3] M Lagerholm, Resource Allocation with Potts Mean Field Neural Network Techniques, thèse de doctorat, Université de Lund, 1998.
- [4] T Nilsson, Comparaison de l'étalonnage ANN mondial PIR de FOSS avec les méthodes de référence : étude pan-européenne sur cinq ans, Livre blanc de FOSS, <http://www.foss.dk/-/media/Files/Documents/IndustrySolution/Papers/Grain/PanEuroStudy.pdf.ashx>, octobre 2011.
- [5] R Malm, M Westerhaus, L Nørgaard, Le PIR et la chimiométrie peuvent-ils être plus précis que la méthode de référence ? Une étude des étalonnages ANN mondiaux pour l'analyse des grains, InFocus 2012, numéro 2.

FOSS

Foss Allé 1

DK-3400 Hilleroed

Denmark

Tel.: +45 7010 3370

Fax: +45 7010 3371

info@foss.dk

www.foss.dk