UNE ANALYSE EMPIRIQUE DES LIENS ENTRE CAPACITE POTENTIELLE ET EFFECTIVE DE MODELES DE RESEAUX DE NEURONES

Remus BRAD
Faculté d'Ingénièrie de l'Université de Sibiu
B-dul Victoriei 5-7, 2400 Sibiu, Romania
tel. 069 216062 ext. 410
rbrad@sibiu.ro

ABSTRACT
The paper is based on a series of studies maded on the learning capabilities of multi-layered perceptrons. The complexity of this nonlinear systems can be varied as we desire, acting for instance on the number of hidden units. In return, we will be confronted with a choice dilemma, concerning the optimal complexity of the system for a given problem. We have defined the potential size as given by the number of hidden units and the effective size as a result of network training on a given problem. By the mean of statistical methods, we have found that the effective number of hidden units is smaller than the potential size, some units have a "binary" activation level or a time constant activation. We also prove that weight initialization to small values is fully recomended and reduce the effective size of the hidden layer.

MOTS CLES: perceptron multicouches, apprentissage, rétropropagation, couche cachée

1. INTRODUCTION

Nous avons étudié une des problématiques de l'apprentissage dans les perceptrons multicouches (PMC). Ces derniers sont des systèmes non linéaire qui calculent une fonction de transfert par composition de fonctions non linéaires élémentaires (cellules). Une de leurs caractéristiques est qu'on peut faire varier la complexité de ces systèmes à volonté en agissant par exemple sur le nombre de cellules cachées. Il se pose alors le problème du choix de la complexité optimale d'un système pour résoudre une application donnée. Cette question générale est fondamentale et dépasse largement le cadre des réseaux de neurones [1].

Nous nous sommes concentrés sur l'étude du lien entre performances d'un système PMC et sa complexité déterminée par son nombre de cellules cachées. La règle générale est que un système trop simple ne réussira pas à apprendre une tâche difficile et à l'inverse, un

système trop complexe tendra à apprendre par coeur l'ensemble des informations présentes dans les données sans synthétiser de règle utile pour caractériser le phénomène. Traditionnellement, dans le monde des réseaux de neurones, on passe beaucoup de temps à trouver un compromis en terme de complexité conduisant à des systèmes performants [3].

2. DONNEES ET ARCHITECTURES

Pour bien préciser le cadre dans lequel les expériences ont été effectuées, nous nous contenterons de décrire brièvement le problème des formes d'ondes de Breiman [2]. Il s'agit d'un problème de classification défini à partir de trois formes d'ondes h1, h2 et h3, dont chacune est un vecteur réel de dimension 21. Ces trois formes sont utilisées pour construire trois classes équiprobables. Les éléments d'une classe sont définis comme une combinaison convexe de deux formes d'onde plus un bruit gaussien. Les trois classes ayant des matrices de variance différentes, les surfaces de séparation sont non linéaires. Chaque onde a été discretisée sur 21 points et donc, nos réseaux auront la même taille en entrée. Une base de 3000 exemples a été utilisée pour engendrer les ensembles d'apprentissages, par une loi de distribution uniforme. La taille de ces ensembles varie en fonction des expériences entre 300 et 1000 exemples. La base de test est composée de 5000 exemples et toutes les performances en test auront comme références cet ensemble.

Ainsi, l'architecture des réseaux employés est 21-X-3 (unité entrée - unité cachée - unité sortie), avec X prenant les valeurs 0, 5, 10, 15, 35 et 60. Les poids à partir desquels l'apprentissage débute sont obtenus par une distribution aléatoire dans l'intervalle [-1,1]. Pour certaines expériences, l'intervalle [-0.01,0.01] a été considéré.

3. LE NOMBRE EFFECTIF D'UNITES CACHEES

La complexité d'un système du type PMC est contrôlée par le nombre de cellules cachées. C'est ce que nous appellerons la *complexité potentielle* du système. Par opposition, la *capacité effective* prendra en compte non seulement le modèle mais son comportement pour une tâche donnée [5]. C'est cette dernière quantité qui nous intéressera pour déterminer le modèle le plus adéquat pour une tâche donnée. La relation entre complexité potentielle et capacité effective est dans ce cadre, cruciale à comprendre. La difficulté du problème vient du fait qu'il est difficile de mesurer ou même simplement de définir de façon universelle ces deux notions [4]. De nombreux travaux explorant différentes directions ont eu lieu sur ce thème ces dernières années. Il est donc intéressant de comprendre le rôle de ces cellules cachées.

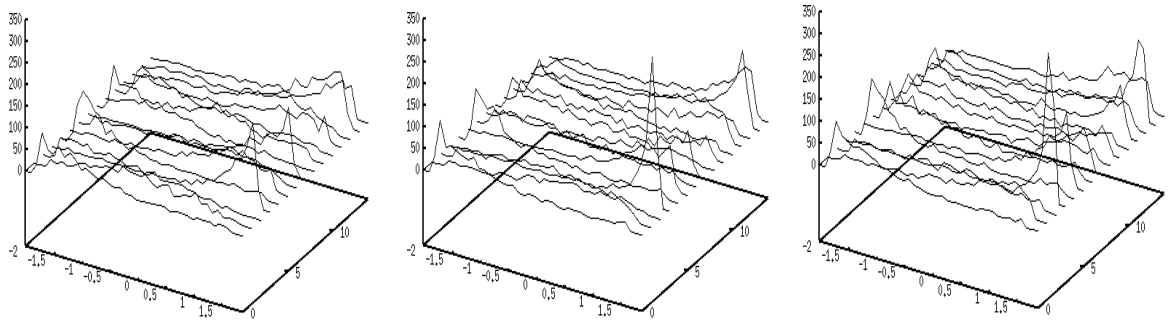
3.1. L'ACTIVATION DES UNITES CACHEES

Chaque cellule participe à la réponse générale du réseau d'une certaine manière et avec une certaine importance. Il est donc important de caractériser ce comportement et son évolution au cours du processus d'apprentissage.

Le processus d'apprentissage que nous avons utilisé débute avec les poids initialisés aléatoirement, dans l'intervalle [-1,1]. Nous allons tout au long de l'apprentissage, sauver les

valeurs d'activation des unités cachées mesurées pour un ensemble de test constitué de 1000 exemples. A partir de ces données, des histogrammes d'activation seront déduits pour chacune des unités. La figure 3.1 montre pour certaines étapes de l'apprentissage, ces histogrammes.

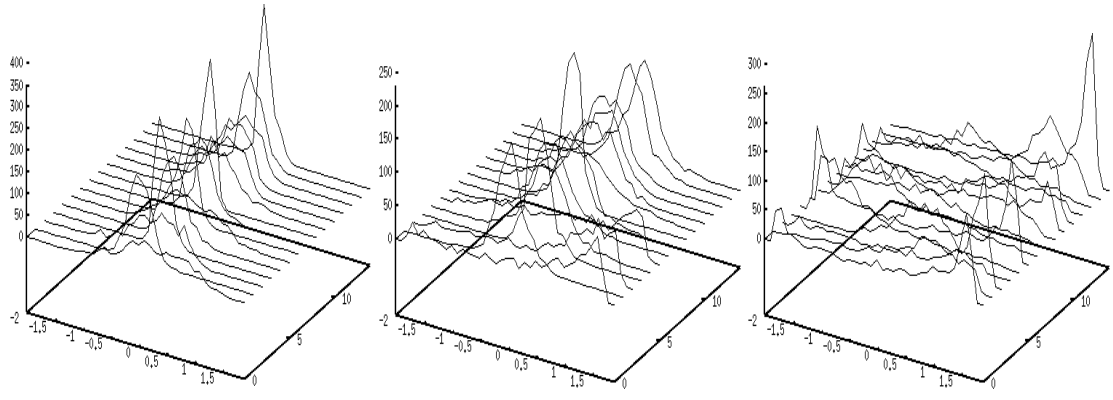
Au fil de l'entraînement, nous sommes en mesure de constater que les valeurs d'activation tendent à devenir binaires, ce qui signifie qu'une unité aura un rôle bien spécifique et non linéaire.



a) b) c)

Figure 3.1. Histogramme d'activations des u.c. après 50 (a), 200 (b) et 1000 (c) itérations, d'un réseau a 15 cellules, base de 300 ex. d'apprentissage, 1000 de test, initialisation avec poids dans [-1,1].

Nous avons refait cette expérience pour le même réseau, mais avec une initialisation beaucoup plus petite des poids, dans l'intervalle [-0.01, 0.01]. La figure 3.2 montre plus clairement la façon dont les valeurs d'activation se changent au cours de l'apprentissage et arrivent aux extrémités. La valeur centrale de 0 est prédominante après 5 pas d'entraînement, mais arrivant à 50, l'intervalle des activations croît considérablement. A partir de 200 jusqu'à 1000 pas, les valeurs ne changent pas trop, elles suivent sûrement un processus de raffinement du rôle joué dans l'apprentissage.



a) b) c)

Figure 3.2. Histogramme d'activations des u.c. après 50 (a), 200 (b) et 1000 (c) itérations, d'un réseau a 15 u.c., base de 300 ex. d'apprentissage, 1000 ex. de test, initialisation avec poids dans [-0.01,0.01].

Dans le cas des grands réseaux, comme ceux ayant 60 unités cachées, la situation est semblable, les activations des poids sont plus prononcées pour les valeurs extrêmes, mais un nombre assez important d'activations ont des valeurs dans l'intervalle possible.

Tous les histogrammes montrent que certaines cellules jouent un rôle plus actif que les autres. Les pics de grande hauteur de certaines unités signifient que la cellule a eu la même réponse tout au long du test et donc elle est invariante et ne sert à rien. Par contre, d'autres

cellules ont un spectre de valeurs plus large, elles sont donc plus "occupées" ou plus actives a participer à la réponse du réseau.

3.2. LA VARIANCE

Les histogrammes présentées ci-dessus, ne sont que des clichés pris lors de l'apprentissage de l'état du réseau. Afin de visualiser de façon plus synthétique le changement pendant l'entraînement, il est possible de caractériser chaque courbe de l'histogramme par un seul nombre dont on peut suivre l'évolution dans le temps. Il s'agit de la variance de l'unité cachée i définie par:

$$\mathrm{var}_i = \frac{1}{N}\sum_p (x_i^p - \overline{x_i})^2$$

où: p - indice sur l'ensemble de test; N - le nombre d'exemples de test; $x^p{}_i$ - activation de la cellule i à l'exemple p; x_i - moyenne sur tous les exemples de test de l'activation de la cellule i:

$$\overline{x_i} = \frac{1}{N}\sum_p x_i^p$$

La variance mesure le taux d'activation. Les figures 3.3, 3.4 présentent l'évolution dans le temps des écarts types, pour les cas discutés ci-dessus.

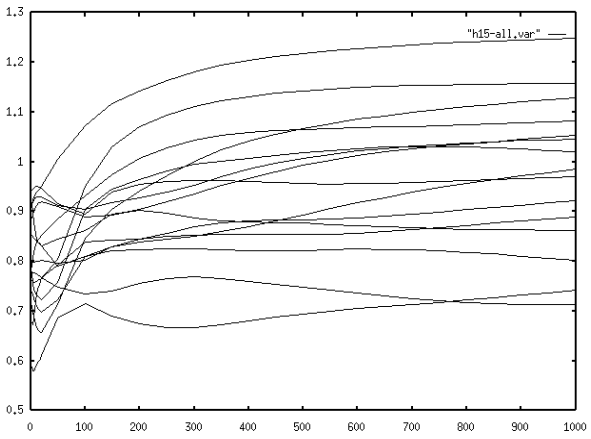


Figure 3.3 Évolution des écarts types des 15 u.c. d'un réseau, poids initiaux dans [-1,1].

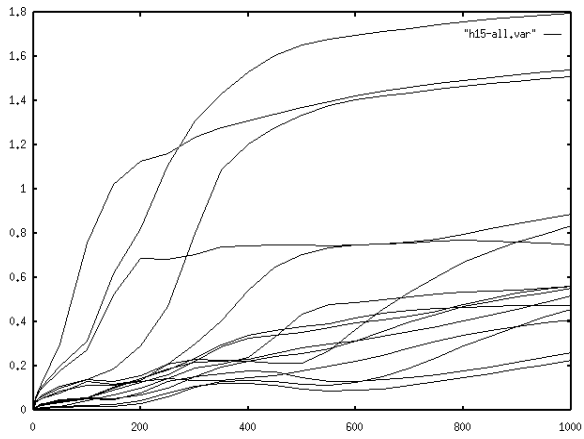


Figure 3.4 Évolution des écarts type des 15 u.c. d'un réseau, poids initiaux dans [-0.01,0.01].

En comparant la figure 3.3 avec la figure 3.4, il est tout a fait visible que l'initialisation a un effet sur les valeurs de début des écarts types, qui est propagé par l'évolution ultérieure.

Nous allons faire une analyse quantitative du phénomène, en partant de la signification de l'écart type. Si celui-ci a une valeur importante, la distribution des éléments à partir desquels il a été calculé, s'étend sur un large domaine. Les valeurs occupent des places dans un intervalle important, centré sur la moyenne [6].

En revenant à notre cas, si l'écart type d'une cellule cachée a une grande valeur, les activations tout au long du test sont bien distinctes et notre cellule participe de façon active à la réponse globale. Par contre, un écart type de dimension réduite, signifie que les activations se regroupent autour de la moyenne et la cellule se comporte comme un invariant, donc inactive.

A partir des figures 3.3 et 3.4, nous pouvons très bien distinguer les unités qui forment la dimension effective du réseau. Dans le cas des réseaux initialisés à partir de grandes valeurs, les cellules participent de façon égale à la réponse globale et il est impossible de considérer a réduire la dimension de la couche cachée. Cet aspect, comme aussi d'autres présentés au cours de l'article, nous encourageons à initialiser les réseaux à des poids de petites valeurs.

3.3. LES VALEURS PROPRES DE LA MATRICE DE COVARIANCE.

Pour avoir une image de la façon dont les cellules agissent, nous allons analyser leur comportement global et pas individuel. Dans cette démarche, notre but sera de trouver la dimension de l'espace engendre par les activations des unités cachées et par l'intermédiaire de cette valeur d'estimer la dimension effective de la couche cachée.

La réponse des cellules cachées peut être vue comme un point dans un espace n-dimensionel. La position de ces points peut-être arbitraire, mais nous pourrons considérer par exemple le cas ou la plupart d'eux sont tout au long d'une droite. Dans ce cas, le réseau pourra être remodelé avec une seule unité cachée. De la même façon, si les points seront dans un plan bidimensionel, le nombre effectif de cellules cachées serait de deux.

Pour essayer de déterminer le nombre d'axes orthogonales sur lesquelles nos points se répandront, nous allons calculer premièrement, la covariance entre deux unités, en considérant une interaction point à point. La covariance entre la cellule i et j est donne par:

$$\mathrm{cov}_{ij} = \frac{1}{N}\sum_{p}(x_i^p - \overline{x_i})(x_j^p - \overline{x_i})$$

où p - indice variant sur tous les exemples N de test; x^p_i - activation de la cellule i a l'exemple p; x_i - la moyenne des activations de la cellule i sur tout les exemples de test.

Deuxièmement, nous allons calculer la dimension effective du nuage de points donnée par le nombre de valeurs propres non-nulles de la matrice de covariance. Les valeurs propres décrivent la dispersion des point au long des axes ainsi considérées. Les figures 3.5 et 3.6 montrent l'évolution avec le temps des valeurs propres pour les exemples considérés.

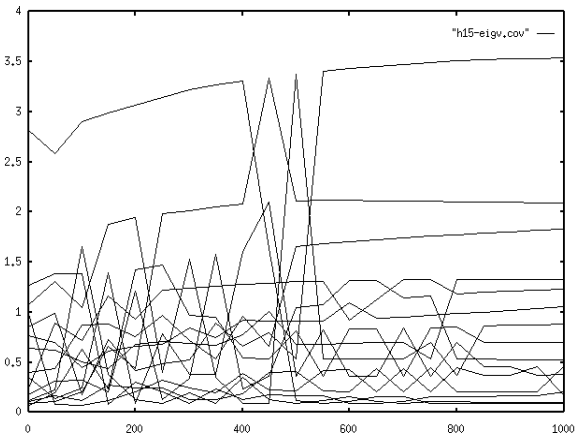


Figure 3.5 L'évolution des valeurs propres de la matrice

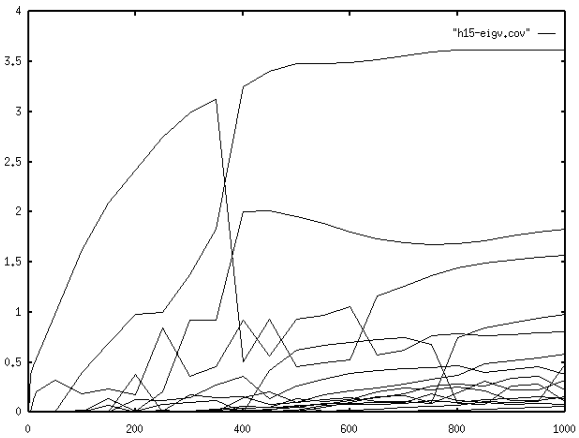


Figure 3.6 L'évolution des valeurs propres de la

matrice de covariance pour 15 u.c. et poids [-1,1]. de covariance pour 15 u.c. et poids [-0.01,0.01]

La figure 3.5 montre que pour une initialisation a grandes valeurs des poids, les valeurs propres sont importantes depuis le début, mais elles ne restent pas constantes, certaines étant en croissance d'autres en décroissance. Ce qui est important, c'est d'observer qu'après 600 pas, l'évolution des valeurs propres est constante, donc le nombre effectif unités cachées participant est stable. La même interprétation revient après l'analyse de la figure 3.6, qui a été obtenue pour des petites initialisations. Certaines valeurs "émergent" depuis le début, d'autres sont encore à zéro, le nombre de valeurs propres qui ont atteint un certain ordre de grandeur définit l'espace engendré par les réponses des unités cachées. Comme nous pouvons le constater, la dimension de cet espace est plus petit que celui définit par la taille de la couche cachée et donc le nombre effectif est plus petit que celui potentiel fixé par l'architecture.

Dans le cas du réseau à 60 unités cachées, les phénomènes sont plus difficiles a suivre, beaucoup de valeurs propres restent ou sont proches de zéro, dans ce cas aussi nous constatons une dimension effective plus petite que celle du réseau.

4. CONCLUSION

A partir des expériences effectuées, nous sommes en mesure de tirer plusieurs conclusions. Nous avons cependant presenté, lors des differents chapitres les remarques à propos des résultats obtenus. Nous pourrons donc conclure, qu'a partir de l'analyse des histogrammes d'activations, de la variance et des valeurs propres de covariance, nous avons démontré que le nombre effectif d'unités cachées et plus petit que celui potentiel, avec un ordre de grandeur faible pour les petits réseaux, mais avec un ordre plus important pour les grands. La plupart des cellules ont des activations "binaires" et ont donc, un comportement constant à la réponse du réseau. Aussi, l'initialisation des poids à des petites valeurs a un effet déterminant sur la taille effective du réseau.

REFERENCES

[1] Cibas T., Gallinari P. et Gascuel O., Experimental Investigation on the Complexity-Performance Relations in Multilayer Perceptrons, a apparaitre a ICANN'95

[2] Gallinari P. and Gascuel O., Statistiques, apprentissage et generalisation; applications aux réseaux de neurones, Support de cours, LAFORIA - Université Paris VI, 1994

[3] Moody J.E., The effective number of parameters: An analysis of generalisation and regularisation in nonlinear learning systems, Advances in Neural Information Processing Systems 4 (NIPS 91), Morgan Kaufmann, pp.847-854, 1992

[4] Weigend A.S. and Rumelhart D.E., Generalisation through minimal networks with application to forecasting, INTERFACE'91 23rd Symposium on Interface: Computing Science and Statistics, Interface Foundation of North America, pp.362-370, 1991

[5] Weigend A.S., On overfitting and Effective Number of Hidden Units, in Proceeding of the 1993 Connectionist Models Summer School, pp.335-342, 1994

[6] Weigend A.S., Time Series Analysis and Prediction, Technical Report CU-CS-744-94