Détection et traitement de "données à problèmes"

Laurent Bougrain

Détection et traitement de "données à problème"

Laurent Bougrain
LORIA-INRIA lorraine,
campus scientifique B.P. 239, 54506 Vandoeuvre-lès-Nancy
e-mail : bougrain@loria.fr

Résumé

Les performances des systèmes se résument bien souvent à une mesure statistique moyenne. Mais l'erreur commise peut varier fortement en fonction de la forme présentée. Ainsi, quelques situations difficiles à traiter peuvent dégrader fortement les performances. Partant de la distribution normale des erreurs généralement produite par les réseaux neuromimétiques, une détection et un traitement particulier sont appliqués aux données qui génèrent les plus fortes erreurs dans un problème de régression afin d'améliorer les performances.

Mots-clés discrimination, réseaux de neurones, régression, téléphonie mobile.

1 Introduction

Les erreurs produites par les systèmes de modélisation ne sont pas uniformes. Même sous l'hypothèse d'une distribution des erreurs normale, les mesures habituelles comme la moyenne des erreurs carrées, la somme ou l'écart type ne rendent pas compte de l'erreur à laquelle on peut s'attendre. Un petit nombre de données, très bruitées, mal acquises ou simplement particulières peuvent engendrer de fortes erreurs et rendre les mesures moins pertinentes. C'est la détection et le traitement de ces données à problème qui sont présentés dans cet article. De cette manière, les performances du système devraient être améliorées, et leurs mesures rendues plus signifi catives.

L'intérêt d'effectuer un apprentissage adapté aux données a été étudié dans des travaux antérieurs [Bougrain et Alexandre, 1999c]. Dans cette étude, le regroupement des données réelles ne se base pas sur un critère de similitude des formes d'entrée mais sur les erreurs qu'elles produisent. Dans la section 2, un modèle prédictif global est appliqué à l'ensemble des données. Section 3, nous décrivons la procédure de partionnement des données en deux corpus. Le premier contient les formes qui produisent une forte erreur, le second celles qui engendrent une faible erreur. Dans la section 4, nous cherchons à discriminer les formes après apprentissage d'un classifi eur neuronal. Cette partie correspond à la phase de détection. Section 5, nous évaluons les performances des réseaux de prédiction spécialisés comparativement au modèle global. Finalement nous concluons cet article en rappelant les principes d'application d'une telle méthode.

L'application test est un problème réel complexe de prédiction de l'atténuation du champ radioélectrique pour la téléphonie mobile. L'atténuation radioélectrique se mesure en décibels (dB). On dispose de 45000 données qui contiennent chasune 32 paramètres d'informations relatives à l'émetteur, au récepteur et au sursol rencontré de l'un à l'autre. Cette application présente l'intérêt d'avoir déjà été abordée sous de nombreux aspects par diverses méthodes d'optimisation [Bougrain *et al.*, 1998; Bougrain et Alexandre, 1999c; Bougrain et Alexandre, 1999b;

Bougrain et Alexandre, 1999a]. Les résultats auxquels la méthode est comparée sont parmi les plus performants.

2 Modèle prédictif général

Un perceptron multicouches (MLP), largement décrit et utilisé dans la littérature [Hertz *et al.*, 1992], est utilisé pour prédire l'atténuation du champ sans distinction particulière. Présentant une architecture en trois couches de taille (32,10,1), entrainé sur un corpus d'apprentissage de 40000 données et testé sur 5000 données, le réseaux de neurone obtient les performances suivantes (tableau 1):

Performances	Apprentissage		Test	
	μ	σ	μ	σ
MLP général	4.13	5.29	4.15	5.44
Incertitude (+/-)	0.20	0.22	0.53	0.61

TAB. 1 – *Erreur sur le modèle général*

N.B. : l'intervalle de confiance est calculé par la formule suivante ([Choukri, 1987])

$$I(\alpha,N) = \frac{T + \frac{Z_\alpha^2}{2N} \pm Z_\alpha \sqrt{\frac{T(1-T)}{N} + \frac{Z_\alpha^2}{4N^2}}}{1 + \frac{Z_\alpha^2}{N}}$$

où N est la taille du corpus, T la performance, et $Z_\alpha = 1.96$ si $\alpha = 95\%$

3 Etude des fortes erreurs

3.1 Une distribution gaussienne des erreurs

On considère le modèle probabiliste de régression le plus simple dans lequel les sorties désirées $\mathbf{t}$ s'obtiennent comme la somme d'une fonction déterministe f appliquée aux paramètres d'entrée $\mathbf{x}$ et d'une variable ϵ correspondant à un bruit gaussien ($\mathbf{t} = f(\mathbf{x}) + \epsilon$. avec $\mathbb{E}(\epsilon) = 0$). Les sorties désirées sont distribuées de manière gaussienne autour de $\mathbb{E}(f(\mathbf{x}))$. On fait l'hypothèse que les variables aléatoires associées aux différents points sont indépendantes et de même loi. Dans ce cas, la mise en oeuvre du principe de maximum de vraisemblance se réduit à trouver l'estimateur des moindres carrés (ie la fonction de régression). Sous ces hypothèses les erreurs commises par la fonction de régression suivent une loi normale.[Bishop, 1995]

Donc, dans le cas d'un modèle neuronal de régression qui utilise la mesure des moindres carrés pour corriger l'erreur faite entre la prédiction et la valeur désirée, les erreurs suivent la même loi que les sorties [Lawrence *et al.*,]. En l'occurrence, si les sorties sont supposées suivre une loi normale, les erreurs suivent également une loi normale. Pour le vérifier, on effectue dans un premier temps un histogramme des erreurs obtenues dans le problème de prédiction sur le corpus de test pour avoir une visualisation de la fonction de densité (figure 1).

La distribution des erreurs s'apparente bien à une fonction gaussienne. On observe l'existence d'erreurs supérieures à 20 dB. Aussi, au delà de la simple altération des performances probables du modèle, certaines données provoquent des erreurs largement supérieures au seuil d'acceptabilité. Une erreur d'estimation supérieure à 15% signifiera dans la majorité des cas que le modèle est

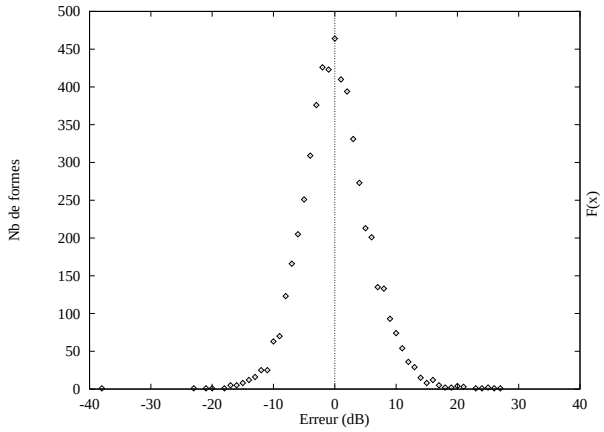

FIG. 1 – *Histogramme des erreurs sur le corpus de test*

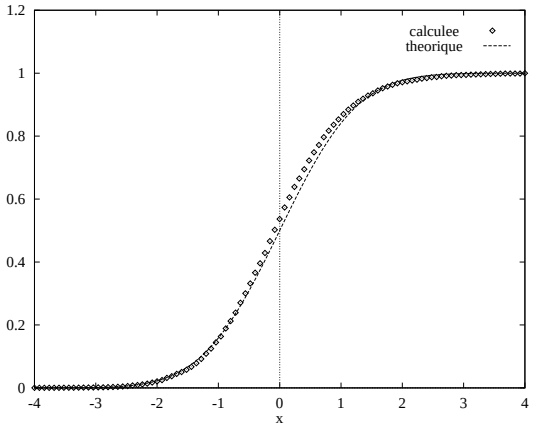

FIG. 2 – *Test de Kolmogorov-Smirnov*

imprécis. La multiplicité de fortes erreurs signifie que le modèle est peu fiable. Le test statistique de Kolmogorov-Smirnov est appliqué pour comparer la fonction de distribution empirique des erreurs à la fonction de distribution d'une loi normale. Il mesure l'écart maximal entre les deux distributions et donne une probabilité de similarité [Press *et al.*, 1992]. Ce test non paramétrique est robuste car il ne fait pas d'hypothèse sur la loi de probabilité de la variable étudiée. Puisqu'il s'applique à un petit échantillon, 100 erreurs sont extraites aléatoirement des erreurs standardisées obtenues sur le corpus de test. La probabilité que les erreurs faites sur le corpus de test suivent une loi normale est de 0.95 avec un écart maximal de 0.05 (figure 2).

3.2 Choix du seuil de séparation des données critiques

Les erreurs générées par le réseau de neurones suivent une loi normale. Il est donc intéressant de définir les grosses erreurs en terme quantitatif et non uniquement qualitatif. C'est à dire que la frontière séparatrice n'est pas une borne fixée par l'application mais par un pourcentage des données. Par exemple, il ne s'agit pas de détecter les formes qui provoquent une erreur supérieure à 10dB mais les formes qui provoquent le plus d'erreur au seuil de 10%. Les procédures de détection et de traitement des formes qui généreront potentiellement de fortes erreurs seront donc indépendantes de l'application.

Si les erreurs obtenues par le modèle suivent une loi de Laplace-Gauss et que p représente le pourcentage des données à forte erreur que l'on désire étudier alors la borne b de la région de rejet est:
$p = P(|x| < b) = P(-b < x < b) = 2 * P(0 < x < b)$.
Ainsi, si l'on veut $p = 10\%$ des données critiques, la table de la loi normale donne $1.64 < b < 1.65$. La correspondance avec la valeur seuil s de l'erreur se fait par déstandardisation de la borne. Par exemple pour un corpus d'erreurs dont les valeurs statistiques sont $\mu = 0.14$ et $\sigma = 5.29$ on a $s = 8.8\ dB$.
En pratique, la distribution des erreurs n'étant pas parfaitement gaussienne, pour une borne théorique de rejet correspondant à 10% des données, la séparation du corpus d'apprentissage affecte en fait 9.9% des données dans le corpus des formes qui génèrent de fortes erreurs.

3.3 Influence des fortes erreurs

Une fois la séparation des formes faite en deux corpus contenant pour l'un les formes pour lesquelles l'erreur de prédiction est inférieure à 8.8 dB et pour l'autre, les formes pour lesquelles l'erreur de prédiction est supérieure à 8.8 dB, on évalue avec le réseau général (section 2) l'erreur de prédiction sur chacun des corpus (tableau 2).

Performances	Apprentissage		Test	
	μ	σ	μ	σ
Faibles erreurs (90% du corpus)	3.33	4.03	3.36	4.08
Fortes erreurs (10% du corpus)	11.31	11.45	11.82	12.23

TAB. 2 – *Erreur sur le modèle généralisé*

4 Détection

Pour qu'un traitement spécifique puisse être appliqué aux données à problème, il faut que ces données soient détectées. Le modèle général de prédiction peut fournir une estimation de la valeur à prédire mais dans la phase d'application l'erreur ne pourra pas être calculée puisque la valeur désirée est inconnue. Il est donc nécesaire d'utiliser un classifieur pour orienter la donnée courante vers le bon estimateur.

Pour avoir une visualisation de la répartition des données à problème, on utilise l'algorithme de projection de Sammon [Sammon, 1969]. Cet algorithme propose une projection non linéaire dans un espace à 2 dimensions où les rapports de distance euclidienne entre les points sont au mieux conservés. Sur la figure 3, les données à problème semblent localisées dans des régions particulières de l'espace.

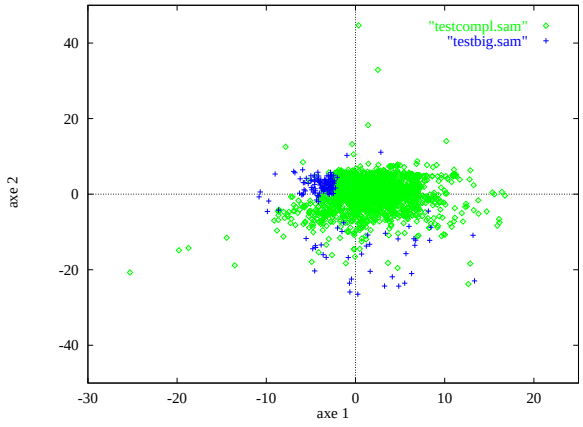

FIG. 3 – *Répartition des fortes erreurs*

Suite au partitionnement des données d'entrée, l'étiquette -1 est donnée au corpus à problème et l'étiquette 1 est donnée à son complémentaire. Un perceptron multicouches apprend la discrimination sur le corpus d'apprentisage étiqueté. D'architecture (32,10,1), le réseau apprend à l'aide de l'algorithme de rétropropagation du gradient. L'estimation de l'erreur est donnée par la formule des moindres carrés plutôt que par celle de l'entropie croisée pour des raisons de simplicité de mise en oeuvre.

Classe		supposée	
		faible erreur	forte erreur
réelle	faible erreur	99.9%	0.1%
	forte erreur	5,6%	94,4%

TAB. 3 – *Affectation du corpus d'apprentissage*

Classe		supposée	
		faible erreur	forte erreur
réelle	faible erreur	99.99%	0.01%
	forte erreur	6,3%	93,7%

TAB. 4 – *Affectation du corpus de test*

5 Modèles prédictifs spécialisés

L'idée est de spécialiser un modèle de régression sur les données qui génèrent de fortes erreurs et de réentrainer le modèle général sur le corpus épuré, ce qui revient à utiliser un second modèle. Donc, un modèle spécialise son apprentissage sur les données à problème, tandis qu'un autre le fait sur le reste des données d'après la discrimination faite par le modèle de prédiction général sur le corpus d'apprentissage.

5.1 Performances limites

Pour avoir une évaluation de l'amélioration optimale que l'on pourrait obtenir en appliquant des modèles spécialisés, on teste ces modèles sur les partitions du corpus de test constituées par le modèle général. Si les performances ne peuvent être améliorées par des réseaux spécialisés lorsque les données sont parfaitement étiquetées, il n'y a aucun espoir qu'elles le soient si les données ne sont pas parfaitement étiquetées. Sinon, l'amélioration obtenue représente la limite vers laquelle on peut tendre à condition que la discrimination soit parfaite (tableau 5).

Performances	Apprentissage			Test		
	μ	σ	p	μ	σ	p
MLP Fortes erreurs	7.80	9.29	0.099	9.10	11.41	0.094
MLP Faibles erreurs	3.36	4.12	0.901	3.41	4.19	0.906
Performances globales	3.8	5.06	1	3.95	5.56	1

TAB. 5 – *Performances limites*

N.B.: Pour rendre compte des performances globales du système, la dernière ligne du tableau exprime les mesures de performances globales en appliquant les formules suivante:
$\mu = E(X) = \sum_k P_k E(X_k)$
et $\sigma = \sqrt{\sum_k P_k(var(X_k) + E^2(X_k)) - E^2(X)}$
où P_k, $E(X_k)$ et $var(X_k)$ sont le poids, l'espérance et la variance du modèle k.

5.2 Performances réelles

Pour valider véritablement la procédure, les corpus utilisés dans la phase de test ont été obtenus par discrimination du corpus de test global en deux sous corpus, l'un contenant les formes que le classifieur estime susceptibles de générer des problèmes, l'autre contenant les formes supposées acceptables par le réseau classique.

Performances	Apprentissage			Test		
	μ	σ	p	μ	σ	p
MLP Fortes erreurs	7.80	9.29	0.099	9.06	11.27	0.088
MLP Faibles erreurs	3.36	4.12	0.901	3.44	4.26	0.912
Performances globales	3.8	5.06	1	3.94	5.51	1

TAB. 6 – *Performances réelles*

La discrimination étant très bonne (tableaux 3 et 4), le partitionnement des données obtenu par le classifieur et quasi identique à celui établi par le réseau général (correspondant au cas idéal puisqu'il n'y a pas d'erreur d'affectation). Dans ce cas, on a intérêt à utiliser la discrimination faite par le modèle général. Ainsi, les résultats liés à l'apprentissage sont les mêmes dans les tableaux 5 et 6. Si l'on avait utilisé le partitionnement du corpus d'apprentissage fait par le classifieur, les résultats seraient proches de ceux obtenus mais légèrement inférieurs. Dans le cas où les performances en discrimination sont moins bonnes il pourrait être préférable d'effectuer l'apprentissage des modèles spécialisés avec les corpus obtenus par le classifieur, pour que la discrimination des données présentées en test soit la même que celle des données d'apprentissage. Les performances limites et réelles sont identiques au regard des incertitudes (un ordre de grandeur des incertitudes est donné dans le tableau 1). Les résultats obtenus avec des modèles spécialisés (tableaux 5 et 6) améliorent ceux de référence obtenus par un modèle général (tableau 1). Le modèle spécialisé sur les données à fortes erreurs permet effectivement de mieux les traiter (voir tableau 2). Par contre, le modèle général entrainé sur toutes les données d'apprentissage disponibles est un meilleur prédicteur des données qui génèrent une faible erreur, sa capacité de généralisation étant plus grande.

6 Conclusion

Certaines données génèrent de fortes erreurs comparativement à la majorité du corpus. Souvent, elles appartiennent à une région de l'espace d'entrée éloignée du barycentre (mauvaise acquisition, données bruitrées, sous corpus particulier). En définissant une procédure de détection pour produire un modèle spécifique à ces données, nous avons amélioré les performances globales du système. Cette procédure est indépendante du problème si l'on se place dans le cadre habituel de la définition d'un modèle probabiliste de régression. Les mesures de performances du système déviennent ainsi plus significatives.

Remerciement

Ce travail a été réalisé dans le cadre d'une collaboration avec le Centre National d'Etudes des Télécommunication au travers du contract n°97 1B008.

Références

[Bishop, 1995] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press-Oxford, 1995.

[Bougrain *et al.*, 1998] L. Bougrain, N. Pican et F. Alexandre. Rôle du contexte dans le modèle owe: un réseau de neurones artifi ciels utilisant des connexions axo-synaptiques. Dans *Proc. of NeuroSciences pour Ingenieur*, Munster, 1998.

[Bougrain et Alexandre, 1999a] L. Bougrain et F. Alexandre. Recurrent neural networks for mobile phone cell planning using topological information. Dans *Proc. of Engineering Applications of Neural Networks*, Varsovie, 1999.

[Bougrain et Alexandre, 1999b] L. Bougrain et F. Alexandre. Unsupervised connectionist algorithms for clustering an environmental data set: a comparison. *Neurocomputing*, 1999. Special issue on NEURAP'98.

[Bougrain et Alexandre, 1999c] L. Bougrain et F. Alexandre. Unsupervised connectionist clustering algorithms for a better supervised prediction: Application to a radio communication problem. Dans *Proc. of International Join Conference on Neural Networks*, Washington, 1999.

[Choukri, 1987] K. Choukri. Quelques approches pour l'adaptation aux locuteurs en reconnaissance automatique de la parole. Rapport Technique ENST-87E026, ENST, 1987. annexe 2.

[Hertz *et al.*, 1992] J. Hertz, A. Krogh et R. G. Palmer. *Introduction to the theory of neural computation*. Addison-Wesley, 1992.

[Lawrence *et al.*,] Steve Lawrence, A.D. Back, A.C. Tsoi et C. Lee Giles. On the distribution of performance from multiple neural network trials. *IEEE Transactions on Neural Networks*, 8(6):1507–1517.

[Press *et al.*, 1992] W.H. Press, S.A. Teukolsky, W.T. Vetterling et B.P. Flannery. *Numerical Recipes*. Cambridge University Press, Cambridge, seconde édition, 1992.

[Sammon, 1969] John W. Jr Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 5:401–409, 1969.