ENBIS 2009
Saint-Etienne, France, 1-3 July 2009

How to optimize sample in active learning : Dispersion, an optimum criterion for classification ?

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.
http://wwwlisc.clermont.cemagref.fr
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

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 adequate for classification problems. Then, we show with theorical and experimental arguments that minimising the dispersion of the sample is a relevant strategy to optimize performance of classification learning.

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

---

1. Introduction

We consider a problem of active learning classification: we suppose we can détermine, with an oracle, the label of any point in a given compact set, and we want generate a sample of a given size which will allow us to get the best approximation of the oracle function. We suppose that this function is deterministic.
This problem may arise in various contexts, but our research is particulary motivated by the resolution of viability problems (shows [7]), which are frequent in economy, ecology or robotics. The goal is to compute policies of actions in order to keep a dynamical system within a given suibset of the state space. Generally, it is supposed that the system badly deteriorates if it crosses the boundary of this constraint set. An essential step to solve these problems is to determine a particular subset of this constraint set, which is called viability kernel. An algorithm [9], computes viability kernels by recursive approximations using statistical learning algorithm (Support Vector Machines (shows [3]) being a particularly relevant learning technique in this context). At each step of the algorithm, one must solve an active classification learning problem: we can compute if any point belongs to the set to approximate (i.e. we have an oracle which needs a lot of costly computer experiments at each point), but to optimise the time of the process, we need to limit the size of the sample as much as possible, and determine the distribution of the sample leading to the best approximation.
A similar problem, which is to determine the best learning set for active learning of functions (regression) is already solved. Indeed, using results about the approximation of integrals, Mary (shows [6]) proves using low discrepancy samples provides the best results for a regression problem. Cervellera & Muselli [4] had already suggested an empirical and theorical demonstration of these results in the specific case of the multi-layer perceptron.
We show that the theoretical approach to obtain generalisation error bounds in regression is not adapted to classification. This result is somehow surprising, because classification can be seen as a particular case of functions approximation.
An analyse in depth suggests that dispersion, i.e. the radius of the higher ball containing no points, is probably an pertinent indicator of quality for samples to be used in active classification.

Indeed, using a simple learning algorithm (as nearest neighbours), we establish a theorical link beetwen generalisation error and dispersion. Moreover, we present experimental results using SVMs that confirm this hypothesis.
In the second part of this document, we present theoretical results on active regression, and we show that, surprisingly, these results cannot be transfered to classification (learning manifold boundaries). In the third part, we show with theorical and experimental arguments, that minimising the dispersion of the sample is the relevant strategy to insure the best results in active classification learning problems and to limit the sample size. In the last part, we discuss about these results and conclude.

2. The results about active regression learning do not apply to active classification learning

We suppose we have $\{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_n, f(x_n))\}$, a set of labelled examples.The objective is to approximate as precisely as possible the function f (supposed to be real), by a function $\hat{f}$, obtained with a learning algorithm (for example an empirical risk minimisation). For any function $\tilde{f}$ of a hypothesis space, we note $L(\tilde{f})$ the error of generalisation defined by :

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

This error will be experimentally estimated by:

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

Error bounds in regression function of discrepancy of learning set:

Applying to statistical learning the Koksma-Hlawka theorem, which limits integral approximation error, we obtain for any function $\tilde{f}$:

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

where $V_{HK}(g)$ is a particular measure of the regularity of the function g: the variation in the sense of Hardy Krause, and D_n^* is the dispersion of a sample of size n.
In this inequality, it is clear that bound is determinist and directly proportional to the discrepancy of the learning set, supposing the variation is finite.

Discrepancy of a sequence:

The discrepancy of a sequence can be viewed as a quantitative measure for good "uniformity" of a sequence. Considering without loss of generality that our sample has to ge taken inside the unit hypercube I^s, of dimension s, it is the maximal difference on all the convex subsets of I^s between the proportion of the points in the convex subset and the volume of the convex subset.
One can show that this definition is closely related to the star discrepancy, in which, instead of considering any convex subset of I^s, we only consider hyperrectangles containing the origin. More formally, we note I^{s*} the set of all subintervals of I^s of the form $\prod_{i=1}^{s}[0, u_i)$, # the operator which, for a sequence $(x_{(n)}) = x_1, \ldots, x_n$ with n elements and a set P, gives the number of element of $x_{(n)}$ in the set P. The Lebesgue measure will be noted λ. We consider only the star discrepancy $D_n^*(x)$of an n-sequence $(x_{(n)})$ defined by (shows FIG. 1)

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

According to Niederreiter (shows [5]), the discrepancy of a low discrepancy sequence decreases to zero as $O\left(\frac{\log^s(n)}{n}\right)$. Note that a regular grid has a discrepancy of order $O\left(\frac{1}{\sqrt{n}}\right)$, which is not low.

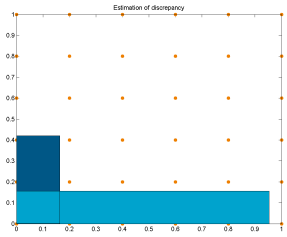


Figure 1: Estimation of discrepancy

Error bounds in regression with a low discrepancy sequence:

Mary applies the theorem of Koksma-Hlawka, which bounds integral approximation, in the context of statistical learning.
He obtains that for any function $\tilde{f}$:

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

In this equation, V_{HK} is a particular measure of variation of a function: the variation of Hardy-Krause. When this variation is bounded for the function $\left|\tilde{f} - f\right|$, it is possible to bound up deterministic generalisation error, with a upper bound in $O\left(\frac{\log^s(n)}{n}\right)$.

Comparison with Vapnik-Chervonenkis bounds (VC):

Within the context of satatistical learning (shows [12]), empirical estimation of a function decreases as $O\left(\frac{\log^s(n)}{n}\right)$ and with a fixed confidence level. It is a convergence in probability. Using low discrepancy sequence, we obtain a determinist upper bound decreasing as $O\left(\frac{\log^s(n)}{n}\right)$. This speed is significantly quicker when the dimension s is small. Furthermore the condition to be in finite VC dimension is substituted by an hypothesis of finite variation of functions. Finally it is not necessary to have a null empirical risk estimation of target function to obtain an upper bound of the error decreasing as $O\left(\frac{1}{n}\right)$instead of $O\left(\frac{1}{\sqrt{n}}\right)$.
Considering the empirical risk minimisation [12], where we consider a sequence of functions which have a finite and increasing VC-dimension, we obtain a stochastic convergence with speed about $0\left(\sqrt{\frac{\log(n)}{n}}\right)$. Using low discrepancy sequence, we always obtain a deterministic convergence about $0\left(\frac{\log^s(n)}{n}\right)$.

The variation of Hardy-Krause of an indicator function is infinite:

Classification is the case where function f takes its values in the set $\{0, 1\}$. Previous results cannot be transposed to this case. Indeed, the variation of indicator functions is generally infinite (shows [11]). So the superior bound in the previous inequality is equal to infinity too.
Moreover, Morokoff and Caflish have demonstrated (shows [10]) that using low discrepancy sequences is not efficient when the integrand function is an indicator function. Numerical tests proved (shows [1]) that, using these sequences, we cannot obtain better results than using a regular grid (which have a higher discrepancy). Therefore the discrepancy does not seem to be the relevant criterion to get optimal samples for classification.

3. Low dispersion is a better criterion of sample quality for active classification learning

Previous results come from multidimensionnal integration. An other possible inspiration comes from numerical optimisation. In this direction, we generaly use an iterative algorithm to approx-

imate the extremum of a non derivable function in a compact set. Approximation error can also be theoretically expressed by a function of dispersion (shows [5]).

Dispersion of a sequence:

We consider the unit cube I^s with the euclidian distance d.
The dispersion of a sequence $x = \{x_1, \ldots, x_n\}$ is defined by:

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

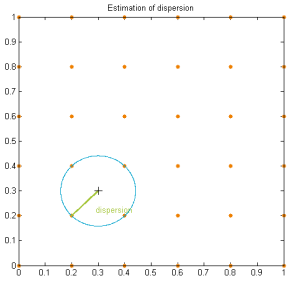
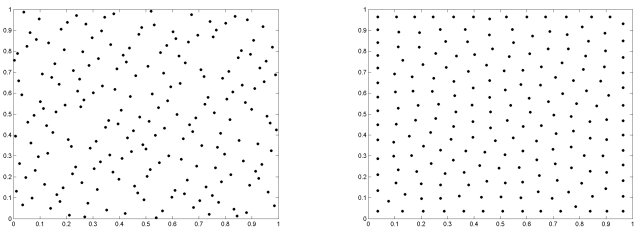


Figure 2: Estimation of dispersion

The dispersion of the sequence is the radius of the biggest empty ball of I^s (shows FIG. 2).

Remarks about discrepancy and dispersion:

Discrepancy and dispersion are not equivalent measures. Indeed, when we add a point in a sequence, its dispersion doesn't change or decrease. Its discrepancy can increase or decrease. Moreover, for an appropriate number of point, the configuration which minimizes the dispersion is a regular grid which doesn't minimize discrepancy.
To illustrate these differences, we have represented in dimension 2 (FIG.3(a)) a low discrepancy sequence of Halton with 190 points and which have a dispersion of 0,11. With an algorithm describted in [2], we have deplaced the points in order to decrease the dispersion. The final result is the figure (FIG.3(b)): dispersion is equal to 0,08. Note on this last figure a tendency of the points form a regular grid, which does not have a low discrepancy.



(a) Low discrepancy (Halton) of dispersion = 0,10.

(b) Halton sequence modified of dispersion = 0,08.

Figure 3: Two sequences of 90 points.

Generalisation error is function of dispersion for a simple classification learning:

The purpose of this paragraph is to established a link beetween generalisation error and dispersion of learning set using a learning process simiular to the nearest neighbours.

Theorem:

Let f a function from I^s to $\{-1,+1\}$. We want to approximate its with a learning set E of dispersion δ. Denoting $B(x,R)$ the ball of center x and radius R.
Let $\chi_{f^+} = \{x \in I^s | f(x) = +1\}$ and $\chi_{f^-} = \{x \in I^s | f(x) = -1\}$. We suppose f has this propriety of regularity: $\exists R$ such

- $\forall x \in \chi_{f^+}, \exists x_0 \in \chi_{f^+} | x \in B(x_0,R)$ 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 $B(x_0,R) \subset \chi_{f^-}$

Let the learning algorithm A approximating the function f by $A(E) = \hat{f}$ as :

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

There is $\lambda > 0$ such as, for any learning set E of dispersion $\delta < R$, the algorithm A gives an approximation of f with a generalisation error $L(A(E))$ such as: $L(A(E)) < \lambda\delta$.

Proof:

Let: $F^+ = \{x \in I^s | \forall x_i^- \in E \cap \chi_{f^-}, d(x_i^-, x) \geq 2\delta\}$ and $F^- = \{x \in I^s | \forall x_i^+ \in E \cap \chi_{f^+}, d(x_i^+, x) \geq 2\delta\}$.

1. Prove that $F^+ \subset \chi_{f^+}$. Let $x \in F^+$. Supposing $x \in \chi_{f^-}$. Regularity hypothesis of f implies: $\exists x' \in \chi_{f^-} | x \in B(x',R)$ et $B(x',R) \subset \chi_{f^-}$. All the more $\exists x'' \in \chi_{f^-} | x \in B(x'',\delta)$ and $B(x'',\delta) \subset \chi_{f^-}$, because $R > \delta$. By definition of the dispersion, $\exists x_0 \in E$, such as $x_0 \in B(x'',\delta)$. Therefore $d(x,x_0) < 2\delta$, it is in contradiction to the hypothesis ($x \in F^+$). So $x \in \chi_{f^+}$. The learning algorithm does not make mistakes on F^+. Hence we have $F^- \subset \chi_{f^-}$.

2. Estimation of learning error: $L(A(E)) = \int_{I^s} \left|f - \hat{f}\right|(x)dx$. On F^+ and F^-, f et $\hat{f}$ are equal, therefore errors are in the set $I^s - F^+ - F^-$, which distings F^+ from F^-. So $L(A(E)) = \int_{I^s - F^+ - F^-} \left|f - \hat{f}\right|(x)dx$. So $L(A(E)) < V(I^s - F^+ - F^-)$, where V is the volum of this set. Let ∂f the boundary beetwen χ_{f^-} and χ_{f^+}, et $M = \{x \in I^s | d(x,\partial f) \leq 2\delta\}$.
   It's evident that $I^s - F^+ - F^- \subset M$. Indeed, $x \notin F^+$ implicates $d(x, E \cap \chi_{f^-}) < 2\delta$, that implicates $d(x,\chi_{f^-}) < 2\delta$. Hence, we demonstrate that $d(x,\chi_{f^+}) < 2\delta$.
   Therefore $L(A(E)) < V(M)$. But $V(M) \leq 4\delta S(\partial f)\left(\frac{R+2\delta}{R}\right)^{s-1} \leq 4\delta S(\partial f) 3^{s-1}$, where $S(\partial f)$ is the integral on the surface ∂f. Regularity condition on f insures that this integral is finite. This factor defined with R allows to bound the volum, supposing the radius of curvature of ∂f is at its minimal value R.

Conclusion:

With this particular algorithm, generalisation error is directly linked to the dispersion of the learning set. We can think, with this result, that dispersion is a pertinent indicator to measure the quality of a learning set in classification.

4. Numerical experiments

We have made numerical experiments on 1.020 classification learning problems in dimension 3. We have generated a learning set with 700 points, made the learning process with the SVMs (shows [3]), and estimated the generalisation error. We have iterated this process when decreasing the dispersion of the sample with an algorithm described in [2].
We can see (FIG.4) a decrease of error rate functions of dispersion decreasing rate. It seems that minimising dispersion is a relevant straegy to insure the best results in classification learning problems.

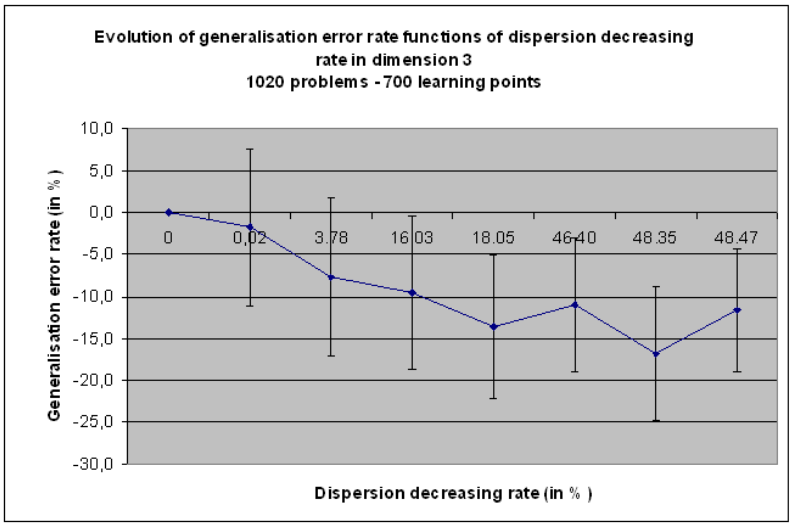


Figure 4: Average evolution of generalisation error rate functions of dispersion decreasing rate (dimension 3, 700 learning points, 1.020 classification problems).

5. Conclusion and discussion

Using works from Mary (shows [6]), we prove that the results about discrepancy established for regression cannot be transfered to classification.

We propose the dispersion as pertinent criterion for optimising the samples for classification. We established a linear link between dispersion and generalisation error in classification within the context of a simple learning algorithm. Moreover, our experimental tests on SVMs also show a link between dispersion on and generalisation error.

Iwata & Ishii (shows [8]) observed experimently a gain of quality in classification with the multilayer perceptron using low discrepancy samples instead of random samples. It does not contradict our results. Indeed the dispersion of a random sample is generally higher than dispersion of low discrepancy sample.

If theses results are confirmed, it will be probably interresting to generate iteratively low dispersion data, with a higher density near the boundary of classification function detected at the previous step. This could enhance significantly the learning performance obtained with a sample of size n, and could so limit costly computer experiments in our application.

Bibliographie

1. Gandar B, Deffuant G, and 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, and Deffuant G. Les suites à dispersion faible comme bases d'exemples optimales en apprentissage. Technical report, Cemagref, 2009.
3. Schölkopf B and Smola AJ. *Learning with Kernels : Support Vector Machines, Regularisation, Optimization, and Beyond*. The MIT Press, 2002.

4. Cervellera C and 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.
5. Niederreiter H. *Random Number Generation and Quasi-Monte Carlo Methods*. Ed. Society for Industrial and Applied Mathematics, 1992.
6. Mary J. *Etude de l'Apprentissage Actif, Application à la Conduite d'Expériences*. PhD thesis, Université Paris XI, 2005.
7. AUBIN JP. *Viability theory*. Birkhäuser, 1991.
8. Iwata K and 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.
9. Chapel L and Deffuant G. SVM viability controller active learning: application to bike control. In *IEEE Approximate Dynamic Programming and Reinforcement Learning. Hawaï, États-Unis*, 2007.
10. W. J. Morokoff and Russel E. Caflisch. Quasi-Monte Carlo integration. *J. Comput. Phys.*, 122(2):218–230, 1995.
11. Owen. Multidimensional variation for quasi-Monte Carlo. 2004. `www-stat.stanford.edu/~owen/reports/ktfang.pdf`.
12. Vapnick VN. *The Nature of Statistical Learning*. Springer-Verlag, 1995.