# Intégration de Connaissances a Priori dans le Principe du Maximum d'Entropie

F. Chakik *** et F. Dornaika *,**

*IKERBASQUE, Basque Foundation for Science
**University of the Basque Country, San Sebastian, Spain
fadi_dornaika@ehu.es
***LaMA Laboratory, Lebanese University, Tripoli, Lebanon
fchakik@ul.edu.lb

**Résumé.** Cet article montre que si l'on dispose d'une connaissance a priori sur le problème en main, l'intégration de cette dernière dans le processus d'apprentissage d'une machine intelligente pour des tâches de classification peut améliorer la performance de cette machine. Nous étudions l'effet de l'intégration de la connaissance a priori de convexité sur le processus d'apprentissage du principe du Maximum d'Entropie (MaxEnt) en utilisant des exemples virtuels. Nous testons les idées proposées sur un problème benchmark bien connu dans la littérature des machines d'apprentissage, le problème de formes d'ondes de Breiman. Nous avons abouti à un taux d'erreur de généralisation de 15.57% qui est très proche du taux d'erreur théorique estimé par Breiman (14%).

## 1 Introduction

On désigne par "connaissances a priori" les informations auxiliaires qui peuvent être utilisées pour aider le processus d'apprentissage. Il y a des informations qui peuvent être données à priori et que l'on appelle *connaissances a priori* Wu et Srihari (2004). Notre travail ne s'intéresse pas à l'extraction des connaissances a priori dans une application donnée, mais plutôt s'intéresse à la manière de les intégrer dans un processus d'apprentissage Lauer et Bloch (2008). Le but est d'améliorer la performance de la généralisation. Afin d'intégrer ces connaissances a priori dans le processus d'apprentissage à partir d'exemples, deux méthodes sont possibles : (1) la représentation des connaissances par des exemples virtuels, qui traduisent cette dernière en un langage compréhensible par l'algorithme d'apprentissage, et (2) l'intégration des connaissances dans l'architecture du classifieur. On appelle exemple virtuel l'exemple qui proviendra d'une connaissance a priori alors qu'un exemple réel proviendrait directement de la fonction à réaliser. Après ce que cette connaissance ait été représentée par des exemples virtuels, nous pourrions mesurer la qualité de l'apprentissage en regardant la performance du système sur un ensemble de ces exemples. Comme un exemple concret sur l'intégration des connaissances a priori dans le processus d'apprentissage, on va illustrer celle concernant la connaissance a priori de convexité dans la tâche de classification des formes d'ondes de Breiman dans le cas d'un classifieur basé sur le principe du Maximum d'Entropie Buck et Macaulay (1991); Chakik et al. (2004).

## 2 Inférence statistique et principe du Maximum d'Entropie

Considérons un ensemble de données formé de vecteurs $X$ de $N$ composantes $x_i$ , $i = 1, \ldots, N$. La modélisation statistique consiste à déterminer une loi de probabilité $P(X)$ qui décrive au mieux la densité de probabilité avec laquelle les variables $X$ ont été tirées. La connaissance dont on dispose pour déterminer un modèle sera traduite par diverses contraintes empiriques sur $P(X)$. Ces contraintes sont définies à l'aide de fonctions $A_{\mathrm{n}}(X)$, $n = 1, \ldots, R$, nommées observables par analogie avec la thermodynamique, et qui refléterons par exemple des relations que l'on juge essentielles entre certaines variables. Un modèle sera d'autant plus complexe demandera d'autant plus de données que le nombre d'observables sera important. Ces fonctions permettent de traduire la connaissance de certaines caractéristiques des variables au moyen d'équations mettant en jeu leurs espérances. Une méthode de recherche concrète est celle qui se base sur le choix d'une loi de probabilité dont l'entropie est maximale, on parle donc du principe du maximum d'entropie. Le principe du maximum d'entropie est une procédure pour induire une distribution de probabilité inconnue à partir d'un ensemble partiel de connaissances (Bessière, 1990), (Buck et Macaulay, 1991). Cette procédure a trouvé un nombre croissant d'applications dans les différentes branches de la science.

Chaque fonction $A_n$ (scalaire ou vecteur) prend la valeur $A_{n\nu}$ pour l'état discret $\nu$. Les $R$ valeurs moyennes, dites théoriques, sont données par

$$\lfloor A_n \rfloor = \sum_{\nu=1}^{M} A_{n\nu} P_\nu \qquad n = 1, \ldots, R$$

où $P_\nu$ répresente la valeur de la densité pour l'état $\nu$. Nous voulons déterminer une distribution de probabilité qui vérifie un ensemble de contraintes. Ces contraintes sont telles que les valeurs théoriques sont égales à leurs estimateurs, déterminés avec l'ensemble de mesures ou de données dont on dispose. En d'autres termes, le MaxEnt est un principe de modélisation de la loi de probabilité qui vérifie les contraintes définies à partir de l'ensemble d'observables et qui maximise l'entropie. En effet, il y a plus d'une loi qui satisfait l'ensemble des contraintes ; on sélectionne celle qui correspond à la distribution la plus étalée possible, compatible avec les observables : elle ne contient pas plus d'information que celle apportée par les mesures.

Dans la procédure du maximum d'entropie, la distribution de probabilité choisie est celle qui maximise l'entropie $S(P_\nu) = -\sum_\nu P_\nu \, log(P_\nu)$ à condition que la contrainte de normalisation $\sum_\nu P_\nu = 1$ et que l'ensemble des $R$ contraintes soient réalisées Chakik et al. (2004). On pourrait montrer que la solution générale pour la densité $P_\nu$ aura la forme suivante ($Z$ est une constante de normalisation) :

$$P_\nu = \frac{1}{Z} \, exp\left[ -\sum_n \overrightarrow{\lambda}_n \cdot A_{n\nu} \right]$$

Les multiplicateurs de Lagrange $\lambda_n$ $(n = 1, \ldots, R)$ doivent satisfaire les contraintes $\lfloor A_n \rfloor \equiv \frac{\partial \log Z(\lambda_1, \ldots, \lambda_R)}{\partial \lambda_n} = \langle A_n \rangle \qquad n = 1, \ldots, R$ où le signe $\equiv$ veut dire identique, $\langle A_n \rangle$ sont les estimateurs empiriques des $A_n$, déterminés à partir des données, que l'on suppose indépendantes, identiquement distribuées. Une fois les valeurs des $\langle A_n \rangle$ connues, les $R$ équations permettent de déterminer les $\lambda_n$ qui, a leur tour, permettent de déterminer la distribution $P_\nu$.

# 3 Problème des formes d'ondes de Breiman

Ce problème est considéré comme un problème standard et classique pour la comparaison de différentes approches et méthodes d'apprentissage de la classification. Il a été introduit par Breiman et al. (1984) pour l'étude des arbres de décision en général, et pour tester la méthode CART en particulier. Ce problème consiste à discriminer entre trois classes de formes d'onde. Chaque forme d'onde simule un phénomène chronologique quantitatif mesuré sur 21 instants régulièrement espacés. C'est un objet caractérisé par un point de $\mathbb{R}^{21}$. Chaque classe consiste en une combinaison convexe aléatoire de deux des ondes de base et est perturbée par un bruit gaussien. Les ondes, notées O(1), O(2) et O(3), sont uni-modales et en déphasage (Figure 1). En l'absence du bruit gaussien, les trois classes sont représentées dans $\mathbb{R}^{21}$ par les trois côtés d'un triangle de sommets O(1), O(2) et O(3) (2). Dans ce cas, le problème sera déterministe et on peut dire que chaque exemple (représenté par un vecteur $X$ à 21 dimensions) appartient à la classe correspondant au côté auquel $X$ appartient. Le bruit gaussien vient perturber ces considérations géométriques et il rend chaque exemple admissible au sens des trois classes et le problème n'est plus déterministe. Ainsi, nous disposons de 11 ensembles d'apprentissage de 300 exemples chacun, et un ensemble de test de 5001 exemples. Notons que le principe d'Entropie Maximale implique l'estimaton d'une densité de probabilité pour chaque classe.

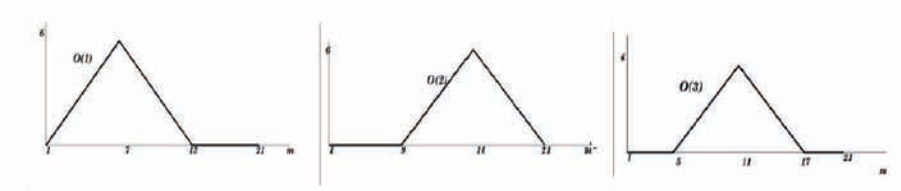


FIG. 1 – *Schématique des trois ondes de Breiman.*

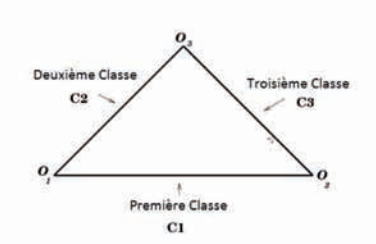


FIG. 2 – *Représentation schématique des trois classes en l'absence du bruit gaussien.*

# 4 Intégration de la connaissance a priori de convexité

Afin de mieux comprendre l'effet des connaissances a priori, nous procédons à intégrer ces dernières dans le processus de classification du MaxEnt d'une façon similaire à celle de

l'idée qui a été proposée par Abumostafa (1993) pour l'intégration des connaissances a priori dans un réseau de neurones. Alors, nous supposons qu'un expert nous a fourni une connaissance a priori sur le phénomène qui a généré les exemples et nous procédons à intégrer cette dernière soit (1) *en créant des exemples virtuels* qui soient en accord avec cette connaissance a priori, soit (2) *dans l'architecture du classifieur*. Par manque de place nous allons présenter uniquement l'apport de l'utilisation des exemples virtuels. Ceci offre au processus d'apprentissage d'apprendre cette connaissance de la même manière qu'il apprend les exemples réels. Un exemple virtuel est généré à partir d'une combinaison convexe de deux exemples (de la base d'apprentissage) de la même classe. Comme nous disposons de 11 ensembles d'apprentissage, nous avons généré les exemples virtuels correspondant à des combinaisons convexes des exemples de chaque ensemble, et puis nous avons calculé les paramètres $\lambda_n$ de chaque densité correspondant aux ensembles augmentés. L'apprentissage se fait sur ces derniers. L'erreur d'apprentissage est définie comme étant l'erreur moyenne commise sur les 11 ensembles d'apprentissage (augmentés). Quant à l'évaluation de l'erreur de test, elle est calculée sur l'ensemble de test pour chaque ensemble d'apprentissage, ensuite on détermine sa valeur moyenne ainsi que sa variance sur les 11 erreurs de test calculées.

Nous avons testé l'influence d'ajout de 20, 50, 70, 100, 200, 300, 600 et 1000 exemples virtuels à la base d'apprentissage d'origine. Ainsi, l'ajout de 20 exemples virtuels veut dire que chaque classe va intégrer 20 nouveaux exemples virtuels. Le tableau 1 donne l'erreur d'apprentissage et de test (en pourcentage) en fonction du nombre d'exemples virtuels ajoutés. Dans ce cas, il y a une densité de probabilité pour chaque classe. Chaque classe possède une seule observation $A_c(X)$ $(c = 1, 2, 3)$ qui est donnée par la distance euclidienne au carré entre le vecteur $X$ à 21 dimensions et le centre de la classe. Il y a trois $\lambda$ à estimer. **(a)** Les trois estimateurs empiriques des observables sont calculés sur les exemples des classes y correspondant. **(b)** Les trois estimateurs empiriques des observables sont calculées avec les exemples des trois classes. Le tableau 2 donne l'erreur d'apprentissage et de test (en pourcentage) en fonction du nombre d'exemples virtuels ajoutés. Dans ce cas, il y a une densité de probabilité pour chaque classe. Chaque classe possède 21 observations $A_{nc}(X)$ $(n = 1, \ldots, 21, c = 1, 2, 3)$. Cette observation est donnée par la difference au carré entre la composante $X_n$ et la composante correspondante du centre de la classe. Il y a trois ensembles de $\lambda_n$ à estimer. Il y a 21 $\lambda$ à estimer pour chaque classe.

Notons qu'on peut se servir d'une connaissance a priori d'une autre manière : Comme l'information contenue dans un ensemble d'apprentissage sera transportée par le processus d'apprentissage aux paramètres $\lambda_c$, alors, la quantité d'information, apportée par l'ajout des exemples virtuels à chacun de ces 11 ensembles d'apprentissage, est définie comme étant la quantité d'information moyenne calculée sur les 11 $\lambda_c$. Donc les paramètres du MaxEnt sont les valeurs moyennes sur les 11 $\lambda_c$. Cela veut dire qu'on essaie de construire un classifieur moyen dont les paramètres sont les valeurs moyennes des 11 valeurs correspondant aux 11 ensembles d'apprentissage. Dans le cadre de notre application et de notre choix d'observables, cela correspond à déterminer une distribution de probabilité gaussienne dont la variance est définie comme une valeur moyenne des variances de 11 distributions de probabilité. La figure 3 montre l'erreur d'apprentissage et de test (en pourcentage) obtenu par le classifieur moyen en fonction du nombre d'exemples virtuels ajoutés.

| Nb. d'exemples virtuels | 0 | 20 | 50 | 70 | 100 | 200 | 300 | 600 | 1000 |
|---|---|---|---|---|---|---|---|---|---|
| Erreur Apprentissage | 20.39 | 18.71 | 17.60 | 17.04 | 15.91 | 14.65 | 13.21 | 13.21 | 12.73 |
| Erreur Test | 20.31 | 19.82 | 19.69 | 19.65 | 19.55 | 18.89 | 18.94 | 18.72 | 18.79 |
| Variance | 2.35 | 2.31 | 2.44 | 2.60 | 2.39 | 2.44 | 2.77 | 2.61 | 2.61 |

**(a)**

| Nb. d'exemples virtuels | 0 | 20 | 50 | 70 | 100 | 200 | 300 | 600 | 1000 |
|---|---|---|---|---|---|---|---|---|---|
| Erreur Apprentissage | 19.75 | 18.23 | 16.79 | 16.17 | 15.39 | 14.51 | 13.03 | 12.53 | 12.40 |
| Erreur Test | 19.43 | 19.55 | 19.68 | 19.65 | 19.91 | 19.75 | 19.70 | 19.87 | 19.92 |
| Variance | 0.74 | 0.76 | 0.77 | 0.76 | 0.82 | 1.00 | 0.86 | 0.82 | 0.95 |

**(b)**

TAB. 1 – *Erreur d'apprentissage et de test (en pourcentage) en fonction du nombre d'exemples virtuels ajoutés. Chaque classe possède une seule observation* $A_c(X)$ $(c = 1, 2, 3)$ *qui est donnée par la distance euclidienne au carré entre le vecteur* $X$ *à 21 dimensions et le centre de la classe. Il y a trois* $\lambda$ *à estimer.* ***(a)*** *Les trois estimateurs empiriques des observables sont calculés sur les exemples des classes y correspondant.* ***(b)*** *Les trois estimateurs empiriques sont calculées avec les exemples des 3 classes.*

| Nb. d'exemples virtuels | 0 | 20 | 50 | 70 | 100 | 200 | 300 | 600 | 1000 |
|---|---|---|---|---|---|---|---|---|---|
| Erreur Apprentissage | 19.12 | 17.32 | 16.22 | 15.33 | 14.59 | 13.21 | 12.34 | 11.76 | 11.31 |
| Erreur Test | 19.62 | 19.27 | 19.23 | 19.00 | 19.17 | 18.68 | 18.83 | 18.83 | 18.56 |
| Variance | 1.11 | 1.12 | 1.22 | 1.01 | 0.99 | 0.88 | 1.09 | 1.00 | 1.02 |

**(a)**

| Nb. d'exemples virtuels | 0 | 20 | 50 | 70 | 100 | 200 | 300 | 600 | 1000 |
|---|---|---|---|---|---|---|---|---|---|
| Erreur Apprentissage | 17.81 | 16.69 | 15.54 | 14.30 | 14.15 | 12.45 | 11.95 | 11.2 | 10.98 |
| Erreur Test | 18.63 | 18.70 | 18.91 | 18.92 | 18.97 | 18.81 | 19.24 | 19.04 | 19.11 |
| Variance | 0.56 | 0.66 | 0.56 | 0.63 | 0.63 | 0.64 | 0.62 | 0.7 | 0.77 |

**(b)**

TAB. 2 – *Erreur d'apprentissage et de test (en pourcentage) en fonction du nombre d'exemples virtuels ajoutés. Chaque classe possède 21 observations* $A_{nc}(X)$ $(n = 1, \ldots, 21, c = 1, 2, 3)$. *Cette observation est donnée par la difference au carré entre la composante* $X_n$ *et la composante correspondante du centre de la classe. Il y a trois ensembles de* $\lambda_n$ *à estimer. Chaque classe a 21* $\lambda$. ***(a)*** *Les 21 estimateurs empiriques des observables sont calculés sur les exemples des classes y correspondant.* ***(b)*** *Les 21 estimateurs empiriques sont calculées avec les exemples des 3 classes.*

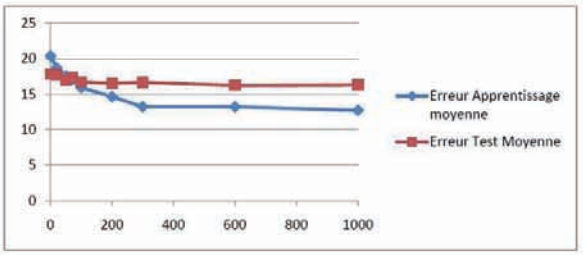


FIG. 3 – *Erreur d'apprentissage et de test (en pourcentage) en fonction du nombre d'exemples virtuels ajoutés. Dans ce cas, les résultats correspondent à un classifieur obtenu en moyennant les parametères de MaxEnt $\lambda_c$ sur les 11 ensembles d'aprrentissage augmentés.*

## 5 Conclusion

Nous avons étudié l'influence d'intégration des connaissances a priori sur la procédure d'apprentissage du maximum d'entropie, tout en supposant que les exemples de la base d'apprentissage suivent une combinaison convexe. Cette connaissance a été représentée par ajout d'exemples virtuels, représentant cette connaissance, à la base d'apprentissage. Nous avons evalué l'influence de cette intégration sur la prédiction du principe du maximum d'entropie.

## Références

Abumostafa, Y. S. (1993). Hints and the VC dimension. *Neural Computation 5*, 278–288.

Breiman, L., J. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. Chapman and Hall.

Buck, B. et V. Macaulay (1991). *Maximum Entropy in Action*. Clarendon Press, Oxford.

Chakik, F., A. Shahin, J. Jaam, et A. Hasnah (2004). An approach for constructing complex discriminating surfaces based on bayesian interference of the maximum entropy. *International Journal of Information Sciences 163*(4).

Lauer, F. et G. Bloch (2008). Incorporating prior knowledge in support vector machines for classification : A review. *Neurocomputing 71*(7–9).

Wu, X. et R. Srihari (2004). Incorporating prior knowledge with weighted margin support vector machines. In *International Conference on Knowledge Discovery and Data Mining*.

## Summary

This paper shows that the integration of prior knowledge into the the maximum entropy principle can improve the learning process. We studied the impact of using the convexity assumption through the use of virtual examples. The proposed schemes are evaluated on a benchmark problem given by Brieman waves. For this problem, we obtained a test recognition error of 15.57% that is very close to the theoretical value (14%).