Sélection de points en apprentissage actif. Discrépance et dispersion : des critères optimaux ?

B. Gandar, Gaëlle Loosli, G. Deffuant

MajecSTIC 2009
Avignon, France, du 16 au 18 novembre 2009

Sélection de points en apprentissage actif Discrépance et dispersion : des critères optimaux ?

Benoît Gandar[1,2], Gaëlle Loosli[2] et Guillaume Deffuant[1]

1 : Cemagref de Clermont-Ferrand, Laboratoire LISC (Laboratoire d'Ingénierie pour les Systèmes Complexes), 24 avenue des Landais, BP 50 085, 63 172 Aubière Cedex 1 - France.
2 : Université Blaise Pascal, Laboratoire LIMOS (Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes), Complexe scientifique des Cézeaux, 63 173 Aubière cedex - France.

Contact : benoit.gandar@cemagref.fr

Résumé
Nous souhaitons générer des bases d'apprentissage adaptées aux problèmes de classification. Nous montrons tout d'abord que les résultats théoriques privilégiant les suites à discrépance faible pour les problèmes de régression sont inadaptés aux problèmes de classification. Nous donnons ensuite des arguments théoriques et des résultats de simulations montrant que c'est la dispersion des points d'apprentissage qui est le critère pertinent à minimiser pour optimiser les performances de l'apprentissage en classification.

Mots-clés : Apprentissage actif, vitesse de convergence, classification, discrépance, dispersion.

Abstract
We want generate learning data appropriated to classification problems. First, we show that theorical results about low discrepancy sequences in regression problems are not adequated for classification problems. Then, we show with theorical and experimental arguments that minimising the dispersion of sample is the relevant strategy to optimize performance of classification learning.

Keywords: Activ learning, statistical learning, classification, discrepancy, dispersion.

---

1. Introduction

1.1. Présentation de l'apprentissage et de la problématique

L'apprentissage est une discipline à la frontière entre les mathématiques, les statistiques, l'informatique et la théorie de l'information. Son objectif est d'extraire et d'exploiter automatiquement l'information présente dans un jeu quelconque de données. Ces données sont de plus en plus abondantes et souvent de plus en plus complexes avec le développement des moyens informatiques, de télécommunication, des capteurs électroniques et autres compteurs. L'apprentissage consiste donc à développer, analyser et implémenter des méthodes qui permettent à une machine (au sens large) de traiter ces données et d'évoluer grâce à un processus d'inférence de décision issu de l'observation de celles-ci.
L'apprentissage permet ainsi de simuler une « intelligence artificielle » à de nombreuses machines ou outils. Souvent invisibles, ses applications sont nombreuses dans notre vie courante : moteur de recherche, aide au diagnostic, bio-informatique, détection de spams, reconnaissance vocale ou de l'écriture, analyse de vidéo, robotique, météorologie, ... D'un point de vue plus scientifique, on peut citer des travaux sur la reconnaissance de forme [4], sur la prédiction de séries caotiques [8], sur l'étude de fiabilité de structure industrielle [5] ou bien sur la détermination de zones de paramètres pour lesquelles un modèle exhibe certains comportements [6,11].
Le jeu de données initial sur lequel se base l'apprentissage possède un rôle important dans la capacité à bien apprendre. Par exemple, un enfant apprendra d'autant plus facilement à lire, que

le manuel de lecture utilisé n'est pas trop compliqué. Cependant, il pourra mieux appréhender la lecture d'un nouveau mot si ce manuel n'a pas été trop simple. Il est donc nécessaire que ce manuel, qui est à la base de l'apprentissage, possède un bon compromis entre simplicité et complexité. Il en est de même dans le cadre de l'apprentissage au sens mathématique et informatique du terme.
Cet article discute des caractéristiques que ce jeu de données, appelé également base d'apprentissage, doit posséder en apprentissage statistique. Les travaux actuels proposent la discrépance comme mesure de qualité de la base d'apprentissage. Nous proposons, en fonction du type d'apprentissage réalisé, un nouveau critère pour mesurer cette qualité : la dispersion.

1.2. L'apprentissage d'un point de vue mathématique

Nous nous plaçons dans le cadre de l'apprentissage statistique. C'est à dire que l'on cherche à apprendre une fonction f inconnue définie sur un espace $\mathbb{R}^s$ à valeur dans $\mathbb{R}$ à partir d'une base d'exemples $\{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_n, f(x_n))\}$. Dans le cadre de l'apprentissage classique, les points (x_i) sont issus d'une loi de probabilité $d\mu(x)$. L'objectif est d'approcher au mieux la fonction f (supposée ici à valeurs réelles), par une fonction $\hat{f}$, obtenue par un algorithme d'apprentissage. Pour une fonction quelconque $\tilde{f}$ de l'espace d'hypothèses, nous noterons $L(\tilde{f})$ l'erreur en généralisation de celle-ci définie par :

$$L(\tilde{f}) = \int \left|\tilde{f} - f\right|(x) d\mu(x).$$

Plus la fonction $\tilde{f}$ est proche de f, plus cette erreur est petite. Cependant, on ne peut la calculer formellement car la fonction f est inconnue. Elle est donc empiriquement estimée par :

$$\hat{L}(\tilde{f}) = \frac{1}{n} \sum_{i=1}^{n} \left|\tilde{f}(x_i) - f(x_i)\right|.$$

La résolution du problème d'apprentissage consiste en général à se donner un espace de fonctions, appellé également espace d'hypothèses, dans lequel nous allons chercher la fonction $\tilde{f}$ qui minimise l'erreur empirique. Nous distinguons deux cas possibles d'apprentissage : l'apprentissage de fonctions ou régression (la fonction f prend ses valeurs dans un espace continu de $\mathbb{R}$) et la classification (la fonction f prend ses valeurs dans un espace discret de $\mathbb{R}$. Dans la suite, nous supposerons que cet ensemble est $\{-1, +1\}$. La classification multi-classes est bien sûr envisageable, mais elle peut être vue comme un cas particulier de la classification bi-classe. Afin de faciliter la compréhension, nous n'étudierons que la classification bi-classe comme cela se fait usuellement.
Dans le cas classique de l'apprentissage, les données sont des réalisations de variables aléatoires, c'est à dire des observations issues de capteurs ou d'enquête. Le mathématicien réalisant l'apprentisage n'a pas d'influence sur leur récolte. Dans le cas de l'**apprentissage actif**, il en est tout autrement : nous supposons qu'il existe un oracle qui détermine ponctuellement l'image de tout point x par la fonction f recherchée. Le plan d'expérience est alors au choix de l'expérimentateur. Le problème devient alors de générer une base d'apprentissage la plus informative possible afin de reconstruire le plus fidèlement possible la fonction f avec une méthode d'apprentissage.

1.3. État de l'art de la sélection de points en apprentissage actif

Le problème de déterminer les meilleurs échantillons pour l'apprentissage actif de fonctions est déjà résolu. En s'appuyant sur les travaux sur les approximations d'intégrales, Mary [9] montre que, pour des problèmes d'apprentissage de fonctions (régression), les bornes théoriques d'erreur en généralisation sont meilleures en utilisant des bases d'apprentissage générées par des suites à discrépance faible qu'en utilisant des tirages aléatoires uniformes. La discrépance est une mesure du critère de bonne « uniformité » d'une suite. Cervellera & Muselli [3] avaient auparavant anticipé une démonstration empirique et théorique de ces résultats dans le cas particulier de l'apprentissage par réseaux de neurones.
Dans ce papier, nous montrons que la démarche théorique permettant d'obtenir les bornes d'erreur en généralisation pour l'approximation de fonction n'est pas adaptée à la classification. C'est un peu surprenant, car celle-ci peut apparaître comme un cas particulier du problème général d'approximation de fonctions. C'est pourquoi nous reprenons dans cet article la démarche de

Mary en détails, en identifiant précisément pourquoi elle ne s'étend pas aux problèmes de classification.
Une analyse plus fine du problème suggère que la dispersion de la base d'apprentissage, c'est à dire le rayon de la boule maximale ne contenant pas de points, est probablement un indicateur pertinent de qualité pour un apprentissage en classification. En effet, en utilisant une procédure d'apprentissage très simple (de type plus proches voisins), nous établissons théoriquement un lien entre l'erreur en généralisation et la dispersion.
Dans la partie 2 de ce document, nous montrons que les résultats théoriques sur la régression ne se transfèrent pas à la classification. Dans la partie 3, nous présentons des arguments théoriques en faveur de la dispersion comme indicateur pertinent de la qualité d'une base d'apprentissage pour les problèmes de classification. Dans la partie 4, nous montrons expérimentalement que les suites à dispersion faible permettent en classification avec l'algorithme des SVMs (Séparateurs à Vaste Marge) d'obtenir de meilleurs résultats en généralisation. Enfin dans une dernière partie, nous discutons ces résultats et concluons.

2. Les résultats sur la discrépance et la régression ne s'étendent pas à la classification

2.1. La notion de discrépance et de suite à discrépance faible

Soit I^s le cube unité en dimension s, I^{s*} l'ensemble des pavés de I^s ancrés à l'origine, # l'opérateur qui pour une suite $(x(n)) = x_1, \ldots, x_n$ à n éléments, et pour un ensemble P, fournit le nombre d'éléments de $x(n)$ appartenant à l'ensemble P. Nous noterons également λ la mesure de Lebesgue. Cette mesure est la mesure la plus utilisée en mathématiques (analyse et probabilité) aibnsi que dans la vie de tous les jours : elle correspond à la " longueur " dans le cas de $\mathbb{R}$, à la " surface " dans $\mathbb{R}^2$, au " volume " dans $\mathbb{R}^3$, etc ...
La discrépance à l'origine d'une suite $(x_{(n)})$, notée $D_n^*(x)$, et définie par :

$$D_n^*(x) = \sup_{P \in I^{s*}} \left| \frac{\#(P, (x_{(n)}))}{n} - \lambda(P) \right|.$$

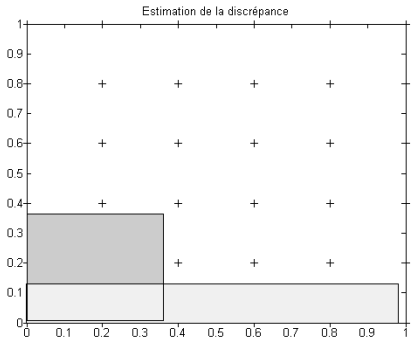


FIG. 1 – Illustration de la discrépance

La discrépance (voir FIG.1) est la différence maximale sur tous les pavés de l'espace entre la proportion de points de la suite appartenant au pavé et la proportion du volume du pavé par rapport au volume total de l'espace. Elle peut être vue comme une mesure de l' « uniformité » de la suite, en prenant en compte toutes les directions de l'espace et le recouvrement de celui-ci par la suite.

Une suite à discrépance faible est définie par une discrépance qui décroît à une vitesse de l'ordre de $O\left(\frac{\log^s(n)}{n}\right)$ [7]. Il est démontré [7] qu'une grille uniforme a une discrépance de l'ordre de $O\left(\frac{1}{\sqrt[s]{n}}\right)$, et n'est donc pas à discrépance faible.

2.2. Les bornes d'erreur en régression

Bornes d'erreur en régression en fonction de la discrépance de l'échantillon :

En s'inspirant des méthodes d'estimation d'intégrales, Mary applique le théorème de Koksma-Hlawka, qui borne l'erreur d'approximation d'intégrales, au cas de l'apprentissage statistique. Il obtient, pour toute fonction $\tilde{f}$:

$$\left| L(\tilde{f}) - \hat{L}(\tilde{f}) \right| \leq V_{HK}\left(\left|\tilde{f} - f\right|\right) D_n^*(x)$$

Dans cette équation, V_{HK} représente une mesure particulière de la variation d'une fonction : la variation de Hardy-Krause, et $D_n^*(x)$ représente la discrépance de la base d'apprentissage de taille n. Lorsque cette variation V_{HK} est bornée pour la fonction $\left|\tilde{f} - f\right|$, et lorsque la discrépance est faible, il est possible de majorer l'erreur en généralisation de manière déterministe, avec un majorant en $O\left(\frac{\log^s(n)}{n}\right)$.

Comparaison avec les bornes de Vapnik-Chervonenkis (VC) :

Dans la théorie de l'apprentissage statistique définie par Vapnik [13], l'estimation empirique d'une fonction convenablement obtenue par un algorithme d'apprentissage, décroît avec une vitesse de l'ordre $O\left(\frac{1}{\sqrt{n}}\right)$ avec un niveau de confiance fixé. Il s'agit d'une convergence en probabilité : on n'est jamais complétement sûr d'être dans l'intervalle considéré.
L'utilisation des suites permet d'obtenir un résultat déterministe, à une vitesse de convergence de l'ordre de $O\left(\frac{\log^s(n)}{n}\right)$. Cette vitesse est significativement plus rapide lorsque la dimension s est petite. De plus, la condition d'être en dimension VC finie est remplacée par une hypothèse de variation finie de fonctions. (La dimension VC d'une famille de fonctions est une mesure la capacité de celle-ci à bien classifier). Nous ne discuterons pas ici des avantages et inconvénients de cette dernière. Enfin, il n'est pas nécessaire que l'estimation empirique du risque de la fonction cible soit nulle, pour obtenir une vitesse de convergence de l'ordre de $O\left(\frac{1}{n}\right)$ au lieu de $O\left(\frac{1}{\sqrt{n}}\right)$. Des résultats plus fins existent dans [9] sous certaines conditions concernant la variation des fonctions contenues dans l'espace d'hypothèse mais ceux-ci ne seront pas présentés dans ce document.
Cette comparaison tourne encore plus à l'avantage des suites à discrépance faible lorsqu'on considère la minimisation du risque empirique [13]. Cette méthode considère une suite emboîtée de familles de fonctions dont la dimension de VC est finie et augmente et obtient un résultat stochastique avec une vitesse de convergence de l'ordre de $O\left(\sqrt{\frac{\log(n)}{n}}\right)$. Cependant, la même méthode en utilisant des suites à discrépance faible donne toujours un résultat déterministe en $O\left(\frac{\log(n)^s}{n}\right)$.

2.3. La variation de Hardy-Krause d'une fonction caractéristique est généralement non bornée

La classification correspond au cas où la fonction f à apprendre prend ses valeurs dans l'ensemble $\{-1, +1\}$: la fonction f est +alors une indicatrice. Les résultats présentés précédemment ne se transposent pas dans ce cas et le problème réside principalement dans l'estimation de la variation de Hardy-Krause d'une fonction f aussi discontinue. Owen a démontré [12] que cette variation est bornée uniquement dans les cas où les discontinuités de la fonction caractéristique f sont parallèles aux axes, c'est à dire que la fonction de classification se compose de l'union de pavés. De telles fonctions ne se rencontrent quasiment jamais dans un problème réel. Dans des cas non dégénérés, la variation de la fonction est infinie, ce qui rend inexploitable l'inégalité bornant l'erreur en généralisation.
Morokoff et Caflish montraient déjà dans [14] que l'utilisation des suites à discrépance faible en intégration n'était pas avantageuse dans le cas où l'intégrande était une fonction indicatrice. Des tests numériques en apprentissage dans [1] montrent que l'utilisation de ces même suites ne permet pas d'obtenir de meilleurs résultats par rapport aux grilles régulières (qui ont une discrépance plus forte). Il semble donc que le critère de qualité d'une base d'apprentissage basée sur la discrépance de celle-ci soit inadapté au le cas de la classification.

3. L'erreur en généralisation en classification est liée à la dispersion pour une procédure d'apprentissage simple

Les résultats précédents s'inspirent de travaux d'estimation d'intégrales utilisant des suites à discrépance faible. Une autre inspiration possible est le domaine de l'optimisation numérique. Dans celui-ci, on utilise généralement un algorithme itératif afin d'approximer l'extremum d'une fonction non différentiable dans un espace euclidien et borné. L'erreur d'approximation commise s'exprime alors de manière théorique comme une fonction de la dispersion des points d'approximation utilisée (voir [7]). En nous inspirant de ces travaux, nous avons essayé d'établir une relation entre la qualité d'apprentissage et la dispersion de la base d'apprentissage. Nous définissons maintenant rigoureusement la notion de dispersion.

Dispersion d'une suite :

Le pavé I^s est muni de la distance *euclidienne* d.
La dispersion d'une suite $x = \{x_1, \ldots, x_n\}$ se définit par :

$$\delta(x) = \max_{y \in I^s} \min_{i=1,\ldots,n} d(y, x_i)$$

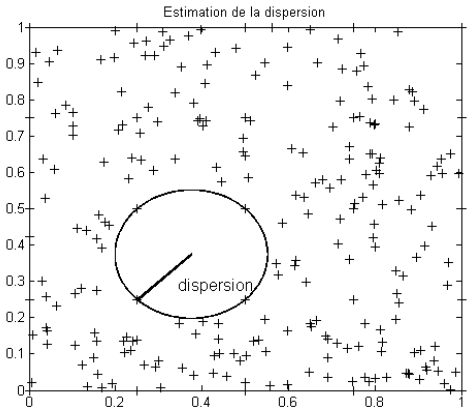
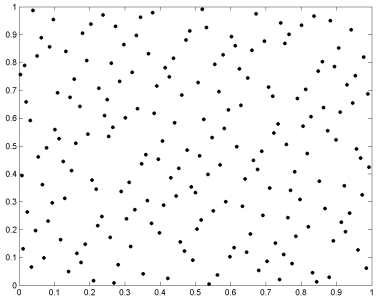


FIG. 2 – Illustration de la dispersion

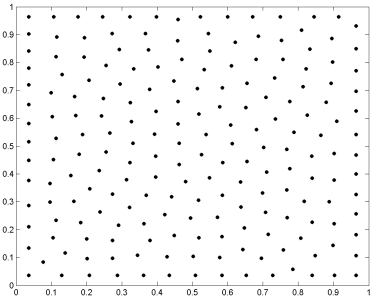
La dispersion (voir FIG. 2) est le rayon de la plus grande boule de l'espace ne contenant aucun point de la suite. Par conséquent, lorsque la dispersion est élevée, la suite de points comporte des « trous » dans le pavé I^s, et une dispersion faible assure une bonne répartition des points dans I^s, un recouvrement de l'espace sans « trous ».

Remarques sur la discrépance et la dispersion

La dispersion et la discrépance ne sont pas des mesures équivalentes (lorsque le nombre de point de la suite est fini, ce qui est le cas dans les applications) : la discrépance est bornée par 1, contrairement à la dispersion. De plus, l'ajout d'un point à une suite diminue ou ne change pas sa dispersion, alors que l'évolution de sa discrépance est quelconque. Enfin pour une suite de taille *adéquate*, la grille régulière minimise la dispersion et ne minimise pas la discrépance.
Pour convaincre le lecteur, nous avons représenté (FIG. 3(a)) une suite à discrépance faible de Halton avec 190 points en dimension 2 qui possède une dispersion de 0,11. A l'aide d'un algorithme explicité dans [2], nous avons déplacé ces points de manière à réduire leur dispersion. Le résultat se trouve sur la figure (FIG. 3(b)), où la suite finale possède une dispersion de 0,08. On peut remarquer sur cette dernière figure une tendance des points à se positionner selon une grille, configuration qui, rappelons le, ne possède pas une faible discrépance.



(a) Suite de Halton (à discrépance faible) de dispersion = 0,10.



(b) Suite de Halton modifiée de dispersion = 0,08.

FIG. 3 – Deux suites de 90 points à dispersion différente.

Lien entre l'erreur en généralisation et la dispersion pour une procédure d'apprentissage particulière :

Le but de cette partie est d'établir un lien entre erreur en généralisation et dispersion en utilisant une procédure d'apprentissage de type plus proches voisins.

Théorème :

Soit f une fonction de I^s à valeur dans $\{-1, +1\}$, dont on cherche l'approximation à partir d'une base d'exemples E de dispersion δ.
Soient $\chi_{f^+} = \{x \in I^s | f(x) = +1\}$ et $\chi_{f^-} = \{x \in I^s | f(x) = -1\}$.
On suppose que f possède la propriété de régularité suivante : $\exists R$ tel que
– $\forall x \in \chi_{f^+}, \exists x_0 \in \chi_{f^+} | x \in B(x_0, R)$ and $B(x_0, R) \subset \chi_{f^+}$
– $\forall x \in \chi_{f^-}, \exists x_0 \in \chi_{f^-} | x \in B(x_0, R)$ and $B(x_0, R) \subset \chi_{f^-}$

Soit l'algorithme d'apprentissage A approximant la fonction f par $A(E) = \hat{f}$ de la manière suivante :

$$\hat{f}(x) = \begin{cases} +1 & \text{si } \forall x_i^- \in E \cap \chi_{f^-}, d(x_i^-, x) \geq 2\delta. \\ -1 & \text{si } \forall x_i^+ \in E \cap \chi_{f^+}, d(x_i^+, x) \geq 2\delta. \\ \text{aléatoire} & \text{sinon} \end{cases}$$

Il existe $\lambda > 0$ tel que, pour toute base d'apprentissage E de dispersion $\delta < R$, la procédure A donne une approximation de f dont l'erreur en généralisation $L(A(E))$ vérifie : $L(A(E)) < \lambda\delta$.
Preuve :
Soient :
$F^+ = \{x \in I^s | \forall x_i^- \in E \cap \chi_{f^-}, d(x_i^-, x) \geq 2\delta\}$.
$F^- = \{x \in I^s | \forall x_i^+ \in E \cap \chi_{f^+}, d(x_i^+, x) \geq 2\delta\}$.

1. Montrons que $F^+ \subset \chi_{f^+}$. Soit $x \in F^+$. Supposons $x \in \chi_{f^-}$. L'hypothèse de régularité de f implique : $\exists x' \in \chi_{f^-} | x \in B(x', R)$ et $B(x', R) \subset \chi_{f^-}$. A fortiori $\exists x'' \in \chi_{f^-} | x \in B(x'', \delta)$ et $B(x'', \delta) \subset \chi_{f^-}$, puisque $R > \delta$. Par définition de la dispersion, $\exists x_0 \in E$, tel que $x_0 \in B(x'', \delta)$. On a alors $d(x, x_0) < 2\delta$, ce qui est contradictoire avec l'hypothèse ($x \in F^+$). Donc $x \in \chi_{f^+}$. L'algorithme d'apprentissage ne fait donc aucune erreur sur F^+. On démontre de même $F^- \subset \chi_{f^-}$.

2. Estimation de l'erreur d'apprentissage : $L(A(E)) = \int_{I^s} \left| f - \hat{f} \right| (x) dx$. Sur F^+ et F^-, f et $\hat{f}$ coïncident, les erreurs sont donc commises dans l'ensemble $I^s - F^+ - F^-$, qui sépare F^+ de F^-. Donc $L(A(E)) = \int_{I^s - F^+ - F^-} \left| f - \hat{f} \right| (x) dx$. Donc $L(A(E)) < V(I^s - F^+ - F^-)$, où V représente le volume de l'ensemble. Soit ∂f la frontière entre χ_{f^-} et χ_{f^+}, et $M = \{x \in I^s | d(x, \partial f) \leq 2\delta\}$. Il est clair que $I^s - F^+ - F^- \subset M$. En effet, $x \notin F^+$ implique $d(x, E \cap \chi_{f^-}) < 2\delta$, ce qui implique $d(x, \chi_{f^-}) < 2\delta$. Le même raisonnement pour F^- conduit à $d(x, \chi_{f^+}) < 2\delta$.

Donc $L(A(E)) < V(M)$. Or $V(M) \leq 4\delta S(\partial f)\left(\frac{R+2\delta}{R}\right)^{s-1} \leq 4\delta S(\partial f)3^{s-1}$, où $S(\partial f)$ représente l'intégrale sur la surface ∂f. La condition de régularité de la fonction assure que cette intégrale est finie. Le facteur faisant intervenir R permet de majorer le volume en supposant que le rayon de courbure de ∂f est à sa valeur minimale de R.

Conclusion

Pour cet algorithme particulier, l'erreur en généralisation est directement liée à la dispersion de la base d'apprentissage. Ce résultat laisse à suggérer que la dispersion est un indicateur pertinent de qualité de la distribution de l'échantillon pour un apprentissage en classification.

Retour sur l'hypothèse de régularité de la fonction

Dans cette approche, conformément à l'intuition, plus la fonction est irrégulière, plus la dispersion de la suite doit être faible et donc, plus le nombre d'exemples d'apprentissage doit être élevé (pour être dans le cas $\delta < R$).

4. Expériences numériques

Nous avons réalisé des tests numériques sur 1020 problèmes de classification avec des SVMs en dimension 3. Nous avons généré une base d'apprentissage de 700 points, avons réalisé les différents apprentissages et estimé les erreurs en généralisation. Ce processus a été itéré en diminuant la dispersion de la base avec un algorithme défini dans [2]. Nous avons représenté l'évolution moyenne du taux d'erreur en apprentissage en fonction du taux de diminution de la dispersion sur ces expériences.

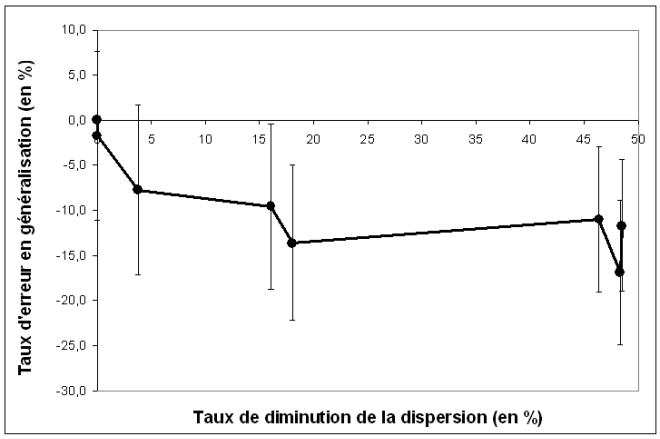


FIG. 4 – Évolution moyenne du taux d'erreur en apprentissage en fonction du taux de diminution de la dispersion (dimension 3, 700 points d'apprentissage, 1020 problèmes de classification).

On remarque (FIG. 4) une diminution du taux d'erreur en fonction du taux de diminution de la dispersion, laissant ainsi supposer que la dispersion est bien un critère intéressant pour évaluer la qualité d'une base d'apprentissage en classification.

5. Conclusion et discussion

En reprenant les travaux de Mary [9], nous avons montré que l'approche sur l'apprentissage de fonctions (régression) basée sur la discrépance de l'échantillon d'apprentissage ne s'appliquait pas au cas de fonctions indicatrices (classification).

En s'appuyant sur les méthodes mathématiques d'optimisation numérique de fonctions non différentiables (ce qui est le cas en classification), nous proposons la dispersion comme indicateur probablement pertinent de la qualité de la base d'apprentissage en classification. Un lien linéaire entre dispersion et erreur en généralisation a été établi dans le cas d'une procédure d'apprentissage simple.

Nous avons mis en évidence expérimentalement l'effet de la dispersion sur la qualité d'apprentissage. Nous pouvons donc légitimement nous demander si cette propriété est propre aux méthodes d'apprentissage utilisées, ou si l'on peut naturellement l'étendre à d'autres types d'apprentissage.

Dans le cadre de travaux sur la classification active avec des réseaux de neurones, Iwaata & Ishii (2002) ont montré de manière empirique dans [10] que des bases d'apprentissage générées par des suites à discrépance faible donnent une erreur en généralisation plus faible que des distributions aléatoires de type uniforme. Pour cela, ils ont comparé des apprentissages dans différentes dimensions (de 2 à 8) réalisés à partir d'exemples issus de distributions uniformes (stochastiques) avec des apprentissages dont les exemples étaient des suites à discrépance faible (de type suite de Faure). Ces résultats ne sont pas contradictoires avec les nôtres. En effet, la dispersion d'une suite à discrépance faible est en général plus petite que celle d'une suite aléatoire.

Enfin, si ces résultats se confirment, il serait alors probablement intéressant de générer des données à faible dispersion de manière récursive en tirant des densités de points plus fortes au voisinage des frontières de la fonction indicatrice détectées à l'itération précédente.

Bibliographie

1. Gandar B, Deffuant G, et Loosli G. Les suites à discrépance faible : un moyen de réduire le nombre de vecteurs supports des svms ? In 12^{ieme} *Journée Scientifique de l'Ecole Doctorale SPI : Apprentissage statistique - Apprentissage symbolique. Annales scientifiques de l'Université Blaise Pascal, Clermont-Ferrand II.*, 2008.
2. Gandar B, Loosli G, et Deffuant G. Les suites à dispersion faible comme bases d'exemples optimales en apprentissage. Technical report, Cemagref, 2009.
3. Cervellera C et Muselli M. Deterministic design for neural network learning : An approach based on discrepancy. In *Proceedings IEEE Transactions on Neural Network*, volume 15, pages 533–544, 2004.
4. Burgues CJC. A tutorial on support vector machines for pattern recognitionrandom number generation and quasi-monte carlo methods. In *Data Mining and Knowledge Discovery*, volume 2, pages 121–167. Ed. Society for Industrial and Applied Mathematics, 1992.
5. Deheeger F et Lemaire M. Support vector machine for efficient subset simulations : smart method.
6. Deffuant G, Martin S, et Chapel L. Utiliser des « supports vector machines » pour apprendre un noyau de viabilité. In *MajecSTIC, Rennes*, 2005.
7. Niederreiter H. *Random Number Generation and Quasi-Monte Carlo Methods*. Ed. Society for Industrial and Applied Mathematics, 1992.
8. Wendt H, Flandrin P, et Abry P. Régressions par machines à vecteurs supports pour la prédiction de séries chaotiques.
9. Mary J. *Etude de l'Apprentissage Actif, Application à la Conduite d'Expériences*. Thèse de doctorat, Université Paris XI, 2005.
10. Iwata K et Ishii N. Discrepancy as a quality measure for advoiding classification bias. In *Proceedings of the 2002 IEEE International Symposium on Intelligent Control. Vancouver, Canada.*, 2002.
11. Chapel L et Deffuant G. Svm viability controller active learning : application to bike control. In *IEEE Approximate Dynamic Programming and Reinforcement Learning. Hawaï, États-Unis*, 2007.
12. Owen. Multidimensional variation for quasi-monte carlo. 2004. `www-stat.stanford.edu/~owen/reports/ktfang.pdf`.
13. Vapnick VN. *The Nature of Statistical Learning*. Springer-Verlag, 1995.
14. Morokoff WJ et Caflisch RE. Quasi monte-carlo integration. 122 :218–230, 1995.