

PLANS D'EXPÉRIENCE PAR COMITÉ DE MODÈLES NEURONAUX

Nicolas Gilardi & Abdelaziz Faraj

*Division Technologie, Informatique, Mathématiques Appliquées,
Institut Français du Pétrole*

1 & 4 avenue Bois Préau, 92500 Rueil-Malmaison, France

tél : +33.1.47.52.55.29 ; fax : +33.1.47.52.70.22

Résumé

Dans cet article, nous présentons une méthode de construction de plan d'expérience pour réseaux de neurones de type Perceptron Multi Couche. Nous nous plaçons dans le cas où l'objectif est de modéliser avec un minimum de mesures un phénomène pour lequel l'information *a priori* est quasi inexistante. La méthode est basée sur celle de la "Requête à un Comité" (Query By Committee ou QBC). Deux autres approches sont utilisées à titre de comparaison : une sélection purement aléatoire des points de mesure et une sélection D-Optimale.

Mots clefs : plans d'expérience non linéaires, réseaux de neurones, Query By Committee, apprentissage actif.

English Abstract : In this paper, we present a way of constructing design of experiments for neural networks model such as Multi-Layer Perceptron (MLP). We are trying to solve the problem of modeling a phenomenon with a minimum of measurements and almost no *a priori* knowledge. Our method is based on Query By Committee (QBC). Two other approaches are used for comparison : a random selection of experiments and a D-Optimal one.

Keywords : non linear design of experiments, neural networks, Query By Committee, active learning.

1 Introduction

Comment modéliser au mieux un phénomène pour lequel on ne possède pas ou peu d'information et dont la mesure est coûteuse, soit en temps, soit en argent ? Ce genre de question se pose souvent dans les sciences expérimentales et le milieu industriel. Formalisés par J. Kiefer et J. Wolfowitz (1959), les plans d'expériences optimaux ont ouvert la voie à la résolution de ce problème. Cependant, cette approche est longtemps restée limitée à des modèles linéaires dans leurs paramètres et dont la structure était supposée connue. D. McKay (1992) et D. Cohn (1994) proposèrent d'étendre les plans d'expérience à des modèles de type réseau de neurones. L'idée principale consiste à s'approcher d'un plan optimal de façon itérative, en linéarisant le modèle neuronal pour se replacer dans une démarche "classique" de plan d'expérience. Mais là encore, l'architecture du réseau doit être fixée *a priori*, ce qui est particulièrement délicat. De plus, l'approximation au premier ordre peut-être très éloignée de la réalité du modèle.

La méthode que nous présentons ici cherche à s'affranchir de l'hypothèse sur la structure du modèle. Pour cela, nous nous basons sur la méthode de Requête à un Comité (Query By Committee ou QBC), proposée pour des problèmes de classification par H. Seung et al. (1992), et étendue à la régression par A. Krogh et J. Vedelsby (1995). Nous partons également

du principe que nous ne possédons aucune information *a priori* sur le phénomène étudié, mis à part son domaine expérimental, le fait qu’il soit modélisable avec un réseau de neurones de type Perceptron Multi-Couche (Multi-Layer Perceptron ou MLP), et la notion plus ou moins arbitraire de ce qu’est un modèle “performant” dans le cadre de l’étude.

Dans un premier temps, nous allons décrire la méthode de la QBC et la façon dont nous l’utilisons. Ensuite, nous présenterons le protocole expérimental et les données utilisés pour comparer la QBC à une approche aléatoire et à une approche D-Optimale. Enfin, nous présenterons les résultats obtenus par les trois méthodes, en terme de taille d’ensemble d’entraînement et d’erreur de test. Nous concluons alors sur l’intérêt de la méthodologie proposée et ses possibles améliorations.

2 Query By Committee

Dans le domaine de l’Apprentissage Statistique, il est généralement admis que la répartition des mesures dans le domaine expérimental se doit d’être purement aléatoire et uniforme. Le nombre de mesures doit, quant à lui, être de taille importante. C’est la façon la plus simple d’obtenir une bonne représentation de la distribution des valeurs à prédire. Cependant, quand la mesure a un coût important, une telle approche devient difficile à mettre en oeuvre. S’inspirant des travaux sur les plans d’expérience linéaires, le domaine de l’Apprentissage Actif (Active Learning) s’est développé avec pour objectif l’élaboration de méthodes de sélection de mesures d’entraînement adaptées aux algorithmes d’apprentissage.

Parmi les nombreuses techniques élaborées depuis, H. Seung et al. (1992) ont développé la méthode QBC.

2.1 Idée Générale

On cherche à modéliser une fonction $f : \mathbb{R}^p \mapsto \mathbb{R}$ pour laquelle on dispose d’un ensemble de n mesures $\mathcal{Z} = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ que nous désignons par *données d’entraînement*. L’objectif est de maximiser l’information de cet ensemble en l’enrichissant de nouvelles données, choisies dans le domaine expérimental \mathcal{X} . De cette façon, on espère obtenir une bonne estimation de f au prix d’un minimum de mesures. Les nouveaux points doivent être choisis sans connaissance *a priori* de leur mesure. Il faut donc s’appuyer sur d’autres critères.

La solution proposée par la QBC consiste à construire m modèles $\hat{f}^{(1)}, \dots, \hat{f}^{(m)}$ à partir de l’ensemble d’apprentissage initial \mathcal{Z} . Ils constitueront notre *comité*. Ces modèles doivent avoir des différences initiales de façon à explorer diverses facettes des données d’entraînement. Une fois construits selon une méthode d’apprentissage classique (voir par exemple C. Bishop (1995)), ils peuvent être utilisés pour prédire la mesure de n’importe quel point du domaine expérimental \mathcal{X} .

On va donc fournir à ce comité de m modèles un certain nombre de points candidats. Pour chaque point $\mathbf{x} \in \mathcal{X}$, nous aurons donc m estimations de la mesure : $\hat{f}^{(1)}(\mathbf{x}), \dots, \hat{f}^{(m)}(\mathbf{x})$. Si tous les modèles prédisent à peu près la même valeur pour $f(\mathbf{x})$, cela signifie qu’ajouter ce point à \mathcal{Z} n’aura pas une grande incidence sur l’entraînement du comité. En revanche, si les modèles sont en désaccord complet sur cette estimation, cela implique que \mathbf{x} est situé dans une région de \mathcal{X} qui n’est pas bien décrite par les données d’entraînement. Mesurer $y = f(\mathbf{x})$ et ajouter cette nouvelle donnée à \mathcal{Z} devrait donc significativement améliorer notre connaissance du phénomène.

Le principe de la QBC est donc de confronter les prédictions des modèles du comité et d’ajouter à l’ensemble d’entraînement les points ayant soulevé le plus grand désaccord. Diverses méthodes peuvent être utiliser pour quantifier ce désaccord $D(\mathbf{x})$. La plus intuitive

a été proposée par A. Krogh et J. Vedelsby (1995). Elle consiste à calculer la variance des estimations :

$$D(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m (\hat{y}^{(i)} - \bar{y})^2 \quad (1)$$

où \bar{y} représente la moyenne des estimations $\hat{y}^{(i)}$. Plus cette variance est élevée, plus le point devrait apporter d'information au comité.

2.2 Mise en oeuvre

Dans le cas d'un comité de MLP, la surface de désaccord définie par D sur \mathcal{X} est lisse. On peut donc choisir de nouveaux points en recherchant ses maxima locaux. Pour cela, on tire aléatoirement un ensemble de points répartis aléatoirement dans \mathcal{X} . Ils seront les points de départ d'optimisations cherchant à maximiser $D(\mathbf{x})$. Les maxima locaux ainsi trouvés sont ensuite triés afin de ne conserver que les plus importants. Cette décision se fait en fonction de la valeur maximale identifiée de D , et du nombre maximum de points que l'on souhaite conserver. On s'arrangera également pour que des points trop près les uns des autres soient fusionnés en un seul.

L'avantage de la QBC par rapport à une approche de type "plan d'expérience" est qu'elle se base sur le comportement statistique de l'apprentissage des modèles du comité, et non sur une structure de modèle prédéterminée. A terme, le désaccord entre les modèles doit s'annuler ou au moins se stabiliser. Si le comité était suffisamment grand et les modèles suffisamment prédictifs, l'ensemble d'entraînement ainsi obtenu devrait donner de bons résultats pour tout modèle entraîné de la même façon que ceux du comité. Cette liberté est très appréciable, par exemple lorsque l'apport de mesures supplémentaires amène à devoir réviser l'idée qu'on se faisait du phénomène.

3 Expériences

Notre objectif dans cet article est de construire un ensemble d'entraînement le plus petit possible permettant d'atteindre une qualité de précision acceptable. Nous allons illustrer cette approche en la comparant à deux autres : une sélection aléatoire, et une sélection D-Optimale de nos données d'entraînement.

Les trois approches seront évaluées sur un ensemble de données en deux dimensions, dont les entrées et les sorties ont été normalisées (moyenne nulle et variance 1). Nous possédons une grille de mesures suffisamment dense pour pouvoir simuler des données continues par interpolation. La procédure de comparaison se déroulera ainsi pour chacune des trois méthodes :

1. A partir d'un ensemble d'entraînement fourni, on génère un plan d'expérience selon la méthode étudiée.
2. Les points mesurés sont ajoutés à l'ensemble d'entraînement.
3. Un MLP est entraîné sur ces données et utilisé pour évaluer un ensemble de test.
4. On retourne en 1 avec le nouvel ensemble d'entraînement.

Dans notre cas, l'ensemble de données initial est identique pour les 3 méthodes. Il contient 40 points tirés aléatoirement du domaine expérimental, selon une méthode d'échantillonnage par hypercube latin. La procédure se poursuit jusqu'à ce que l'ensemble d'entraînement contienne environ 80 points.

L'évaluation des performances des différentes méthodes est basé sur deux critères. Tout d'abord, la qualité de l'estimation basée sur un plan optimal doit être meilleure ou similaire

à celle basée sur un plan aléatoire. Ensuite, à qualité de prédiction égale, le plan optimal doit être plus petit que le plan aléatoire. La sélection aléatoire va donc nous servir de référence lors de nos comparaisons. La sélection D-Optimale représentera un “état de l’art” face auquel nous pourrions situer la QBC.

3.1 Description des Méthodes

3.1.1 Sélection aléatoire

En partant des données initiales, nous ajoutons à chaque itération un nombre de nouveaux points constant. Ceux-ci sont tirés de façon aléatoire dans le domaine expérimental, selon une loi uniforme. Le nombre de neurones cachés du modèle est choisi par cross-validation sur l’ensemble d’entraînement, puis les paramètres sont calculés par descente de gradient sur l’ensemble d’entraînement complet.

3.1.2 Plan D-Optimal

On détermine l’architecture initiale du modèle neuronal par cross-validation sur les données initiales. On détermine ensuite les paramètres du modèle sur la totalité de ces données. On calcule ensuite la matrice jacobienne J de ce modèle sur un ensemble candidat. On choisit ensuite un sous-ensemble C de J tel que $\det(C^T \times C)$ soit maximum, ce qui revient à minimiser la variance des paramètres du modèle. La taille de ce sous-ensemble est de l’ordre du nombre de paramètres du modèle (cf. J.P. Gauchi (1997)). Les points obtenus sont retirés de l’ensemble candidat et ajoutés à l’ensemble d’entraînement. Les paramètres du modèle prédictif, dont le nombre de neurones cachés a été fixé au préalable, sont calculés sur cet ensemble.

3.1.3 Sélection par QBC

On choisit la taille de notre comité dont on construit indépendamment les différents modèles. On recherche ensuite les maxima locaux de la surface de désaccord. Leur nombre varie selon la forme de la surface de désaccord et la tolérance de l’utilisateur sur le désaccord minimum à considérer. Finalement, on intègre ces nouveaux points à l’ensemble d’entraînement sur lequel le modèle prédictif est construit.

3.2 Résultats

Afin de comparer au mieux nos méthodes, nous avons entraîné 20 MLP sur chacun des plans proposés. Ensuite, nous avons calculé la racine de l’erreur quadratique moyenne pour chaque MLP. La médiane de ces erreurs est la valeur choisie pour nos comparaisons.

Les figures 1a et 1b nous montrent les résultats obtenus pour deux ensembles de données initiaux différents. Si la QBC semble se comporter correctement dans les deux cas, on ne peut pas en dire autant de la D-Optimalité. Elle semble en effet incapable de construire un ensemble d’entraînement efficace. En fait, les résultats de l’approche D-Optimale sont tellement mauvais que l’hypothèse d’une erreur expérimentale ne peut être écartée. Nous n’irons donc pas plus loin dans la comparaison avec cette méthode tant que de nouvelles expériences fiables n’auront pas été effectuées.

Les figures 1 nous montrent également que la nature du plan initial a une influence sur les performances de la QBC. Cependant, sur les 10 ensembles testés, les plans de la QBC ont toujours réussi à atteindre des qualités de prédictions supérieures ou égales au plan

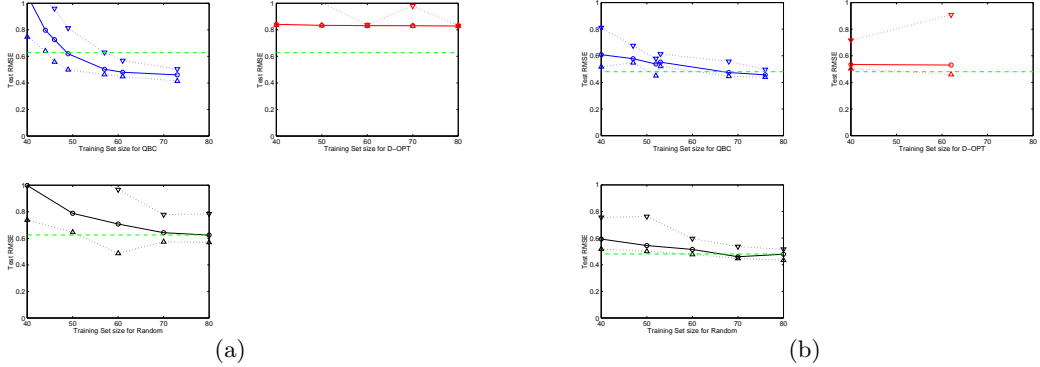


FIG. 1 – Erreur de test pour deux ensembles de données initiaux différents. L'axe des X représente la taille du plan QBC (en haut à gauche), D-Optimal (en haut à droite) et aléatoire (en bas à gauche). Les lignes pleines joignent les erreurs de test médianes pour le plan considéré. Les courbes pointillées sont les erreurs de test min et max. La ligne horizontale est l'erreur médiane du plan aléatoire à 80 points.

aléatoire à 80 points. Cela prouve que cette méthode ne détériore pas davantage un plan initial médiocre.

	a) Moyenne des erreurs	b) Moyenne des tailles
Aléatoire	0.489	80
D-Optimal	0.582	N/A
QBC	0.444	66

TAB. 1 – Résumé des 10 expériences. La colonne a) est la moyenne des meilleures erreurs de test obtenues. La colonne b) est la taille moyenne du plan nécessaire pour atteindre l'erreur de test du plan aléatoire à 80 points

Le tableau 1 est un résumé des résultats obtenus à partir des 10 plans initiaux. La première colonne fait référence au premier critère de qualité défini dans la section 3 : la qualité prédictive d'un modèle entraîné sur un plan donné. On y apprend que la QBC génère les plans les plus performants, même s'ils sont à peine meilleurs que des plans aléatoires. Il faut tout de même préciser que sur les 10 expériences, la QBC était meilleure que l'aléatoire dans 7 cas, et équivalent dans 3 cas.

La seconde colonne du tableau 1 nous montre la taille du plan que chaque méthode doit produire pour atteindre les performances prédictive du plan aléatoire à 80 points. Cette fois, les performances de la QBC sont nettement plus significatives. Les résultats de la D-Optimalité ne sont pas montrés car elle n'a fournit qu'une seule fois un plan plus performant que l'aléatoire.

4 Conclusions

Après avoir présenté le principe de la QBC, nous avons montré que cette méthode peut-être utilisée pour construire des plans d'expérience pour MLP. Les points les plus intéressants de cette méthode sont qu'aucune information *a priori* sur la structure du modèle n'est nécessaire, et qu'elle construit, à performances égales, des plans nettement plus petits qu'une approche aléatoire. Son principal inconvénient est que la procédure peut être longue lorsque

la taille du comité devient grande. Le gain en performances prédictive par rapport au plan aléatoire est également inférieur à celui escompté. Enfin, le nombre de “campagnes de mesures” peut être grand si les surfaces de désaccord ont peu de maxima locaux.

Par conséquent, cette méthode doit être améliorée. Plusieurs pistes peuvent être explorées, en particulier pour définir la surface de désaccord. Un meilleur choix de celle-ci devrait améliorer à la fois les performances prédictives et réduire la taille du plan. Les différents hyper-paramètres de la méthodes (tolérance sur la distance minimale, tolérance sur la valeur du désaccord) doivent également être étudiés plus en détail afin de trouver une relation entre leurs valeur et le nombre de nouveaux points choisis.

Pour finir, de nouvelles expériences sont nécessaires pour obtenir une comparaison fiable entre la QBC et la D-Optimalité

Remerciements

Les auteurs tiennent à remercier les participants du Consortium Neuropex pour les intéressantes discussions sur les plans d’expériences non-linéaire qui ont motivé ce travail de recherche.

Bibliographie

- [1] C. Bishop (1995) *Neural Networks for Pattern Recognition*; Clarendon Press; Oxford.
- [2] D. Cohn (1994) *Neural Networks Exploration Using Optimal Experiment Design*; *Advances in Neural Information Processing Systems* 6.
- [3] J.P. Gauchi (1997) *Plans d’Expériences Optimaux pour Modèles de Régression Non Linéaire*; *Plans d’Expériences, Applications à l’Entreprise*, Chap. 8; Drosesbeke, Fine et Saporita éditeurs, Paris.
- [4] J. Kiefer and J. Wolfowitz (1959), *Optimum Designs in Regression Problems*; *Annals of Mathematical Statistics*.
- [5] A. Krogh and J. Vedelsby (1995) *Neural Networks Ensembles, Cross Validation, and Active Learning*; *Advances in Neural Information Processing Systems* 7; Cambridge MA.
- [6] D. MacKay (1992) *Information-based Objective Functions for Active Data Selection*; *Neural Computation* 4.
- [7] H. Seung, M. Opper, and H. Sompolinsky (1992) *Query By Committee*; in *proceedings of the Fifth Workshop on Computational Learning Theory*; San Mateo CA.