Comparaison de différentes approches de l'évaluation supervisée

Comparing various approaches of supervised evaluation

Sylvain Ferrandiz

Édition électronique

URL : <http://journals.openedition.org/msh/11126>

DOI : 10.4000/msh.11126

ISSN : 1950-6821

Éditeur

Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée

Date de publication : 30 décembre 2009

Pagination : 59-77

ISSN : 0987-6936

Référence électronique

Sylvain Ferrandiz, « Comparaison de différentes approches de l'évaluation supervisée », *Mathématiques et sciences humaines* [En ligne], 187 | Automne 2009, mis en ligne le 15 décembre 2009, consulté le 23 juillet 2020. URL : <http://journals.openedition.org/msh/11126> ; DOI : <https://doi.org/10.4000/msh.11126>

COMPARAISON DE DIFFÉRENTES APPROCHES DE L'ÉVALUATION SUPERVISÉE

Sylvain FERRANDIZ¹

RÉSUMÉ – *La sélection d'instances pour la classification suivant le plus proche voisin est un problème classique d'apprentissage statistique. Placé dans le cadre de la sélection d'hypothèses, ce problème possède deux caractéristiques : premièrement l'ensemble des hypothèses est structuré et, deuxièmement, il dépend des données. Nous proposons des critères d'évaluation non paramétriques tenant compte de ces caractéristiques, notre objectif étant de comparer des sous-ensembles d'instances de cardinaux différents sans recourir à l'ajustement d'un paramètre extérieur. Pour cela, nous nous intéressons aux approches régularisantes de l'évaluation.*

Nous explorons successivement trois approches de l'évaluation statistique régularisée : l'approche SRM (pour Structural Risk Minimisation), l'approche BIC (pour Bayesian Information Criterion) et l'approche MDL (pour Minimum Description Length). Nous proposons ainsi trois nouveaux critères d'évaluation régularisée de l'intérêt d'un sous-ensemble d'instances. Tous trois autorisent la comparaison d'ensembles d'instances de tailles différentes. Tous trois sont non paramétriques.

Nous procédons à une comparaison qualitative des critères, basée sur des données réelles et synthétiques, et démontrons le fait suivant : le critère MDL est plus fin que le critère BIC, lui-même plus fin que le critère SRM.

MOTS CLÉS – Bayesianisme, Classification supervisée, Longueur de description, Plus proche voisin, Risque structurel

SUMMARY – Comparing various approaches of supervised evaluation
Instance selection for the nearest neighbor rule is a classical topic in statistical learning. Within the context of hypothesis selection, the characteristics of this problem is that: the set of hypotheses is structured and depends on the data. We thus propose specific nonparametric criteria. We aim at comparing sets of instances of varying size without introducing an extra parameter. Balancing approaches give tools to solve this problem.

Three approaches are successively considered : the SRM (standing for Structural Risk Minimization) approach, the BIC (standing for Bayesian Information Criterion) approach and the MDL (standing for Minimum Description Length) approach. The exploration of each one leads to the definition of a regularized criterion. Each criterion allows to compare sets of instances of various size. Each criterion is nonparametric.

We make use of real and synthetic datasets to prove the following point : the MDL criterion is finer than the BIC criterion which, in turn, is finer than the SRM criterion.

KEYWORDS – Bayesianism, Description length, Nearest neighbor, Structural risk, Supervised classification

¹Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion, sylvain.ferrandiz@gmail.com

1. INTRODUCTION

L'apprentissage statistique consiste à sélectionner à partir de données la meilleure hypothèse parmi tout un ensemble. Le critère d'évaluation utilisé est central. Différentes approches de l'évaluation existent, comme les approches basées sur le risque empirique et les approches bayésiennes. Dans cet article, nous les explorons en nous focalisant sur un problème d'apprentissage précis : la sélection d'instances pour la classification suivant le plus proche voisin.

En classification supervisée, la règle de classification suivant le plus proche voisin [Fix, Hodges, 1951] attribue à un nouvel individu l'étiquette de l'individu le plus proche parmi ceux constituant l'échantillon. La mise en œuvre de cette modélisation soulève trois questions fondamentales : celle des variables descriptives à considérer, celle de la métrique à employer et celle des individus de l'échantillon à conserver. Nous nous intéressons dans cet article uniquement à la dernière. Ainsi, nous supposons fixés par ailleurs l'ensemble des variables décrivant les individus et la métrique mesurant la similitude entre deux individus.

De manière usuelle pour un problème d'apprentissage statistique, le développement d'une méthode de sélection d'instances passe par les trois étapes suivantes :

- définition de l'ensemble des hypothèses à considérer
- définition d'un critère d'évaluation sur l'ensemble des hypothèses considéré
- définition d'un algorithme de recherche de la meilleure hypothèse au sens du critère.

Nous traitons ici uniquement les deux premiers aspects. Le lecteur intéressé par la partie algorithmique de la sélection d'instances peut prendre comme points de départ [Liu, Motoda, 2001 ; Wilson, Martinez, 2000].

L'article est organisé comme suit. Dans la Section 2, nous présentons la question de la sélection d'instances pour la règle du plus proche voisin comme un problème de sélection d'hypothèses. Nous en présentons les caractéristiques principales et posons les notations. Dans les sections suivantes nous proposons trois critères d'évaluation des hypothèses en adoptant successivement trois approches.

Dans la Section 3, nous adoptons une approche structurale de type SRM basée sur le risque empirique et obtenons un premier critère en tirant partie des travaux de [Cannon *et al.*, 2002]. Nous illustrons son comportement sur des jeux de données synthétiques. Dans la Section 4, nous interprétons de manière heuristique l'approche bayésienne BIC décrite dans [Schwartz, 1978] et obtenons un second critère. Son comportement est illustré sur des jeux de données synthétiques et réels. Dans la Section 5, nous rappelons l'approche bayésienne et le critère introduits dans [Ferrandiz, Boullé, 2006]. Nous la rangeons ici dans la catégorie des approches MDL, pour la distinguer de l'approche bayésienne BIC précédente. Son apport est illustré sur des jeux de données réels. Nous concluons dans la Section 6.

2. ENSEMBLE D'HYPOTHÈSES STRUCTURÉ ET DÉPENDANT DES DONNÉES

La sélection d'instances pour la règle de classification suivant le plus proche voisin constitue un exemple de problème de sélection d'hypothèses où l'ensemble des hypothèses est structuré et dépend des données. C'est ce que nous précisons ici, ainsi que les notations.

Commençons par une description macroscopique. L'ensemble des variables descriptives et la métrique étant fixés, les individus sélectionnés (ou : prototypes) induisent une partition de Voronoi des individus. Les individus se répartissent dans les groupes de la partition et on obtient pour chaque groupe une distribution des étiquettes. La question de la sélection d'instance repose alors sur la définition et l'optimisation d'un critère d'évaluation basé sur les distributions de la cible obtenues pour quantifier l'intérêt du regroupement.

Soyons maintenant plus précis et fixons les notations. Soit Ω l'univers, *i.e.* l'ensemble des individus sur lesquels portent les mesures. Soit $\mathbb{L} = \llbracket 1, J \rrbracket$ l'ensemble des J étiquettes. On note Y la variable *cible* qui à tout individu associe une étiquette dans $\mathbb{L}$. Soit $\mathbb{X}$ l'espace de représentation muni d'une métrique δ et X la variable *descriptive* qui à tout individu de Ω associe des mesures dans $\mathbb{X}$. On suppose à disposition un *échantillon* du couple (X, Y) . Autrement dit, on dispose des valeurs $(X(\omega_n), Y(\omega_n))$ pour un ensemble $\mathbb{I}$ de N individus $\{\omega_n; 1 \leq n \leq N\}$. On note, pour tout $n \in \mathbb{I}$, $x_n = X(\omega_n)$ et $y_n = Y(\omega_n)$. Enfin, $E = (x_n, y_n)$ désigne l'échantillon.

On ne considère ici que des méthodes à base de plus proche voisin. Sur le plan formel, on pourrait se passer de X et supposer simplement qu'on dispose d'une métrique sur Ω . Néanmoins, les formalismes des différentes approches (SRM, BIC et MDL) ne sont pas spécifiques à ces approches et leur présentation dans un contexte générique nécessite de faire apparaître l'espace de description $\mathbb{X}$. En conséquence, nous parlerons fréquemment et abusivement de distance entre individus, de partition de l'univers, alors qu'il serait plus en accord avec le formalisme de parler de distance entre les représentations des individus et de partition de l'espace de représentation.

Soit $H \subset \mathbb{I}$. Pour tout $k \in H$ (par abus de notation), on appelle *cellule de Voronoi* de k relativement à H l'ensemble

$$V(k, H) = \left\{ \omega \in \Omega; k = \arg \min_{k' \in H} \delta(X(\omega), x_{k'}) \right\} \quad (1)$$

Les éléments de $V(k, H)$ sont les éléments de Ω pour lesquels k est le plus proche élément de H . L'individu k est appelé *prototype* de la cellule $V(k, H)$.

La famille $(V(k, H))_{k \in H}$ est appelée *diagramme de Voronoi* associé à H et on la note $V(H)$. A condition de résoudre les cas d'égalité lorsque plusieurs prototypes sont à égale distance d'un même individu, un diagramme de Voronoi réalise une partition de l'ensemble des individus Ω (par abus de langage). Nous adoptons le traitement des égalités décrit dans [Devroye *et al.*, 1996]. À tout sous-ensemble H de $\mathbb{I}$ nous sommes alors en mesure d'associer une partition de Ω , qu'on qualifie de *partition de Voronoi*. Un exemple d'une telle partition est donné à la Figure 1.

FIGURE 1. Exemple de diagramme de Voronoi dans le plan euclidien.

Nous notons $\mathcal{H}(\mathbb{I})$ l'ensemble des partitions de Voronoi obtenu. Cet ensemble est de cardinal fini au plus égal à 2^N . Notons, pour $1 \leq K \leq N$, $\mathcal{H}^{(K)}(\mathbb{I})$ l'ensemble des partitions de Voronoi en K cellules. Sélectionner un ensemble de K individus à conserver pour la classification suivant le plus proche voisin, c'est donc choisir, au moins implicitement, une partition de Voronoi en K cellules de l'ensemble Ω . La sélection d'instances est en ce sens un problème de sélection d'hypothèse, les hypothèses étant des partitions de Ω .

Particularités. L'ensemble $\mathcal{H}(\mathbb{I})$ des hypothèses que nous considérons dépend des données. De plus, il est structuré : $\mathcal{H}(\mathbb{I}) = \bigcup_K \mathcal{H}^{(K)}(\mathbb{I})$.

Objectif. Nous nous fixons pour objectif de définir un critère c évaluant la pertinence de toute hypothèse :

$$\begin{aligned} c : \mathcal{H}(\mathbb{I}) &\rightarrow \mathbb{R}_+ \\ H &\mapsto c(H) \end{aligned} \tag{2}$$

De plus, nous souhaitons obtenir un critère non paramétrique.

L'optimisation d'un tel critère fournit une solution sans que l'utilisateur ait besoin de préciser *a priori* le nombre K de prototypes. Le nombre de prototypes est un résultat. Un critère non paramétrique évite tout travail supplémentaire consommateur de temps et de données, comme par exemple la mise en œuvre d'une procédure de validation croisée [Stone, 1974].

L'obtention d'un tel critère passe par l'adoption d'une approche régularisante de l'évaluation, comme l'approche structurale ou les approches bayésiennes.

3. APPROCHE SRM : PROPOSITION D'UN RISQUE STRUCTUREL

Pour un classifieur $f : \mathbb{X} \rightarrow \mathbb{L}$, le risque empirique est le nombre d'erreurs de classification réalisées par f sur un échantillon. De manière générale, on obtient

ainsi une estimation biaisée de la probabilité d'erreur d'un classifieur [Hastie *et al.*, 2001]. Pour fiabiliser l'estimation, la théorie de l'apprentissage statistique [Vapnik, 1996] propose de pénaliser le risque empirique par un terme structurel en relation avec la capacité du classifieur. Nous présentons brièvement les concepts clés de cette théorie, leur adaptation au cas d'un ensemble d'hypothèses dépendant des données et proposons un nouveau critère pour les partitions de Voronoi.

3.1. RÉGULARISATION STRUCTURELLE DU RISQUE EMPIRIQUE

La minimisation du risque empirique constitue un problème *mal-posé* [De Vito *et al.*, 2005] : la stabilité de la solution n'est pas garantie. Depuis les travaux de Tikhonov [1963], on sait que certains problèmes sont ramenés à des problèmes bien posés si l'on minimise un risque régularisé. Dans le cas du risque empirique, cela revient à minimiser une expression de la forme

$$R^*(f) = R_{emp}(f, E) + \lambda \Omega(f) \quad (3)$$

où $R_{emp}(f, E)$ est le risque empirique de f calculé sur l'échantillon E , Ω est une *fonctionnelle régularisante* et λ un *coefficient de régularisation*.

Le travail de Vapnik permet de ramener le problème de l'apprentissage par minimisation du risque empirique à un problème bien posé. Il propose une fonctionnelle régularisante pour le risque empirique. Ce travail conduit au principe de minimisation du risque structurel que nous présentons maintenant.

Supposons les réalisations de $E = (x_n, y_n)_N$ i.i.d. Plaçons-nous dans le cas d'une cible binaire : Y est à valeurs dans $\{0, 1\}$. Soit $\mathcal{C}$ un ensemble de classifieurs. Le *coefficient de pulvérisation* de $\mathcal{C}$, noté $\mathcal{N}^{\mathcal{C}}(E)$, est le cardinal de l'ensemble $\{(f(x_1), \dots, f(x_N)); f \in \mathcal{C}\}$. Intuitivement, ce coefficient quantifie le nombre de partitions de l'échantillon en deux groupes que réalisent les classifieurs de $\mathcal{C}$.

La *fonction de croissance* associée à $\mathcal{C}$ est la fonction de $\mathbb{N}$ dans $\mathbb{R}$, notée $\mathcal{G}^{\mathcal{C}}$, définie par

$$\forall N \in \mathbb{N}, \mathcal{G}^{\mathcal{C}}(N) = \log \sup_{E, \#E=N} \mathcal{N}^{\mathcal{C}}(E) \quad (4)$$

La *VC-dimension* de $\mathcal{C}$, notée $VC(\mathcal{C})$, est alors la plus grande valeur de N pour laquelle $\mathcal{G}^{\mathcal{C}}(N) = N \log 2$. La VC-dimension peut être infinie.

Le principe de minimisation du risque structurel consiste à minimiser conjointement le risque empirique et la borne sur l'écart entre cette erreur empirique et l'erreur théorique. La VC-dimension joue un rôle central dans l'expression de cette borne.

– *Principe de minimisation du risque structurel* [Vapnik, 1996]. Soit $E = (x_n, y_n)_N$ un échantillon de taille N . Soit $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(K)}, \dots$ une suite infinie croissante d'ensembles de classifieurs telle que $\mathcal{C} = \bigcup \mathcal{C}^{(K)}$ et telle que $VC(\mathcal{C}^{(K)})$ est finie pour tout K . Pour $f \in \mathcal{C}$, on note $VC(f)$ la VC-dimension du plus petit ensemble $\mathcal{C}^{(K)}$ contenant f . Pour un réel η fixé, le principe de minimisation du risque structurel (principe SRM en abrégé) préconise de sélectionner le classifieur f_{opt} tel que

$$f_{opt} = \arg \min_{f \in \mathcal{C}(E)} R_{emp}(f, E) + \sqrt{\frac{VC(f)(1 + \log(2N/VC(f))) - \log(\eta/4)}{N}} \quad (5)$$

Notons que la suite des VC-dimensions est croissante. Le risque empirique diminue avec l'augmentation de la VC-dimension alors que, N étant fixé, la régularisation augmente avec elle. La somme de ces deux termes définit le *risque structurel*.

En l'état, le principe SRM (pour Structural Risk Minimization) ne permet pas de travailler avec un ensemble de classifieurs dépendant des données.

3.2. CAS DE LA DÉPENDANCE AUX DONNÉES

L'ensemble des hypothèses que nous considérons est dépendant des données, ce dont ne tient pas compte la théorie initiale. Une adaptation des concepts de la théorie de Vapnik est proposée dans [Cannon *et al.*, 2002] afin de traiter ce cas. Nous la présentons brièvement.

Soit $\mathcal{C}$ l'ensemble des classifieurs binaires. Pour un classifieur $f : \mathbb{X} \rightarrow \{0, 1\}$ binaire quelconque, on note $I(f) = \{x \in \mathbb{X}; f(x) = 1\}$. La dépendance aux données est formalisée par l'existence d'une application associant à tout échantillon E un ensemble $\mathcal{C}(E)$ inclus dans $\mathcal{C}$ (par exemple, nous verrons ci-après le cas de l'application qui à un échantillon E associe l'ensemble des classifieurs par plus proche voisin obtenus par sélection d'individus dans E).

On suppose E i.i.d et Y binaire. Si $M \leq N$, on note $\mathcal{N}_M(E)$ le cardinal suivant :

$$\mathcal{N}_M(E) = \#\{E_M^{(x)} \cap I(f); E_M \subset E, \#E_M = M \text{ et } f \in \mathcal{C}(E_M)\} \quad (6)$$

où $E_M^{(x)}$ désigne un sous-ensemble de $\{x_n\}$ de taille M . Intuitivement, ce coefficient quantifie le nombre de partitions en deux groupes des individus de l'échantillon E réalisées par les classifieurs définis à partir d'un sous-échantillon de taille M inclus dans E .

Pour $M \leq N$, on définit le *coefficient de pulvérisation* $\mathcal{S}_{M/N}$ de $\mathcal{C}$ par la relation

$$\mathcal{S}_{M/N} = \sup_{E, \#E=N} \mathcal{N}_M(E) \quad (7)$$

Supposons donnée une application de $\mathcal{C}$ dans $\mathbb{N}$. Pour un échantillon E et $K \in \mathbb{N}$, notons $\mathcal{C}^{(K)}(E)$ l'ensemble des éléments de $\mathcal{C}(E)$ dont l'image est K par cette application. Notons $\mathcal{C}^{(K)}$ l'union des $\mathcal{C}^{(K)}(E)$ sur l'ensemble des échantillons et, pour $M \leq N$, $\mathcal{S}_{M/N}^{(K)}$ le coefficient de pulvérisation de $\mathcal{C}^{(K)}$.

Dès lors, il est possible de mettre en place un processus de décision analogue au principe SRM pour traiter des ensembles d'hypothèses dépendants des données.

– *Principe SRM avec dépendance aux données* [Cannon *et al.*, 2002]. Il est préconisé de sélectionner le classifieur f de $\mathcal{C}$ minimisant la quantité $R_{emp}(f) + r(K, M)$, où

$$r(K, M) = \sqrt{2 \frac{\log eM}{M \log M} \log \mathcal{S}_{M/N}^{(K)}} \quad (8)$$

avec $M = N/2$ et K le plus petit élément de $\llbracket 1, M \rrbracket$ tel que $f \in \mathcal{C}^{(K)}$. e désigne le nombre de Néper.

3.3. ÉVALUATION SRM DES PARTITIONS DE VORONOI

Soit H un ensemble de K individus de $\mathbb{I}$. Nous lui associons le classifieur f_H suivant. Pour un individu quelconque à classifier, on détermine le plus proche parmi les éléments de H . Notons-le k . On attribue à l'individu l'étiquette correspondant à la classe majoritaire dans $V(k, H)$. Nous notons $\mathcal{C}^{(K)}(\mathbb{I})$ l'ensemble des classifieurs obtenus à partir d'un sous-ensemble de cardinal au plus K de $\mathbb{I}$ (on reconnaît le $\mathcal{C}^{(K)}(E)$ de la section précédente après abus de notation). En reprenant les notations précédentes, $\mathcal{C}^{(K)}$ désigne l'union des $\mathcal{C}^{(K)}(\mathbb{I})$.

PROPOSITION 1 (Borne sur le coefficient de pulvérisation). *Pour $K \leq M$, le coefficient de pulvérisation de $\mathcal{S}_{M/N}^{(K)}$ de $\mathcal{C}_{(K)}$ est borné par $\binom{N}{M} \sum_{k=1}^K \binom{N}{k}$.*

Preuve. Par définition du coefficient de pulvérisation :

$$\mathcal{S}_{M/N}^{(K)} = \sup_{E, \#E=N} \mathcal{N}_M^{(K)}(E) \quad (9)$$

E étant un échantillon de cardinal N et $\mathcal{N}_M^{(K)}(E)$ le cardinal $\mathcal{N}_M(E)$ défini relativement à $\mathcal{C}^{(K)}$. On a la majoration suivante du cardinal $\mathcal{N}_M^{(K)}(E)$:

$$\begin{aligned} \mathcal{N}_M^{(K)}(E) &= \#\{E_M^{(x)} \cap I_f; E_M \subset E, \#E_M = M \text{ et } f \in \mathcal{C}_K(E_M)\} \\ &\leq \# \bigcup_{E_M \subset E, \#E_M=M} \bigcup_{f \in \mathcal{C}_K(E_M)} \{E_M \cap I_f\} \\ &\leq \binom{N}{M} \sum_{k=1}^K \binom{N}{k} \end{aligned} \quad (10)$$

3.4. CRITÈRE SRM

En accord avec l'approche SRM adaptée au cas d'un ensemble d'hypothèses dépendant des données, nous proposons d'évaluer une hypothèse H par le risque structural du classifieur f_H associé :

$$c^{SRM}(H) = R_{emp}(f_H) + \sqrt{\frac{4 \log e N}{N \log N} \left(\log \binom{N}{N/2} + \log \sum_{k=1}^K \binom{N}{k} \right)} \quad (11)$$

Nous illustrons le comportement de ce critère à l'aide d'une expérience sur données synthétiques. Nous générons uniformément 500 points dans un carré et considérons deux classes que nous distribuons uniformément et indépendamment. Dans ce cas, la variable cible est indépendante des variables descriptives.

Afin de visualiser le comportement du critère, nous définissons la notion de *trajectoire* d'optimisation. Les 500 individus sont tout d'abord ordonnés, aléatoirement.

FIGURE 2. Trajectoire d'optimisation du critère SRM : cas d'une cible indépendante des variables explicatives. La meilleure partition au sens du critère SRM est celle constituée d'un unique groupe.

La suite obtenue est parcourue et l'individu considéré à chaque étape est éliminé. À chaque étape, la partition basée sur l'ensemble des individus restant est évaluée. La suite de valeurs ainsi obtenue définit ce que nous appelons une trajectoire d'optimisation. La Figure 2 montre une trajectoire d'optimisation du critère SRM pour le jeu de données considéré. Notons que l'heuristique est sous-optimale mais qu'elle satisfait pleinement au but fixé dans cet article d'une comparaison qualitative des critères.

Cet exemple montre l'intérêt d'un critère régularisé non paramétrique. Parce qu'un tel critère permet de comparer des partitions de différentes tailles sans aucun paramètre à ajuster, son optimisation conduit à une sélection automatique du nombre de prototypes. Lorsque cela conduit à sélectionner la partition constituée par un seul groupe, ce qui est le cas ici, l'indépendance de la cible vis-à-vis des variables explicatives est caractérisée.

Nous illustrons maintenant les limites du critère SRM à travers une seconde expérience. Nous utilisons de nouveau un jeu de données synthétiques. Nous générons uniformément 2000 points dans un carré et considérons deux classes dont la distribution conditionnelle est $(0.9, 0.1)$ dans les coins supérieur droit et inférieur gauche, $(0.6, 0.4)$ dans les autres coins. Dans ce cas, la classe majoritaire est la même en tout point du carré bien que sa densité varie. Nous visualisons sur la Figure 3 une trajectoire obtenue pour le critère SRM.

De nouveau, l'optimisation du critère SRM conduit à sélectionner la partition de Voronoi constituée par une unique cellule. Mais, cette fois-ci, la cible n'est pas indépendante des variables explicatives. Parce qu'il emploie le risque empirique

FIGURE 3. Trajectoire d'optimisation du critère SRM : cas d'une classe toujours majoritaire. La meilleure partition au sens du critère SRM est celle constituée d'un unique groupe. La cible présente des variations de densité que le critère ne peut capturer.

comme mesure de qualité, le critère SRM ne mesure que le caractère majoritaire d'une cible. Il est aveugle aux différences de densité conditionnelle du moment que la classe majoritaire reste majoritaire. Le risque empirique et sa version régularisée manquent donc de finesse.

De plus, si le critère SRM repose sur des bases théoriques solides, celles de la théorie de l'apprentissage statistique, il a fallu faire des hypothèses pour l'obtenir. L'utilisation du critère n'est donc fondée que si :

- le problème de classification est binaire,
- la partition évaluée est constituée par moins de $N/2$ cellules,

Dans tout autre cas, son utilisation n'est plus qu'heuristique.

4. APPROCHE BIC : PROPOSITION D'UN CRITÈRE HEURISTIQUE

On se place dans cette section dans le cadre de la sélection de modèles. Celle-ci a pour objectif la comparaison de densités de probabilité aux formes différentes, par exemple des mélanges de gaussiennes et des modèles markoviens. Dans ce cas, la question première est celle du choix du modèle (gaussien ou markovien, dans l'exemple). Dans cette optique, un *modèle* est un *ensemble* d'hypothèses.

Dans l'article [Schwartz, 1978], une approche bayésienne de la sélection de modèle est proposée. Nous la rappelons ici. Cette approche peut être appliquée de manière heuristique afin d'opérer une sélection d'hypothèse. Nous appliquons cette

approche heuristique et proposons un nouveau critère pour notre problème de sélection d'hypothèse.

4.1. SÉLECTION BAYÉSIENNE DE MODÈLES ET CRITÈRE BIC DE SCHWARTZ

On considère un ensemble $\mathcal{H}$ d'hypothèses qui se décompose en l'union d'un nombre fini M d'ensembles de densités $\mathcal{H}_1, \dots, \mathcal{H}_M$, appelés *modèles*. Chaque modèle correspond à une forme de densité particulière. La question statistique posée est alors la détermination d'un modèle adéquat à partir d'un échantillon de données.

L'approche bayésienne de la sélection de modèle consiste à sélectionner le modèle le plus probable connaissant les données :

$$m_{opt} = \arg \max_{1 \leq m \leq M} p_E(m) \quad (12)$$

La question de la définition de la distribution de probabilité $p_E(m)$ est alors centrale. La formule de Bayes permet de définir p_E de manière indirecte :

$$p_E(m) = \frac{p(m)p_m(E)}{p(E)} \quad (13)$$

Un *a priori* uniforme étant posé sur les modèles (*i.e.* $p(m) = 1/M$) et constatant que la distribution $p(E)$ est indépendante du modèle, on est ramené à définir le terme $p_m(E)$ et à sélectionner le modèle

$$m_{opt} = \arg \max_{1 \leq m \leq M} p_m(E) \quad (14)$$

Dans [Schwartz, 1978], les modèles sont supposés paramétriques : pour $m \in \llbracket 1, M \rrbracket$, chaque élément de $\mathcal{H}_m$ est déterminé par un unique vecteur $\theta_m \in \Theta_m$ de paramètres réels de dimension d_m . La *probabilité a posteriori* $p_m(E)$ du modèle m est définie en intégrant la vraisemblance sur les éléments du modèle m puis en utilisant la formule de Bayes :

$$\begin{aligned} p_m(E) &= \int_{\Theta_m} p_m(E, \theta) d\theta \\ &= \int_{\Theta_m} p_m(E/\theta) q_m(\theta) d\theta \end{aligned} \quad (15)$$

Sous des hypothèses de dérivabilité de la vraisemblance vis-à-vis des paramètres, l'intégrale définissant $p_m(E)$ est approximée, et on obtient en éliminant les termes constants (*i.e.* ne dépendant pas de N)

$$\log p_m(E) = \log L(\hat{\theta}_m, E) - \frac{d_m}{2} \log N \quad (16)$$

où $\log L(\hat{\theta}_m, E)$ est la log-vraisemblance de l'hypothèse la plus vraisemblable dans $\mathcal{H}_m$, de paramètre $\hat{\theta}_m$.

Cette approche admet une interprétation heuristique. En effet, bien qu'issue du domaine de la sélection de modèles, l'approche de Schwartz a souvent été interprétée

et utilisée pour faire de la sélection d'hypothèses, de la manière suivante : pénaliser la log-vraisemblance par le nombre de paramètres multiplié par $-\frac{1}{2} \log N$ et choisir l'hypothèse maximisant le critère obtenu. Le critère obtenu, souvent qualifié de "BIC" (pour Bayesian Information Criterion), consiste en une pénalisation de la log-vraisemblance par une quantité dépendante du nombre de paramètres du modèle.

4.2. ÉVALUATION BIC HEURISTIQUE DES PARTITIONS DE VORONOI

Nous appliquons à notre tour l'approche heuristique inspirée par l'approche BIC de Schwartz. Nous commençons par définir la log-vraisemblance conditionnelle (analogue supervisé de la vraisemblance). Puis nous introduisons l'ensemble des densités conditionnelles à considérer, déterminons pour ces densités la formule analytique de la log-vraisemblance conditionnelle et terminons par l'évaluation du nombre de paramètres caractérisant une hypothèse.

Si E est un échantillon de N réalisations i.i.d. et pour une densité conditionnelle $f(\cdot/\cdot)$ sur $\mathbb{X} \times \mathbb{Y}$, on appelle *vraisemblance conditionnelle* de f relativement à E la quantité :

$$L(f, Y/X) = \prod_{n=1}^N f(y_n/x_n) \quad (17)$$

Pour faciliter les manipulations algébriques, on considère plutôt la *log-vraisemblance conditionnelle* $\log L(f, Y/X)$.

Associions maintenant à toute partition de Voronoi une densité de probabilité conditionnelle. Soit H un sous-ensemble de $\mathbb{I}$ de cardinal K et considérons la partition de Voronoi associée. Dans chaque cellule, on calcule les fréquences d'apparition des étiquettes sur les individus appartenant à la cellule. Pour un individu quelconque, on détermine sa cellule d'appartenance (*i.e.* on recherche le plus proche parmi les K prototypes) et on lui attribue la distribution fréquentielle des étiquettes de cette cellule. Nous notons $\mathcal{D}^{(K)}(\mathbb{I})$ l'ensemble des densités conditionnelles ainsi obtenues.

Soit $f \in \mathcal{D}^{(K)}(\mathbb{I})$ une densité conditionnelle associée à une partition de Voronoi en K cellules. La log-vraisemblance de f s'écrit :

$$\log L(f, Y/X) = \sum_{k=1}^K \sum_{j=1}^J N_{kj} \log N_{kj} - \sum_{k=1}^K N_k \log N_k \quad (18)$$

où J est le nombre de classes cibles, N_{kj} le nombre d'individus de l'échantillon dans la k^{eme} cellule portant la j^{eme} étiquette et N_k le nombre d'individus de l'échantillon dans la k^{eme} cellule.

En effet, la distribution des étiquettes est constante sur chaque cellule. Sur la k^{eme} cellule, elle vaut $\frac{1}{N_k}(N_{k1}, \dots, N_{kJ})$. Ainsi,

$$\begin{aligned} \log L(f, Y/X) &= \sum_{n=1}^N \log f(y_n/x_n) \\ &= \sum_{j=1}^J \sum_{k=1}^K N_{kj} \log \frac{N_{kj}}{N_k} \end{aligned} \quad (19)$$

FIGURE 4. Trajectoire d'optimisation du critère BIC : cas d'une classe toujours majoritaire avec des différences de densité. La meilleure partition au sens du critère BIC est constituée de sept groupes. Contrairement au critère SRM, le critère BIC est sensible aux variations de densité conditionnelle.

Pour une densité f de $\mathcal{D}^{(K)}(\mathbb{I})$ associée à un ensemble H de K prototypes, le nombre de paramètres la caractérisant est $J(K - 1) + 1$: on dispose de K degrés de liberté pour le choix des prototypes et de $(J - 1)(K - 1)$ degrés de liberté pour la spécification de la table de contingence (K distributions de J classes à spécifier, en connaissance des effectifs dans les groupes et des effectifs dans les classes).

4.3. CRITÈRE BIC

En suivant l'interprétation heuristique de l'approche bayésienne de Schwartz, nous proposons de pénaliser la log-vraisemblance d'une hypothèse par $-\frac{1}{2} \log N$ fois le nombre de paramètres, et donc de minimiser :

$$c^{BIC}(H) = \frac{J(K - 1) + 1}{2} \log N - \sum_{k=1}^K \sum_{j=1}^J N_{kj} \log N_{kj} + \sum_{k=1}^K N_k \log N_k \quad (20)$$

Montrons que le critère BIC procède à une évaluation plus fine que le critère SRM. Reprenons pour cela le protocole expérimental décrit dans la Section 2 permettant d'obtenir une trajectoire d'optimisation du critère BIC. Reprenons également le jeu de données de la seconde expérience, où l'une des deux classes est partout majoritaire mais présente des différences de densité. La trajectoire du critère BIC est reportée sur la Figure 4.

Le critère BIC procède à une évaluation plus fine que celle issue du critère SRM. L'optimisation du critère BIC mène à une solution basée sur 7 prototypes, là où l'emploi du critère SRM conduisait à la solution constituée par un unique groupe.

FIGURE 5. Trajectoire d'optimisation du critère BIC, cas du jeu de données réelles à 28 classes cibles Abalone. La meilleure solution est basée sur deux prototypes. La pénalisation BIC est très forte au regard de la vraisemblance.

Le critère BIC est sensible à toute variation de densité de la cible et pas seulement à celles conduisant à un changement de classe majoritaire. Le terme de vraisemblance tient compte de la distribution de toutes les étiquettes dans chaque groupe alors que le risque empirique focalise sur l'une des étiquettes dans chaque groupe, celle majoritaire.

Nous illustrons maintenant un défaut du critère BIC à l'aide d'une seconde expérience. Nous utilisons cette fois-ci un jeu de données réelles de l'UCI [Blake, Merz, 1996] : Abalone. C'est un problème de classification à 28 classes cibles. Une trajectoire d'optimisation du critère BIC est reportée sur la Figure 5.

Ce cas illustre la conséquence pratique de l'application heuristique de l'approche de Schwartz à la sélection d'hypothèses : le terme de pénalisation n'est pas bien calibré relativement à la vraisemblance. Sur le jeu de données Abalone, il domine fortement celle-ci et conduit à sélectionner un faible nombre de prototypes (2). La pénalisation du critère BIC manque donc de finesse.

5. ÉVALUATION BAYÉSIENNE DESCRIPTIVE

Dans [Ferrandiz, Boullé, 2006], une approche bayésienne de type MDL est adoptée. Un critère est proposé et nous rappelons dans cette section les étapes menant à son obtention : réécriture de l'inférence bayésienne dans le cas d'un ensemble d'hypothèses dépendant des données, proposition d'une distribution *a priori* sur l'ensemble des hypothèses, proposition d'une fonction de vraisemblance de la cible connaissant une hypothèse.

5.1. SÉLECTION BAYÉSIENNE D'HYPOTHÈSE ET ENSEMBLE D'HYPOTHÈSES DÉPENDANT DES DONNÉES

Si $\mathcal{H}$ est un ensemble d'hypothèses, le principe du maximum *a posteriori*, abrégé en MAP, préconise de sélectionner l'hypothèse la plus probable connaissant les données :

$$H_{opt} = \arg \max_{H \in \mathcal{H}} p_E(H) \quad (21)$$

Comme précédemment, l'inférence bayésienne ramène cette question à la définition d'une distribution de probabilité *a priori* p sur $\mathcal{H}$ et d'une fonction de vraisemblance q_H sur les données, pour tout $H \in \mathcal{H}$, du fait de la décomposition :

$$p_E(H) = p(H)q_H(E) \quad (22)$$

Dans notre cas, l'ensemble des hypothèses est fini et l'expression $\sum_H p(H)q_H(E)$ est définie et finie. On peut donc supposer que p_E est une distribution de probabilité, dite *a posteriori* et revenir à des notations plus classiques : pour toute hypothèse H , la quantité $p_E(H)$ est appelée *probabilité de H sachant E* et on la note $p(H/E)$.

L'ensemble d'hypothèses $\mathcal{H} = \mathcal{H}(\mathbb{I})$ dépend partiellement des données. Pour une hypothèse $H \in \mathcal{H}(\mathbb{I})$, nous proposons de réécrire la probabilité *a posteriori* $p(H/E)$ de la manière suivante :

$$\begin{aligned} p(H/E) &= \frac{p(H, E)}{p(E)} \\ &= \frac{p(\{x_n\})p(H/\{x_n\})p(\{y_n\}/H, \{x_n\})}{p(E)} \end{aligned} \quad (23)$$

En éliminant les termes ne dépendant pas de l'hypothèse H , le principe MAP s'écrit alors :

$$H_{opt} = \arg \max_{H \in \mathcal{H}(\mathbb{I})} p(H/\{x_n\})p(\{y_n\}/H, \{x_n\}) \quad (24)$$

L'inférence bayésienne passe par la proposition d'une distribution de probabilité $p_{\{x_n\}} := p(\cdot/\{x_n\})$ (toujours dite *a priori*) sur l'ensemble des hypothèses et d'une distribution $p_{H, \{x_n\}} := p(\cdot/H, \{x_n\})$ (toujours appelée *fonction de vraisemblance*) sur l'ensemble des étiquetages pour toute hypothèse $H \in \mathcal{H}(\mathbb{I})$.

5.2. PROPOSITION D'UNE DISTRIBUTION A PRIORI

Nous rappelons ici la distribution *a priori* $p_{\{x_n\}}$ proposée dans [Ferrandiz, Boullé, 2006]. Soit $H \in \mathcal{H}(\mathbb{I})$ une hypothèse, K le nombre de cellules.

La formule des probabilités itérées permet d'écrire :

$$p_{\{x_n\}}(H) = p_{\{x_n\}}(K, H) = q_{\{x_n\}}(K)p_{\{x_n\}}(H/K) \quad (25)$$

Pour obtenir une distribution *a priori* sur les hypothèses, il suffit alors de proposer une distribution $q_{\{x_n\}}$ sur le nombre de prototypes puis une distribution $p_{\{x_n\}}(\cdot/K)$ sur les partitions composées de K cellules.

Nous considérons le nombre K de prototypes comme un élément de l'ensemble $\llbracket 1, N \rrbracket$ et choisissons la distribution uniforme sur cet ensemble :

$$q_{\{x_n\}}(K) = \frac{1}{N} \quad (26)$$

La partition de Voronoi H en K cellules est définie par l'ensemble des K prototypes. Nous considérons cet ensemble de prototypes comme un élément de l'ensemble des K -combinaisons avec répétition dans un ensemble à N éléments. C'est un ensemble de cardinal $\binom{N+K-1}{K}$. Nous adoptons un *a priori* uniforme sur cet ensemble :

$$p_{\{x_n\}}(H/K) = \frac{1}{\binom{N+K-1}{K}} \quad (27)$$

5.3. PROPOSITION D'UNE FONCTION DE VRAISEMBLANCE

Soit $H \in \mathcal{H}(\mathbb{I})$ une hypothèse et soit K son nombre de cellules. Nous rappelons ici la fonction de vraisemblance $p_{H, \{x_n\}}$ proposée dans [Ferrandiz, Boullé, 2006].

Soit Y_k l'ensemble des étiquettes y_n telle que l'individu n appartient à la k^{eme} cellule de Voronoi $V(k, H)$ ($1 \leq k \leq K$). Nous supposons l'égalité suivante valide, synonyme d'indépendance conditionnelle :

$$p_{H, \{x_n\}}(\{y_n\}) = \prod_{k=1}^K p(Y_k) \quad (28)$$

Intuitivement, l'attribution des étiquettes est considérée indépendante d'une cellule à l'autre. C'est une hypothèse plus faible que l'hypothèse i.i.d classique entre les individus.

Soit $k \in \llbracket 1, K \rrbracket$. Notons N_{kj} le nombre d'individus dans la cellule k portant l'étiquette j . La formule des probabilités itérées permet d'écrire :

$$\begin{aligned} p(Y_k) &= p((N_{kj})_j, Y_k) \\ &= q((N_{kj})_j) p(Y_k / (N_{kj})_j) \end{aligned} \quad (29)$$

Pour obtenir une fonction de vraisemblance, il suffit alors de proposer une distribution q sur les fréquences puis une distribution $p(\cdot / (N_{kj})_j)$ sur les étiquetages, pour chaque cellule k .

Nous considérons le vecteur des fréquences $(N_{k1}, \dots, N_{kJ})$ des J classes cibles dans la cellule k comme un élément de l'ensemble des vecteurs de dimension J à coefficients entiers dont la somme vaut N_k . Cet ensemble a pour cardinal $\binom{N_k+J-1}{J-1}$. Nous choisissons la distribution uniforme sur cet ensemble :

$$q(N_{k1}, \dots, N_{kJ}) = \frac{1}{\binom{N_k+J-1}{J-1}} \quad (30)$$

Nous considérons l'ensemble Y_k comme un élément de l'ensemble des applications de $V(k, H) \cap H$ dans $\mathbb{L}$ dont N_{kj} individus ont pour image j ($1 \leq j \leq J$). Dénombrer cet ensemble revient à dénombrer les répartitions de N_k individus dans

FIGURE 6. Opposé de la log-vraisemblance et entropie. Ces quantités sont mesurées sur une trajectoire d'optimisation pour le jeu de données Spam. L'hypothèse d'indépendance sous-jacente à l'obtention du critère BIC est plus forte que celle relative au critère MDL. Cela se traduit par des valeurs du critère MDL plus faibles.

J boîtes sous contrainte d'en placer exactement N_{kj} dans la j^{eme} ($1 \leq j \leq J$). C'est un problème multinomial, le nombre de possibilités étant quantifié par le coefficient multinomial $\frac{N_k!}{N_{k1}! \dots N_{kJ}!}$. Nous obtenons, en choisissant la distribution uniforme :

$$p(Y_k / (N_{kj})_j) = \frac{N_{k1}! \dots N_{kJ}!}{N_k!} \quad (31)$$

5.4. LE CRITÈRE BAYÉSIEN DESCRIPTIF

Nous avons adapté le cadre de la sélection bayésienne d'hypothèses au cas d'un ensemble d'hypothèses dépendant des données. Nous avons explicité la probabilité *a posteriori* d'une hypothèse H . Maximiser cette probabilité revient à minimiser l'opposé de son logarithme, et nous proposons de minimiser le critère suivant :

$$c^{MDL}(H) = \log N + \log \binom{N + K - 1}{K} + \sum_{k=1}^K \log \binom{N_k + J - 1}{J - 1} + \log \frac{N_k!}{N_{k1}! \dots N_{kJ}!} \quad (32)$$

D'après les travaux de Shannon [1948], l'opposé du logarithme d'une probabilité s'interprète comme une longueur de mot d'un code. L'approche bayésienne adoptée entre dans le cadre des approches MDL (pour Minimum Description Length). C'est pourquoi nous qualifions le critère de MDL, dans le but de le différencier du critère BIC issu d'une approche bayésienne de la sélection de modèle. Notre objectif n'étant pas ici d'approfondir les relations entre bayésianisme et principe MDL, le lecteur intéressé peut se reporter à [Grünwald *et al.*, 2005] et [Vitanyi, Li, 2000].

Notons que le dernier terme du critère est maximal pour une distribution uniforme des étiquettes dans la cellule et nul pour une distribution pure dans la cellule (*i.e.* tous les individus de la cellule portent la même étiquette). C'est un comportement analogue à celui de l'entropie de Shannon et c'est pourquoi nous qualifions ce terme d'*entropique*.

D'après l'approximation de Stirling ($\log x! \approx x \log x - x + O(\log x)$), pour une partition de Voronoi $H \in \mathcal{H}(\mathbb{I})$ et lorsque N tend vers l'infini, on a :

$$\sum_{k=1}^K \log \frac{N_k!}{N_{k1}! \dots N_{kJ}!} \approx - \sum_{k=1}^K \sum_{j=1}^J N_{kj} \log N_{kj} + \sum_{k=1}^K N_k \log N_k \quad (33)$$

Asymptotiquement, le dernier terme du critère MDL et la log-vraisemblance du critère BIC sont égaux au signe près. Cela peut paraître étonnant, dans la mesure où les critères reposent sur des hypothèses d'indépendance différentes : indépendance des cellules pour le critère MDL, et indépendance des individus pour le critère BIC. C'est donc au niveau non-asymptotique qu'il faut chercher la différence.

Reprenons le protocole expérimental décrit précédemment et permettant d'obtenir une trajectoire d'optimisation de chacun des critères MDL et BIC. Nous nous intéressons ici uniquement au dernier terme du critère MDL et à l'opposé de la log-vraisemblance du critère BIC. Sur la Figure 6, nous reportons la différence entre la trajectoire de ces deux quantités pour le jeu de données Spam de l'UCI [Blake, Merz, 1996].

Dans cette expérience, N est fixé. Elle montre la différence pratique entre l'hypothèse i.i.d. régissant le critère BIC et l'hypothèse d'indépendance des cellules conduisant au critère MDL. La première est plus forte que la seconde, ce qui se manifeste par une valeur de l'opposé de la log-vraisemblance plus élevée que celle du dernier terme du critère MDL. Dans la mesure où ce sont ces deux termes qui quantifient l'information apportée par la partition sur la variable cible, on conclut de la manière suivante : le critère MDL quantifie mieux l'information sur la cible que le critère BIC grâce à une hypothèse d'indépendance plus faible. Le critère MDL montre déjà un peu plus de finesse que le critère BIC.

Mais c'est surtout la pénalisation du critère MDL qui est plus fine. Nous reportons sur la Figure 7 la trajectoire d'optimisation du critère MDL sur le jeu de données réelles Abalone exploité précédemment. La pénalisation du critère MDL est moins abrupte, plus équilibrée vis-à-vis du terme entropique. Le critère MDL est pénalisé plus finement que le critère BIC. Il en est plus discriminant. Le critère MDL conduit à sélectionner une partition en 13 cellules (2 cellules pour le critère BIC) et produit ainsi une connaissance plus riche. Cette expérience a été menée sur plusieurs autres jeux de données, et il a été constaté que le critère MDL extrait toujours une information plus fine que le critère BIC. Par exemple, sur le jeu de données à 10 classes cibles PenDigits [Blake, Merz, 1996], le critère BIC est minimal pour une partition en 38 cellules là où le critère MDL est minimal pour 89 cellules.

Pour extraire une connaissance plus riche, il est donc préférable d'optimiser le critère MDL.

FIGURE 7. Trajectoire d'optimisation du critère MDL sur le jeu de données Abalone. La meilleure solution au sens du critère MDL est composée de 13 prototypes. Une pénalisation plus fine et une hypothèse d'indépendance plus faible permettent d'extraire une information plus riche qu'avec le critère BIC (2 prototypes).

6. CONCLUSION

La sélection d'instances pour la classification suivant le plus proche voisin constitue un problème d'apprentissage statistique nécessitant l'évaluation d'hypothèses dans un ensemble à la fois structuré et dépendant des données. Afin d'aboutir à une décision complètement automatique, nous avons proposé trois nouveaux critères d'évaluation non paramétrique. Nous avons pour cela exploré trois approches de l'évaluation statistique régularisée : l'approche SRM, l'approche BIC et l'approche MDL. Pour chacune d'elles, nous avons proposé un nouveau critère.

Nous avons procédé à une comparaison qualitative de ces critères. À l'aide de jeux de données synthétiques et réels, nous avons illustré les différences de comportement des trois critères. La différence principale réside dans la finesse de l'évaluation et la richesse de la connaissance qu'on retire de l'utilisation d'un critère particulier.

L'approche structurelle, consistant à régulariser le risque empirique, fournit le critère le moins fin. En effet, le risque empirique se concentre uniquement sur les différences de classe majoritaire et est aveugle aux différences de densité conditionnelle. C'est un handicap, par exemple lorsque la classe d'intérêt est minoritaire, comme dans les problèmes de détection de fraude ou de caractérisation d'un comportement d'attrition ou d'appétence.

Les approches bayésiennes, une fois adaptées au cas des hypothèses dépendantes des données, fournissent des critères plus fins susceptibles de détecter des variations de densité conditionnelle. Des deux approches BIC et MDL, la seconde conduit au

critère le plus fin. En effet, elle repose sur une hypothèse d'indépendance plus faible et une régularisation plus fine, ce qui lui permet de mieux cerner l'information sur la cible apportée par un ensemble de prototypes.

Au final, c'est donc l'optimisation du critère MDL qui permet d'extraire l'information la plus riche, de manière non paramétrique.

BIBLIOGRAPHIE

- BLAKE C.L., MERZ C.J., *UCI repository of machine learning databases*, 1996, <http://www.ics.edu/mllearn/MLRepository.html>.
- CANNON A., ETTINGER J., HUSH D., SCOVEL C., "Machine learning with data dependent hypothesis classes", *Journal of machine learning research* 2, 2002, p. 335-358.
- DEVROYE L., GYÖRFI, LUGOSI G., *A probabilistic theory of pattern recognition*, Springer, 1996.
- FERRANDIZ S., BOULLÉ M., "Supervised evaluation of Voronoi partitions", *Journal of intelligent data analysis* 10(3), 2006, p. 269-284.
- FIX E., HODGES J., *Discriminatory analysis. Nonparametric discrimination: consistency properties*, Technical report 4, project number 21-49-004, USAF School of Aviation Medicine, Randolph Field (TX), 1951.
- GRÜNWALD P.D., MYUNG I.J., PITT M.A., *Advances in minimum description length: theory and applications*, MIT Press, 2005.
- HASTIE T., TIBSHIRANI R., FRIEDMAN J., *The elements of statistical learning*, Springer, 2001.
- LIU H., MOTODA H., *Instance selection and construction for data mining*, Norwell (MA), Kluwer, 2001.
- SCHWARTZ G., "Estimating the dimension of a model", *Annals of statistics* 6(2), 1978, p. 461-464.
- SHANNON C.E., "A mathematical theory of communication", Technical report, *Bell systems technical journal*, vol. 27, 1948, p. 379-423.
- STONE M., "Cross-validatory choice and assessment of statistical predictions", *Journal of the royal statistical society* 36(2), 1974, p. 111-147.
- TIKHONOV A.N., "On solving ill-posed problem and method of regularization", *Doklady Akademii Nauk, USSR* 153, 1963, p. 501-504.
- VAPNIK V., *The nature of statistical learning theory*, New-York, Springer-Verlag, 1996.
- VITANYI P.M.B., LI M., "Minimum description length induction, bayesianism, and Kolmogorov complexity", *IEEE Transactions of information theory* 46, 2000, p. 446-464.
- DE VITO E., ROSASCO L., CAPONNETTO A., DE GIOVANNINI U., ODOE F., "Learning from examples as an inverse problem", *Journal of machine learning research* 6, 2005, p. 883-904.
- WILSON D.R., MARTINEZ T.R., "Reduction techniques for instance-based learning algorithms", *Machine learning* 38(3), 2000, p. 257-286.