Prédictions contrôlées en apprentissage automatique

Alexander Gammerman & Vladimir Vovk

Computer Learning Research Centre, Department of Computer Science
Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK
alex@cs.rhul.ac.uk, vovk@cs.rhul.ac.uk

Résumé: Les récentes avancées obtenues en apprentissage automatique rendent possible la conception d'algorithmes efficients de prédiction pour des ensembles de données à grand nombre de paramètres. Cet article décrit une nouvelle méthode pour contrôler les prédictions élaborées par de nombreux algorithmes, incluant les machines à vecteurs support, la régression ridge à noyau, les plus proches voisins par noyau et bien d'autres méthodes correspondant à l'actuel état de l'art. Les prédictions contrôlées pour les étiquettes de nouveaux objets comportent des mesures quantitatives de leur précision et de leur fiabilité. Nous prouvons que ces mesures sont valides sous hypothèse de randomisation, traditionnelle en apprentissage automatique : les objets et leurs étiquettes sont supposés indépendamment générés par la même distribution de probabilité. En particulier, il devient possible de contrôler (aux fluctuations statistiques près) le nombre de prédictions erronées en choisissant un niveau de confiance approprié. La validité étant assurée, l'objectif restant pour les prédictions contrôlées est l'efficience : prendre au mieux les caractéristiques des nouveaux objets ainsi que l'information disponible pour produire des prédictions aussi précises que possible. Ceci peut être obtenu avec succès en utilisant toute la puissance des méthodes modernes d'apprentissage automatique.

Mots-clés: prédicteurs conformes, apprentissage en ligne, étrangeté d'une prédiction, induction, transduction

Abstract: Recent advances in machine learning make it possible to design efficient prediction algorithms for data sets with huge numbers of parameters. This article describes a new technique for `hedging' the predictions output by many such algorithms, including support vector machines, kernel ridge regression, kernel nearest neighbours, and by many other state-of-the-art methods. The hedged predictions for the labels of new objects include quantitative measures of their own accuracy and reliability. These measures are provably valid under the assumption of randomness, traditional in machine learning: the objects and their labels are assumed to be generated independently from the same probability distribution. In particular, it becomes possible to control (up to statistical fluctuations) the number of erroneous predictions by selecting a suitable confidence level. Validity being achieved automatically, the remaining goal of hedged prediction is efficiency: taking full account of the new objects' features and other available information to produce as accurate predictions as possible. This can be done successfully using the powerful machinery of modern machine learning.

Keywords: conformal predictors, on line procedure, strangeness, induction, transduction

1 Introduction

Les deux principaux aspects du problème de prédiction, la classification supervisée et la régression, sont des thèmes usuels de la statistique et de l'apprentissage automatique. Les techniques classiques de classification supervisée et de régression sont capables de traiter des ensembles de données conventionnels de faible dimension et petite taille ; cependant, les tentatives faites pour appliquer ces techniques à des ensembles actuels de données de grande dimension se heurtent à de sérieuses difficultés conceptuelles et calculatoires. Plusieurs nouvelles méthodes, et tout d'abord les machines à vecteur support [42,43] et autres méthodes à noyau, ont été récemment développées en apprentissage automatique, avec l'objectif explicite de traiter des ensembles de données de grande taille et de grande dimension.
Un inconvénient caractéristique de ces nouvelles méthodes est l'absence de mesure utilisable de confiance pour les prédictions. Par exemple, certaines limites supérieures strictes obtenues par la théorie d'apprentissage PAC (probablement approximativement correcte) pour la probabilité d'erreur dépasse souvent 1, même pour des ensembles de données relativement simples ([51], p. 249).

Dans cet article, nous décrivons une approche efficiente pour 'contrôler' les prédictions obtenues par les méthodes nouvelles et traditionnelles d'apprentissage automatique, c'est-à-dire, pour les compléter avec des mesures de leur précision et de leur fiabilité. Lorsqu'elles sont choisies correctement, ces mesures ne sont pas seulement valides et informatives, mais elles prennent aussi pleinement en compte les caractères spécifiques de l'objet à prédire.

Nous appelons prédicteurs conformes nos algorithmes pour obtenir des prédictions 'contrôlées' ; ils sont introduits formellement dans la section 3. Leur plus importante propriété est la validité automatique sous hypothèse de randomisation (qui sera discutée brièvement). Très simplement, nous entendons par validité le fait que les prédicteurs conformes ne surestiment jamais leurs prédictions en termes de précision et de fiabilité. Cette propriété, établie dans les sections 3 et 5, est formalisée en horizon fini, sans recourir à une forme asymptotique. L'établissement de la validité des prédicteurs conformes dépend d'une hypothèse requise par beaucoup d'autres algorithmes d'apprentissage automatique, et que nous appellerons hypothèse de randomisation : les objets et leurs étiquettes sont supposés être obtenus indépendamment à partir d'une même distribution de probabilité. On doit reconnaître qu'il s'agit d'une hypothèse forte, et des domaines en apprentissage automatique émergent en reposant sur d'autres hypothèses (telle que l'hypothèse de Markov en apprentissage renforcé ; voir par exemple [36]) ou même s'affranchissent complètement de toute hypothèse stochastique (apprentissage compétitif en ligne ; voir par exemple [6, 47]). Notre contrainte est cependant plus faible qu'une hypothèse de modèle statistique paramétrique, parfois complétée par une distribution a priori sur l'espace des paramètres, assez usuelle en théorie statistique de la prédiction. Elle n'apparaît pas abusivement restrictive si l'on tient compte de la puissance des propriétés qui peuvent être prouvées sous cette hypothèse.

Nous savons ainsi que les prédicteurs conformes disent vrai. Il est clair que c'est insuffisant : le vrai peut être sans valeur informative et par conséquent inutile. Nous évoquerons les différentes mesures d'apport informatif des prédicteurs conformes comme leur efficience. Etant donné qu'on peut prouver la validité des prédicteurs conformes, l'efficience est la seule caractéristique à laquelle nous nous intéresserons en construisant des prédicteurs conformes pour résoudre des problèmes spécifiques. Virtuellement, tout algorithme de classification supervisée ou de régression peut être transformé en prédicteur conforme, et ainsi la plupart des méthodes formant l'arsenal moderne de l'apprentissage automatique peut être rapporté à un cadre de prédiction conforme efficiente.

Nous commençons la partie principale de l'article, dans la section 2, avec la description d'un prédicteur idéal basé sur la théorie algorithmique de randomisation de Kolmogorov. Ce prédicteur 'universel ' donne les meilleures prédictions 'contrôlées' mais, malheureusement s'avère hors de portée des calculs. Nous chercherons, pour notre part, à approximer au mieux ce prédicteur universel.

Dans la section 3, nous introduisons formellement la notion de prédicteur conforme, et nous établissons un résultat simple concernant sa validité. Dans cette section, nous décrivons aussi brièvement les résultats expérimentaux décrivant la méthodologie de la prédiction conforme.

Dans la section 4 nous considérons un exemple mettant en évidence comment les prédicteurs conformes réagissent aux violations de notre modèle stochastique de génération des données (dans le cadre de notre hypothèse de randomisation). Si le modèle coïncide avec le véritable mécanisme stochastique, nous pouvons construire un prédicteur conforme optimal, qui s'avère au moins aussi bon que le prédicteur de confiance bayésien optimal (les définitions formelles sont données plus loin). Lorsque le mécanisme stochastique s'écarte significativement du modèle, les prédicteurs conformes restent valides, mais leur efficience en souffre inévitablement. Le prédicteur optimal bayésien produit d'abord des résultats trompeurs qui peuvent paraître superficiellement aussi bons que lorsque le modèle est correct.

Dans la section 5 nous décrivons le dispositif « en ligne » pour le problème de prédiction, et dans la section 6 nous le mettons en contraste avec le dispositif « regroupé » plus standard. La notion de validité introduite dans la section 3 s'applique à chacun des dispositifs, mais elle peut être renforcée dans le dispositif « en ligne » : nous pouvons maintenant prouver que le pourcentage de prédictions erronées peut être proche d'un niveau de confiance choisi avec une forte probabilité. Pour le dispositif « regroupé » la propriété de validité forte des prédicteurs conformes reste empirique. Dans la section 6, nous discutons des limitations du dispositif « en ligne » et nous introduisons des dispositifs intermédiaires entre les dispositifs « regroupé » et « en ligne ». Pour une large part, les prédicteurs conformes pour les dispositifs intermédiaires vérifient encore la propriété de validité forte.

La section 7 est réservée à la discussion de la différence entre les deux formes d'inférence à partir de données empiriques, l'induction et la transduction (soulignée par Vladimir Vapnik [42, 43]). La prédiction conforme est transductive, mais on peut arriver à une amélioration significative d'efficience calculatoire en combinant à des éléments inductifs (Section 8).

Nous montrons donc comment des méthodes très répandues d'apprentissage automatique peuvent être utilisées comme algorithmes sous-jacents en prédiction « contrôlée ». Nous ne donnons pas la description complète de ces méthodes, et nous renvoyons le lecteur aux descriptions existantes facilement accessibles. Cet article est cependant auto-suffisant puisqu'il explique toutes les caractéristiques des algorithmes utilisés pour aboutir à des prédictions « contrôlées ».
Nous espérons que le lecteur pourra ainsi appliquer nos techniques de contrôle aux méthodes d'apprentissage automatique qu'ils utilisent.

2 Prédictions contrôlées idéales

Le problème le plus élémentaire de l'apprentissage automatique est peut-être définit comme suit. On se donne un *ensemble d'apprentissage* formé d'*exemples* $(x_1,y_1),...,(x_l,y_l)$ (1), chaque exemple (x_i,y_i), i = 1, …, l, étant formé d'un *objet* x_i (habituellement le vecteur de ses attributs) et son *étiquette* y_i ; le problème posé est la prédiction du label y_{l+1} d'un nouvel objet x_{l+1}. Deux cas particuliers importants sont ceux où les étiquettes appartiennent *a priori* à un ensemble fini relativement petit (classification supervisée), et où les étiquettes peuvent être tout nombre réel (régression).
Le but usuel de la classification supervisée est de produire une prédiction $\hat{y}_{l+1}$ qu'on souhaite coïncider avec la véritable étiquette y_{l+1}, et l'objectif usuel de la régression est d'obtenir une prédiction $\hat{y}_{l+1}$ qu'on souhaite proche de la véritable étiquette y_{l+1}. Dans le cas de la classification supervisée, notre but est de compléter la prédiction $\hat{y}_{l+1}$ avec une mesure de sa fiabilité. Dans le cas de la régression, nous souhaitons disposer de mesures de la précision et de la fiabilité de notre prédiction. En raison de l'équilibre entre précision et fiabilité, on ne peut améliorer la première qu'au détriment de la seconde et vice-versa. Nous recherchons des algorithmes qui réalisent le meilleur équilibre et une mesure permettant de quantifier l'équilibre obtenu.
Commençons tout d'abord avec la classification supervisée. L'idée est d'examiner chaque étiquette possible Y et de regarder comment la séquence qui en résulte $(x_1,y_1),...,(x_l,y_l),(x_{l+1},Y)$ (2) se comporte par rapport à l'hypothèse de randomisation (si elle se conforme à cette hypothèse, nous dirons qu'elle est aléatoire ; ceci sera formalisé plus loin dans cette section). Le cas idéal est celui où tous Y sauf une conduisent à des séquences (2) non aléatoires ; on peut alors prendre l'étiquette Y restante comme prédiction de confiance pour y_{l+1}.
Dans le cas de la régression, nous pouvons considérer l'ensemble de tous les Y conduisant à une séquence (2) aléatoire comme notre 'ensemble de prédiction'. Un obstacle évident vient de ce que l'ensemble de tous les Y possibles est infini et que l'on ne peut donc pas développer explicitement tous les Y, mais nous verrons dans la section suivante des moyens de surmonter cette difficulté.Nous voyons ainsi que la question de la prédiction contrôlée est intimement associée au test du caractère aléatoire. Différentes versions de la notion universelle de randomisation ont été définies par Kolmogorov, Martin-Löf et Levin (cf. par exemple [24]) en se basant sur l'existence de machines de Turing universelles. Adaptée à notre contexte, la définition de Martin-Löf est la suivante. Soit Z l'ensemble de tous les exemples possibles (qu'on suppose être un espace mesurable) ; puisque chaque exemple est formé d'un objet et d'une étiquette, alors $Z = X \times Y$, où X est l'ensemble de tous les objets possibles, et Y, $|Y| > 1$, est l'ensemble de toutes les étiquettes possibles. Nous noterons Z^* l'ensemble de toutes les suites finies d'exemples. Une fonction $t : Z^* \to [0;1]$ est un test de randomisation si

1. pour tout $\varepsilon \in (0,1)$, tout $n \in \{1,2,...\}$ et toute distribution de probabilité P sur Z, on a
$$P^n\{z \in Z^n : t(z) \le \varepsilon\} \le \varepsilon \ ; \qquad (3)$$
2. t est semi-calculable supérieurement.

La première condition signifie la validité du test de randomisation ; si, par exemple, on a $t(z) \le 1\%$ pour l'ensemble de données z, alors soit l'ensemble n'a pas été obtenu par générations indépendantes à partir de la même distribution de probabilité P, ou bien un événement rare (de probabilité au plus égale à 1%, pour toute P) est survenu. La seconde condition signifie que le test doit pouvoir être calculable au sens faible (la calculabilité au sens usuel ne peut être requise, puisque le test universel peut seulement être semi-calculable supérieurement : il peut toujours être appliqué pour mettre en évidence *toutes* les structures dans les suites de

données qui les rendent non aléatoires). Martin-Löf (développant des idées antérieures de Kolmogorov) a prouvé qu'existe, à un facteur constant près, un plus petit test de randomisation.
Choisissons un plus petit test de randomisation, que nous appelons le *test universel*, et pour une suite de données nous désignons par *niveau de randomisation* de cette suite la valeur prise par ce test pour cette suite. Une suite aléatoire est alors une suite dont le niveau de randomisation n'est pas petit ; ceci est plutôt informel, mais il est clair que pour des suites finies de données on ne peut pas partager par une valeur nette toutes ces suites entre suites aléatoires et non aléatoires (comme ce qui est défini par Martin-Löf pour les suites infinies). Si t est un test de randomisation, non nécessairement universel, la valeur qu'il prend pour une suite de données sera appelée le *niveau de randomisation détecté par t*.

Remarque Le terme 'aléatoire' est utilisé dans la littérature avec au moins deux sens distincts. Dans cet article, nous avons besoin des deux mais, heureusement, la différence n'apparaît pas dans notre cadre actuel. Tout d'abord l'aléatoire se réfère à l'hypothèse de génération indépendantes des exemples à partir de la même distribution de probabilité ; ceci est l'origine de notre 'hypothèse de randomisation'. D'autre part, une suite de données est dite aléatoire par rapport à un modèle statistique si le test universel (généralisant la notion de test universel défini ci-dessus) ne détecte aucune différence de conformation entre les deux. Etant donné que le seul modèle statistique qui nous intéresse dans cet article est celui qui recouvre l'hypothèse de randomisation, nous avons une parfaite cohérence entre les deux acceptions.

Prédiction assortie de confiance et de crédibilité
Une fois choisi un test de randomisation t, universel ou non, il peut être utilisé pour la prédiction contrôlée. On a deux présentations naturelles de l'ensemble des résultats de telles prédictions : dans cette sous-section, nous décrivons celle qui peut seulement être utilisée en classification supervisée. Si le test de randomisation n'est pas calculable, on peut imaginer un oracle pour répondre aux questions relatives à ses valeurs.
Etant donné l'ensemble d'apprentissage (1) et l'objet test x_{l+1}, on peut opérer ainsi :

- envisager toutes les valeurs possibles $Y \in \mathbf{Y}$ pour l'étiquette y_{l+1} ;
- déterminer le niveau de randomisation détecté par t pour chaque solution (2) ;
- prédire l'étiquette Y correspondant à la solution de plus haut niveau de randomisation détecté par t ;
- poser comme *confiance* de cette prédiction le complément à 1 du deuxième plus haut niveau de randomisation détecté par t ;
- poser comme *crédibilité* de cette prédiction le niveau de randomisation détecté par t de la prédiction choisie Y (c'est-à-dire le plus grand niveau de randomisation détecté par t sur toutes les étiquettes possibles).

Pour comprendre le sens intuitif de cette confiance, prenons un niveau de signification conventionnel par exemple de 1% (dans la terminologie de cet article, ceci correspond à un niveau de confiance de 99% , soit 100% moins 1%). Si la confiance de notre prédiction est 99% ou plus, et que la prédiction est erronée, la suite de données relève d'un ensemble choisi *a priori*, de probabilité au plus égale à 1% (l'ensemble de toutes les séquences de données dont le niveau de randomisation détecté par t ne dépasse pas 1%)
Intuitivement, une faible crédibilité signifie soit que l'ensemble d'apprentissage est non aléatoire, ou bien que l'objet testé n'est pas représentatif de l'ensemble d'apprentissage (par exemple, l'ensemble d'apprentissage contient des représentations des chiffres, et l'objet testé est une lettre).

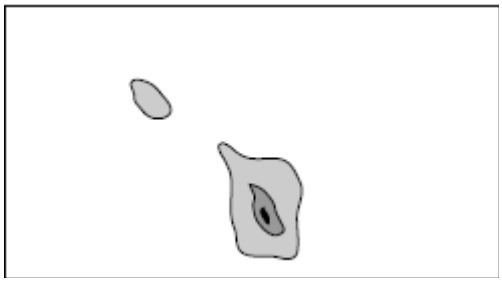

Figure 1 : Un exemple de famille emboîtée d'ensembles de prédiction (prédiction au hasard en noir, prédiction de confiance en gris sombre, et prédiction de confiance élevée en gris léger).

Prédicteurs de confiance

Dans les problèmes de régression, la confiance, ainsi que définie dans la précédente sous-section, n'est pas une grandeur utilisable ; elle prendrait la valeur 0. Une meilleure approche consiste à choisir une gamme de niveaux de confiance notés $1-\varepsilon$, et pour chacun d'eux à spécifier un *ensemble de prédiction* $\Gamma^{\varepsilon} \subseteq Y$, l'ensemble des étiquettes jugées possibles au niveau de confiance $1-\varepsilon$. Nous considérerons toujours des ensembles de prédictions emboîtés : $\Gamma^{\varepsilon_1} \subseteq \Gamma^{\varepsilon_2}$ lorsque $\varepsilon_1 \geq \varepsilon_2$. Un *prédicteur de confiance* est une fonction qui organise chaque ensemble d'apprentissage, chaque nouvel objet, et chaque niveau de confiance $1-\varepsilon$ (nous supposons que ε peut prendre toute valeur de (0 ,1)) dans l'ensemble de prédiction correspondant Γ^{ε}. Un prédicteur de confiance sera dit *valide* lorsque la probabilité que la véritable étiquette soit en dehors de l'ensemble de prédiction Γ^{ε} ne dépasse pas ε, pour tout ε.
Nous pouvons, par exemple, choisir les niveaux de confiance 99%, 95% et 80% et considérer l'ensemble de prédiction à 99% soit $\Gamma^{1\%}$ comme la prédiction de haute confiance, l'ensemble de prédiction à 95% soit $\Gamma^{5\%}$ comme l'ensemble de prédiction de confiance, et l'ensemble de prédiction à 80% soit $\Gamma^{20\%}$ comme la prédiction de hasard. La figure 1 montre ce à quoi peut ressembler une telle famille d'ensembles de prédiction dans le cas d'un espace *Y* rectangulaire. La prédiction de hasard cible bien l'objectif, mais nous savons que ce type de prédiction peut être erroné avec une probabilité de 20%. La prédiction de confiance est bien plus large. Si nous voulons avoir une prédiction de haute confiance (avec une probabilité d'erreur de 1%) nous devons accepter une précision beaucoup plus faible ; il y a même une localisation de prédiction possible totalement distincte que nous ne pouvons écarter à ce niveau de confiance.
Etant donné un test de randomisation, universel ou non, nous pouvons définir le prédicteur de confiance correspondant comme suit : pour tout niveau de confiance $1-\varepsilon$, l'ensemble de prédiction correspondant est formé par les *Y* tels que le niveau de randomisation de la solution (2) détectée par le test soit plus grand que ε. La condition de validité (3) pour les tests statistiques implique alors qu'un prédicteur de confiance défini de la sorte est toujours valide.
Le prédicteur de confiance basé sur le test universel (le *prédicteur universel de confiance*) est intéressant pour son étude mathématique (cf. par exemple [50], Section 4), mais il n'est pas calculable, et ne peut donc être utilisé en pratique. Notre objectif dans les sections qui suivent est de lui trouver des approximations calculables.

3 Prédiction conforme

Dans cette section, nous expliquons comment utiliser les tests de randomisation pour la prédiction. Le lien entre tests et prédiction est bien connu et il a longuement été discuté par les philosophes [32] et les statisticiens (cf. par exemple [9], section 7.5). Dans cette section nous regardons comment certains algorithmes très répandus de prévision peuvent être transformés en tests de randomisation, et de la sorte être utilisés pour obtenir des prédictions contrôlées.
Pour commencer, nous partons des machines à vecteur support ([42, 43], avec l'idée de revenir à la méthode du portrait généralisé [44]). Supposons que $Y = \{-1,+1\}$ (problème de classification supervisée binaire). A chaque ensemble d'exemples $(x_1,y_1),...,(x_n,y_n)$ (4), on associe un problème d'optimisation dont la résolution produit des nombres positifs (ou nuls) $\alpha_1,...,\alpha_n$ (multiplicateurs de Lagrange). Ces nombres déterminent la règle de prédiction utilisée par la machine à vecteur support (SVM) (cf. [43], Chapitre 10, pour les détails), mais ils sont aussi intéressants par eux-mêmes. Chaque α_i, i = 1, …, n, montre la particularité (étrangeté) de l'exemple correspondant (x_i,y_i) au sein de l'ensemble (4). Si α_i = 0, alors (x_i,y_i) s'ajuste parfaitement à l'ensemble (4) (en fait tellement bien que de tels exemples sont de fait non informatifs, et que la machine à vecteur support les ignore pour la prédiction). Les éléments tels que $\alpha_i > 0$ sont appelés vecteurs supports et une valeur élevée de α_i indique que l'exemple correspondant (x_i,y_i) est un élément éloigné (outlier).
En appliquant cette procédure à la complémentation (2) dans le rôle de l'ensemble (4) (de telle sorte que $n = l + 1$), on trouve les éléments $\alpha_1,...,\alpha_{l+1}$ correspondants. Si *Y* était différent de l'étiquette réelle y_{l+1}, on devrait avoir (x_{l+1},Y) comme élément éloigné dans l'ensemble (2) et ainsi α_{l+1} élevé en comparaison

avec $\alpha_1,...,\alpha_l$. Une façon naturelle pour comparer α_{l+1} aux autres valeurs α est d'examiner le rapport $p_Y = \frac{\left|\{i = 1,...,l+1 : \alpha_i \geq \alpha_{l+1}\}\right|}{l+1}$ (5) que nous appellerons la *p*-valeur associée au label possible *Y* pour x_{l+1}.
Ainsi la *p*-valeur est la proportion des valeurs de α qui sont au moins aussi élevée que le dernier α.
La méthodologie des SVM (telle que décrite dans [42, 43]) s'applique directement seulement au problème de classification supervisée binaire, mais le cas général peut se ramener au cas binaire par les procédures classiques 'un-contre-un' ou 'un-contre-le-reste'. Ceci nous permet de définir les valeurs d'étrangeté $\alpha_1,...,\alpha_{l+1}$ pour le problème général de classification supervisée (cf. [51], p. 59, pour les détails), qui à leur tour permettent de déterminer les *p*-valeurs (5).
La fonction qui associe à chaque suite (2) la *p*-valeur correspondante, définie par l'expression (5), est un test de randomisation (ce résultat provient du Théorème 1 établi dans la section 5 ci-après). Par conséquent les *p*-valeurs, qui sont nos approximations aux niveaux de randomisation correspondants, peuvent être utilisées pour la prédiction contrôlée ainsi que nous l'avons décrit dans la section précédente. Par exemple dans le cas de la classification supervisée binaire, si la *p*-valeur p_{-1} est faible alors que p_1 ne l'est pas, nous prédirons 1 avec une confiance $1 - p_{-1}$ et une crédibilité p_1. La crédibilité classique sera 1 : pour la plupart des ensembles de données le pourcentage des vecteurs supports est faible ([43], chapitre 12), et on s'attend alors à ce que $\alpha_{l+1}= 0$ lorsque $Y = y_{l+1}$.

Remarque Lorsque l'ordre des exemples n'est pas pris en compte, nous parlons de l'ensemble des données (4) comme d'un ensemble, bien que comme objet mathématique il soit plutôt un multi-ensemble puisqu'il peut contenir plusieurs fois le même exemple. Nous continuerons à utiliser cette terminologie informelle (pour être précis, on devrait dire multi-ensemble de données plutôt qu'ensemble de données).

0	1	2	3	4	5	6	7	8	9	Etiqu. réelle	Prédi-ction	Confi-ance	Crédi-bilité
0,01	0,11	0,01	0,01	0,07	0,01	100	0,01	0,01	0,01	6	*6*	99,89	100
0,32	0,38	1,07	0,67	1,43	0,67	0,38	0,33	0,73	0,78	6	*4*	98,93	1,43
0,01	0,27	0,03	0,04	0,18	0,01	0,04	0,01	0,12	100	9	*9*	99,73	100

Table 1: Quelques exemples tests de l'ensemble des données: les p-valeurs (en %) des chiffres 0-9, l'étiquette réelle et la prédiction, ainsi que la confiance et la crédibilité (en %).

La table 1 illustre les résultats de prédiction contrôlée pour un ensemble de données bien connu de chiffres manuscrits, appelé ensemble de données USPS [23]. Cet ensemble de données contient 9298 chiffres représentés chacun comme une matrice de 16x16 pixels; il est partagé en un ensemble d'apprentissage de taille 7291 et un ensemble test de taille 2007. La table montre, pour plusieurs exemples tests, les *p*-valeurs pour chaque étiquette possible, l'étiquette réelle, l'étiquette prédite, la confiance et la crédibilité calculées par une méthode SVM avec un noyau polynômial de degré 5. Pour interpréter les valeurs de cette table, on se rappellera qu'une confiance élevée (proche de 100%) signifie que toutes les étiquettes autres que celle qui est prédite sont peu vraisemblables. Si le premier exemple était prédit de manière erronée, ceci voudrait dire qu'un événement rare (de probabilité inférieure à 1%) est survenu; de la sorte, nous espérons une prédiction correcte (ce qu'elle est). dans le cas du second exemple, la confiance est aussi élevée (plus de 95%), mais on voit que la crédibilité es basse (moins de 5%). A partir de cette valeur de la confiance nous pouvons dire que les étiquettes autre que 4 sont exclues au niveau 5%, mais l'étiquette 4 est aussi exclue au niveau 5%. Ceci montre que l'algorithme de prédiction n'était pas capable d'extraire suffisamment d'information de l'ensemble d'apprentissage pour nous permettre d'affecter en confiance cet exemple; l'étrangeté des étiquettes autres que 4 peut provenir du fait que l'objet lui-même est étrange; cet exemple test est peut-être fortement différent de tous les exemples de l'ensemble d'apprentissage. Sans surprise, la prédiction pour ce second exemple est fausse.
En général une confiance élevée montre que toute alternative à l'étiquette prédite est peu vraisemblable. Une faible crédibilité signifie que tout l'exemple est suspect; comme nous l'avons déjà mentionné, nous obtiendrons une faible crédibilité si le nouvel exemple est une lettre alors que tous les exemples de l'ensemble d'apprentissage sont des chiffres. La crédibilité sera faible également si le nouvel exemple est un chiffre écrit de manière inhabituelle. Il faut noter que la crédibilité habituelle ne sera pas basse lorsque l'ensemble des données provient d'une génération indépendante à partir d'une même distribution: la probabilité que la crédibilité ne dépasse pas un certain seuil ε (tel que 1%) est au plus égale à ε. En résumé

nous pouvons accepter une prédiction si (1) sa confiance est proche de 100% et (2) sa crédibilité n'est pas trop basse (pas en dessous de 5%).
Beaucoup d'autres algorithmes de prédiction peuvent être pris comme algorithmes sous-jacents de prédiction contrôlée. Par exemple, on peut utiliser la technique des plus proches voisins pour associer

$$\alpha_i := \frac{\sum_{j=1}^{k} d_{ij}^{+}}{\sum_{j=1}^{k} d_{ij}^{-}}, \quad i = 1,\ldots, n, \qquad (6)$$

aux éléments (x_i, y_i) de l'ensemble (4), d_{ij}^{+} étant la $j^{\text{ème}}$ plus courte distance de x_i aux autres objets étiquetés comme x_i, et d_{ij}^{-} étant la $j^{\text{ème}}$ plus courte distance de x_i aux autres objets étiquetés différemment de x_i; le paramètre $k \in \{1,2,\ldots\}$ dans l'équation (6) est le nombre de plus proches voisins pris en compte. Les distances peuvent être calculées dans un espace de représentation (c'est-à-dire que la distance entre $x \in X$ et $x' \in X$ peut être envisagée comme $\|F(x) - F(x')\|$, F passant de l'espace des objets X dans un espace de représentation, usuellement de Hilbert), et ainsi la définition (6) peut aussi s'adapter aux plus proches voisins par noyau.

L'intuition qui sous-tend l'équation (6) est la suivante : un objet x_i étiqueté y tendra à être entouré par les autres objets étiquetés par y, et dans ce cas la valeur α_i correspondante sera faible. Dans le cas peu usuel où des objets étiquetés différemment de y sont plus proches que des objets étiquetés y, α_i augmentera. Par conséquent les coefficients α traduisent l'étrangeté des exemples.
Les p-valeurs calculées à partir de l'équation (6) peuvent également être utilisées pour des prédictions contrôlées. C'est un fait empirique général que la précision et la fiabilité des prédictions contrôlées sont liées au taux d'erreur de l'algorithme sous-jacent. Par exemple, pour les données USPS, l'algorithme du plus proche voisin (celui pour lequel $k = 1$) donne un taux d'erreur de 2,2%, et les prédictions contrôlées basées sur l'équation (6) sont de confiance élevée (au moins 99%) pour plus de 95% des exemples tests.

Définition générale

On peut généraliser la notion de prédicteur conforme comme suit. Une *mesure de non-conformité* est une fonction qui, à toute séquence de données (4), associe une suite de nombres $\alpha_1,\ldots,\alpha_n$ appelés *scores de non-conformité*, de telle sorte qu'interchanger deux exemples (x_i, y_i) et (x_j , y_j) conduit à interchanger les scores de non-conformité correspondants α_i et α_j, sans modifier les autres.
Le prédicteur conforme correspondant relie chaque ensemble de données (1), $l = 0,1, \ldots$, chaque nouvel objet x_{l+1}, et chaque niveau de confiance $1 - \varepsilon \in (0,1)$ à l'ensemble de prédiction

$$\Gamma^{\varepsilon}(x_1, y_1, \ldots, x_l, y_l, x_{l+1}) := \{Y \in \mathbf{Y} : p_Y > \varepsilon\} \qquad (7)$$

où p_Y est défini par l'équation (5), et $\alpha_1,\ldots,\alpha_{l+1}$ désignent les scores de non-conformité correspondant à la séquence (2).
Nous avons déjà remarqué que l'association à chaque solution (2) de la p-valeur (5) donne un test de randomisation ; ceci est vrai en général. Il s'ensuit que pour tout l la probabilité de l'événement $y_{l+1} \in \Gamma^{\varepsilon}(x_1, y_1, \ldots, x_l, y_l, x_{l+1})$ est au moins égale à $1 - \varepsilon$.
Cette définition est adaptée aussi bien à la classification supervisée qu'à la régression, mais dans le premier cas, on peut résumer les ensembles de prédictions par deux nombres

- la confiance $\sup\{1 - \varepsilon : |\Gamma^{\varepsilon}| \leq 1\}$
- et la crédibilité $\inf\{\varepsilon : |\Gamma^{\varepsilon}| = 0\}$

Régression et efficience calculatoire

Comme nous l'avons déjà mentionné, les algorithmes décrits ne peuvent être directement appliqués au cas de la régression, même si le test de randomisation est efficient au plan calculatoire: nous ne pouvons pas considérer toutes les valeurs possibles Y pour y_{l+1} en raison de leur infinité. Cependant on peut trouver des modes de calcul efficients pour déterminer les ensembles de prédiction Γ^{ε}. L'idée est que si les α_i sont

définis comme les résidus $\alpha_i := \left| y_i - f_Y(x_i) \right| (10)$ où $f_Y : X \to \mathbf{P}$ est une fonction de régression ajustée à l'ensemble complet (2), alors ces α_i peuvent avoir une expression simple en fonction de *Y*, donnant ainsi une méthode efficiente de calcul des ensembles de prédiction (par les équations (5) et (7)). Cette idée a été implémentée ([28]) dans le cas où f_Y est obtenu par régression ridge, ou régression ridge par noyau, l'algorithme de prédiction contrôlée qui en résulte étant appelé « *ridge regression confidence machine* ». Une description plus complète (et ses modifications dans le cas où les résidus simples (10) sont remplacés par les variantes telles que les résidus 'studentisés') est présentée dans [51], Section 2.3.

4 Approche bayésienne des prédictions conformes

Les méthodes bayésiennes sont devenues très utilisées à la fois en apprentissage automatique et en statistique grâce à leur puissance et à leur adaptabilité, et dans cette section nous examinons comment le schéma bayésien peut être utilisé pour construire des prédicteurs conformes efficients. Nous nous retreindrons ici à décrire des résultats expérimentaux obtenus par ordinateur ([26]) à partir d'ensemble de données artificiels, car pour des données réelles, on ne dispose pas de moyen pour vérifier la validité de l'hypothèse bayésienne.
On suppose que $X = \mathbf{R}^p$ (chaque objet est un vecteur à *p* composantes réelles) et notre modèle pour générer les données est $y_i \cdot x_i + \xi_i \quad i = 1,2,...,$ (11) où les ξ_i sont des aléas gaussiens centrés réduits indépendants, le vecteur poids $w \in \mathbf{R}^p$ étant distribué selon une loi multinormale $N\left(0;(1/a)I_p\right)$ (nous utilisons la notation I_p pour la matrice unité (p x p), et la notation $N(0;A)$ pour la loi de Gauss à *p* dimensions de vecteur moyen 0 et de matrice de variance-covariance *A*) ; *a* est une constante positive. Dans le modèle de génération utilisé pour nos données expérimentales, la constante *a* était égale à 1.
Le meilleur ajustement (au sens des moindres carrés) du modèle (11) à une ensemble de données (4) est obtenu par la procédure de régression ridge de paramètre *a* (cf. par exemple [51], Section 10.3 pour plus de détails). L'utilisation des résidus (10) avec f_Y obtenu par la régression ridge de paramètre *a* conduit à un prédicteur conforme efficient que nous appellerons *machine de confiance à régression ridge de paramètre a.* Chaque ensemble de prédiction produit par cette machine de confiance à régression ridge sera remplacé par son enveloppe convexe, dite *intervalle de prédiction* correspondant.
Pour tester la validité et l'efficience de cette machine de confiance à régression ridge, nous avons procédé comme suit. Dix vecteurs $w \in \mathbf{R}^5$ ont été générés indépendamment à partir de la distribution multinormale $N(0;I_5)$. Pour chacun de ces dix vecteurs *w*, 100 objets d'apprentissage et 100 objets tests ont été générés indépendamment à partir de la distribution uniforme sur $[-10; 10]^5$ et pour chaque objet *x* son label *y* a été généré selon $w \cdot x + \xi$, les ξ étant tous gaussiens centrés réduits, indépendants. Pour chacun des 1000 objets tests et chaque niveau de confiance $1-\varepsilon$, l'ensemble de prédiction Γ^ε de son label a été obtenu à partir de l'ensemble d'apprentissage correspondant en utilisant la machine de confiance à régression ridge de paramètre $a = 1$. La ligne pleine de la figure 2 trace le niveau de confiance en fonction du pourcentage d'exemples tests dont les étiquettes ne sont pas couvertes par les ensembles de prédictions correspondants pour ce niveau de confiance. Puisque les prédicteurs conformes sont toujours valides, le pourcentage situé en dehors de l'intervalle de prédiction ne doit jamais dépasser le complément à 100 du niveau de confiance, aux fluctuations statistique près, et ceci est confirmé sur le graphique.

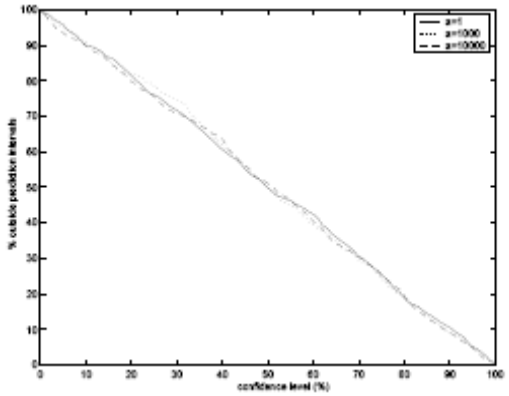

Figure 2: Validité de la machine de confiance à régression ridge.

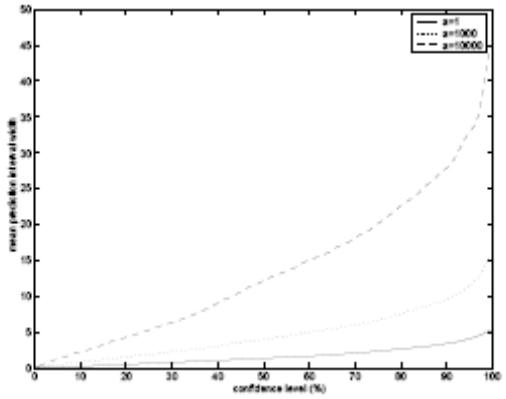

Figure 3: Efficience de la machine de confiance à régression ridge.

Une mesure naturelle de l'efficience des prédicteurs conformes est la largeur moyenne des intervalles de prédiction, pour différents niveaux de confiance : l'algorithme sera d'autant plus efficient que les intervalles de prédiction produits seront petits. Le tracé plein de la figure 3 présente la largeur moyenne (pour tous les exemples tests) des intervalles de prédiction produits à un niveau de confiance en fonction du niveau de confiance.

Puisque nous connaissons le mécanisme de génération des données, l'approche par les prédicteurs conformes peut apparaître quelque peu redondante : pour tout objet test nous pourrions plutôt rechercher la distribution de probabilité conditionnelle de son étiquette, qui est gaussienne, et donner comme ensemble de prédiction Γ^ε le plus petit (c'est-à-dire centre sur la moyenne de la distribution conditionnelle) intervalle de probabilité conditionnelle $1-\varepsilon$. Les figures 4 et 5 sont les analogues des figures 2 et 3 pour ce *prédicteur de confiance bayésien optimal* ; le tracé plein de la figure 4 montre sa validité.

Il est intéressant de noter que les lignes pleines des figures 5 et 3 ont la même apparence, en prenant en compte les différences d'échelles des axes verticaux. La machine de confiance à régression ridge apparaît d'aussi bonne qualité que le prédicteur bayésien optimal ; ceci est un phénomène général ; il est aussi illustré dans le cadre de la classification supervisée, par la construction ([51], section 3.3) d'un prédicteur conforme asymptotiquement aussi bon que le prédicteur de confiance bayésien optimal.

La similarité entre les deux algorithmes disparait lorsque l'on donne de mauvaises valeurs pour a. Par exemple regardons ce qui se passe si nous indiquons dans l'algorithme que la moyenne de $\|w\|$ est seulement de 1% de ce qu'elle est en réalité (ceci revient à prendre a = 10000). La machine de confiance à régression ridge reste valide (cf. le tracé à tirets de la figure 2), mais son efficience se détériore (le tracé à tirets de la figure 3). L'efficience du prédicteur de confiance bayésien optimal (tracé à tirets de la figure 5) est durement affectée, et ses prédictions deviennent invalides (le tracé à tirets de la figure 4 s'écarte significativement de la diagonale, particulièrement pour les niveaux de confiance élevés : ainsi seulement environ 15% des étiquettes sont dans les intervalles de prédiction à 90%). Le pire pouvant se produire à la machine de confiance à régression ridge est que ses prédictions deviennent inutilisables (au moins sans intérêt), alors que les prédictions bayésiennes optimales peuvent devenir trompeuses.

Les figures 2-5 montrent aussi les tracés pour la valeur intermédiaire a = 1000. Des résultats similaires mais pour des ensembles de données différents sont présentés dans la section 10.3 de [51]. Un schéma général de la prédiction conforme de type bayésien est donné dans [51], pp 102-103.

5 Prédiction « en ligne »

Dans la section 3 nous avons établi la validité des prédicteurs conformes avec le sens que la probabilité d'erreur $y_{l+1} \notin \Gamma^\varepsilon\left(x_1, y_1, \ldots, x_l, y_l, x_{l+1}\right)$ (12) ne dépasse pas ε, pour un niveau de confiance $1-\varepsilon$. Le terme probabilité dans le sens 'probabilité inconditionnelle' est ramené à l'acception fréquentiste : si on répète plusieurs générations de séquences $\left(x_1, y_1, \ldots, x_l, y_l, x_{l+1}, y_{l+1}\right)$, la proportion de celles qui vérifient (12) sera au plus égale à ε, aux fluctuations statistiques près. Affirmer que nous contrôlons ainsi le nombre d'erreurs serait exagéré en raison du caractère artificiel de ce procédé de génération répétée d'un ensemble

d'apprentissage et d'un nouvel exemple test. Peut-on dire que le niveau de confiance $1-\varepsilon$ traduit une borne du nombre d'erreurs d'un protocole d'apprentissage ? Dans cette section, nous donnons une réponse positive pour le protocole en ligne le plus répandu, et dans la section suivante nous montrerons comment on peut généraliser aux autres protocoles.

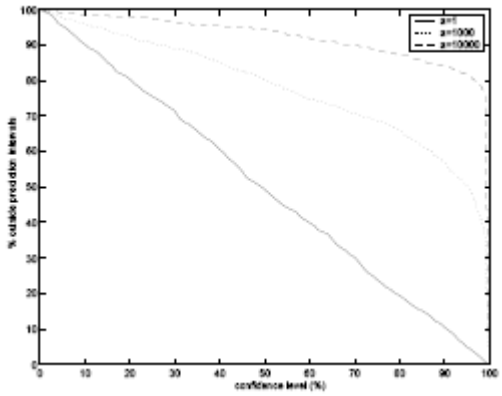

Figure 4 : Validité du prédicteur de confiance bayésien optimal

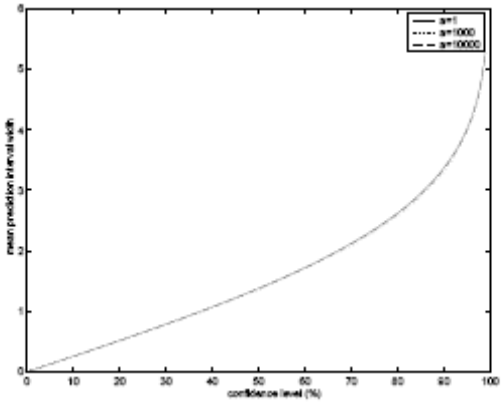

Figure 5 : Efficience du prédicteur de confiance bayésien optimal

En apprentissage en ligne les exemples se présentent un par un. A chaque instant, on observe un objet et on prédit son étiquette. Puis on observe son étiquette réelle et on passe à l'exemple suivant. Nous commençons donc par observer le premier exemple x_1 et par prédire son étiquette y_1. On observe ensuite l'étiquette y_1 et le second objet x_2 pour prédire son étiquette y_2, et ainsi de suite. A l'étape n, ayant observé les exemples précédents $\left(x_1, y_1, \ldots, x_{n-1}, y_{n-1}\right)$, on observe le nouvel objet x_n et notre but est de prédire son étiquette y_n.

La qualité de nos prédictions doit s'améliorer avec l'accumulation des exemples traités. Ceci est le sens de notre apprentissage.

Notre prédiction pour y_n est une famille emboîtée d'ensemble de prédictions $\Gamma_n^\varepsilon \subseteq Y, \varepsilon \in (0,1)$.

Le processus de prédiction en ligne peut être synthétisé par le protocole suivant :

$Err_0^\varepsilon := 0;$

$Mult_0^\varepsilon := 0;$

$Emp_0^\varepsilon := 0;$

FOR n = 1, 2, … :

objet $x_n \in X$;

prédictions $\Gamma_n^\varepsilon \subseteq Y, \ \forall\ \varepsilon \in (0,1)$;

étiquette $y_n \in Y$;

$$err_n^\varepsilon := \begin{cases} 1 \text{ si } y_n \notin \Gamma_n^\varepsilon \\ 0 \text{ sinon} \end{cases} \quad \varepsilon \in (0,1);$$

$$Err_n^\varepsilon := Err_{n-1}^\varepsilon + err_n^\varepsilon, \varepsilon \in (0,1);$$

$$mult_n^\varepsilon := \begin{cases} 1 \text{ si } \left|\Gamma_n^\varepsilon\right| > 1 \\ 0 \text{ sinon} \end{cases} \varepsilon \in (0,1);$$

$$Mult_n^\varepsilon := Mult_{n-1}^\varepsilon + mult_n^\varepsilon, \varepsilon \in (0,1);$$

$$emp_n^\varepsilon := \begin{cases} 1 \text{ si } \left|\Gamma_n^\varepsilon\right| = 0 \\ 0 \text{ sinon} \end{cases} \varepsilon \in (0,1);$$

$$Emp_n^\varepsilon := Emp_{n-1}^\varepsilon + emp_n^\varepsilon, \varepsilon \in (0,1);$$

END FOR.

La famille Γ_n^ε est supposée emboîtée: $\Gamma_n^{\varepsilon_1} \subseteq \Gamma_n^{\varepsilon_2}$ si $\varepsilon_1 \geq \varepsilon_2$. Dans ce protocole nous retenons également pour chaque niveau de confiance $1-\varepsilon$ les valeurs cumulées d'ensembles de prédiction erronés Err_n^ε, d'ensemble de prédiction multiples (c'est-à-dire d'ensembles de prédiction contenant plus d'une étiquette) $Mult_n^\varepsilon$, et d'ensembles de prédiction vides Emp_n^ε. Nous discuterons le sens de chacune de ces valeurs.

Le nombre de prédictions erronées est une mesure de validité de nos prédicteurs de confiance : nous cherchons à ce que $Err_n^\varepsilon \leq \varepsilon n$, aux fluctuations statistique près. Dans la figure 6, nous pouvons voir les courbes $n \mapsto Err_n^\varepsilon$ pour un prédicteur conforme particulier et à trois niveaux de confiance : le tracé plein pour 99%, le tracé en tiret-point pour 95% et le pointillé pour 80%. Le nombre d'erreurs croît linéairement, et les pentes sont environ 0,2 pour le niveau de confiance 80%, 0,05 pour le niveau de confiance 95%, et 0,01 pour le niveau de confiance 99%. Nous allons montrer que ce n'est pas fortuit.

Le nombre de prédictions multiples $Mult_n^\varepsilon$ est une mesure intéressante de l'efficience dans le cas de la classification supervisée : nous souhaitons le plus possible de singletons dans nos prédictions. La figure 7 montre la courbe des erreurs cumulées $n \mapsto Err_n^{2,5\%}$ (tracé plein) et celle des prédictions multiples $n \mapsto Mult_n^{2,5\%}$ (en pointillés) pour un niveau de confiance fixé de 97,5%. On peut voir qu'environ 250 (à peu près 2,5%) étaient erronées et environ 300 (à peu près 3%) des prédictions multiples.

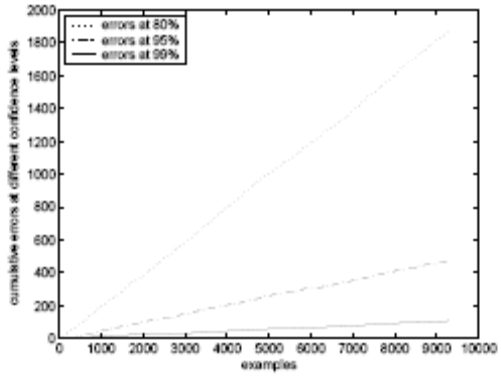

Figure 6 : Nombres cumulés d'erreurs pour un prédicteur conforme (le prédicteur conforme du plus proche voisin) appliqué en mode en ligne au données USPS (9298 chiffres manuscrits, après permutation aléatoire) pour les niveaux de confiance 80%, 95% et 99%.

On voit donc que par le choix de ε, on peut contrôler le nombre d'erreurs. Pour une valeur faible de ε (par rapport à la complexité de l'ensemble des données) ceci peut conduire à devoir donner parfois des prédictions multiples. D'autre part, pour une valeur élevée de ε, on peut aboutir à des ensembles de prédiction vides à certaines étapes, comme on peut le voir en bas à droite de la figure 7 : lorsque le prédicteur ne fait plus de prévisions multiples, il commence à faire occasionnellement des prédictions vides (courbe en tiret-point). Une prédiction vide avertit que l'objet à prédire est inhabituel (sa crédibilité, telle que définie en section 2, est au plus de ε).

Ce serait une erreur que de se concentrer sur un seul niveau de confiance $1-\varepsilon$. Un ensemble de prédiction Γ_n^ε vide ne signifie pas que l'on ne peut faire aucune prédiction : on doit simplement s'attacher à d'autres niveaux de confiance (par exemple examiner la gamme de valeurs de ε pour laquelle Γ_n^ε est un singleton). De même un ensemble de prédiction Γ_n^ε multiple ne signifie aucunement que toutes les étiquettes de Γ_n^ε sont aussi vraisemblables : une légère augmentation de ε peut conduire à l'élimination de certaines étiquettes. Bien entendu, la prise en compte de la continuité de la suite des ensembles de prédictions pour toutes les valeurs $\varepsilon \in (0,1)$ est trop difficile et un compromis raisonnable consiste à se focaliser sur quelques niveaux classiques, comme sur la figure 1.

Par exemple, la Table 2 donne les *p*-valeurs pour les différents types de douleur abdominale évoqués pour un patient donné à partir de ses symptômes. On voit qu'au niveau de confiance 95% l'ensemble de prédiction est multiple, {cholécystite, dyspepsie}. Si on relâche le niveau de confiance à 90%, l'ensemble de prédiction se réduit à un singleton {dyspepsie} ; et d'autre part au niveau de confiance 99%, l'ensemble de prédiction s'élargit à {appendicite, douleur abdominale non spécifique, cholécystite, pancréatite, dyspepsie}. Une telle information détaillée, en combinaison avec la propriété de validité, est particulièrement intéressante en médecine (certaines des premières applications de la prédiction conforme ont été développées dans les champs de la médecine et de la bioinformatique : cf. par exemple [3, 35]).

Dans le cadre de la régression, on aura habituellement $Mult_n^\varepsilon = n$ et $Emp_n^\varepsilon = 0$, ce qui rend ses mesures peu utiles pour évaluer l'efficience. Des mesures plus utiles, telles que celles utilisées dans la section précédente, pourraient, par exemple, prendre en compte les largeurs des intervalles de prédiction.

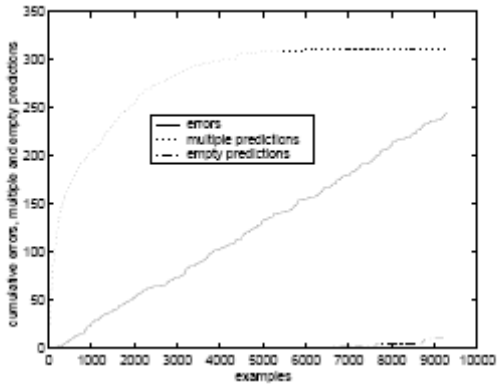

Figure 7 : les performances en ligne du prédicteur conforme au plus proche voisin au niveau de confiance 97,5%, dans le cas des données USPS (après permutation aléatoire).

Table 2 : un exemple tiré d'un ensemble de données enregistrées sur des patients souffrant de douleur abdominale aiguë [15]: les *p*-valeurs pour les neuf groupes de diagnostic possibles (appendicite APP, diverticulose DIV, ulcère perforé PPU, douleur abdominale non spécifique NAP, cholécystite CHO, occlusion intestinale INO, pancréatite PAN, colique néphrétique RCO, dyspepsia DYS), et l'étiquette réelle.

APP	DIV	PPU	NAP	CHO	INO	PAN	RCO	DYS	Etiq. réelle
1,23%	0,36%	0,16%	2,83%	5,72%	0,89%	1,37%	0,48%	80,56%	DYS

Analyse théorique

En regardant les figures 6 et 7, on peut être tenté de croire que la probabilité d'erreur à chaque étape du protocole en ligne est e et que les erreurs surviennent indépendamment aux différentes étapes. Ceci n'est pas exact comme le révèle un examen attentif du coin en bas à gauche de la figure 7. Cependant, ceci devient exact (comme noté dans [48]) si on redéfinit les *p*-valeurs (5) selon

$$p_Y = \frac{\left|\left\{i : \alpha_i \geq \alpha_{l+1}\right\}\right| + \eta\left|\left\{i : \alpha_i = \alpha_{l+1}\right\}\right|}{l+1} \quad (13)$$

avec $i = 1, \ldots, l+1$, et $\eta \in (0,1)$ obtenu par génération aléatoire dans une distribution uniforme sur (0,1) (les valeurs de η doivent être indépendantes entre elles, et indépendantes de tout autre élément ; en pratique elles sont obtenues par des générateurs de nombres pseudo-aléatoires). La seule différence entre les équations (5)

et (13) est la prise en compte des éléments de queue $\alpha_i = \alpha_{l+1}$ dans l'expression (13). En remplaçant l'expression (5) par l'expression (13) dans la définition des prédicteurs conformes, on obtient la notion de *prédicteur conforme lissé*.
On peut maintenant établir comme suit la propriété de validité des prédicteurs conformes lissés :

Théorème 1 On suppose que les exemples $(x_1, y_1), (x_2, y_2), \ldots$ sont générés indépendamment à partir de la même distribution de probabilité. Pour tout prédicteur conforme lissé opérant selon le protocole de prédiction en ligne, et pour tout niveau de confiance $1-\varepsilon$, les variables aléatoires $err_1^\varepsilon, err_2^\varepsilon, \ldots$ sont indépendantes et prennent la valeur 1 avec la probabilité ε.

En combinant le théorème 1 et la loi forte des grands nombres, on voit que l'événement $\lim_{n\to\infty} \frac{Err_n^\varepsilon}{n} = \varepsilon$ est de probabilité 1 pour les prédicteurs conformes lissés (qui sont donc 'bien calibrés'). Puisque le nombre d'erreurs faites par un prédicteur conforme ne dépasse jamais le nombre d'erreurs faites par le prédicteur conforme lissé correspondant, l'événement $\limsup_{n\to\infty} \frac{Err_n^\varepsilon}{n} \le \varepsilon$ est de probabilité 1 pour les prédicteurs conformes (qui sont alors 'conservativement bien calibrés').

6 Formateurs lents, formateurs paresseux et mode regroupé.

Dans le mode en ligne strict, envisagé dans la section précédente, nous avons une information en retour immédiate (la véritable étiquette) pour tout exemple soumis à prédiction, et ceci crée des interrogations pour l'application pratique de ce scénario. Imaginons par exemple un centre de tri postal utilisant un algorithme de prédiction en ligne pour l'identification du code postal ; supposons que l'information en retour relative à la véritable étiquette provienne d'un 'formateur' humain. Si cette information est donnée pour tout objet x_i, aucun point n'est alors tributaire de l'algorithme de prédiction ; nous pouvons utiliser autant que nous le voulons l'information donnée par le 'formateur'. Il serait intéressant que l'algorithme de prédiction fonctionne encore bien, et en particulier soit valide, si l'affectation par une formateur humain n'était faite que, par exemple, une fois tous les dix objets (scénario du formateur 'paresseux'). Alternativement, même si l'algorithme de prédiction nécessite la connaissance de toutes les étiquettes, il peut cependant être utile si les étiquettes ne sont pas données immédiatement mais avec un certain délai (formateurs 'lents'). Dans notre exemple de tri postal un tel délai peut permettre la connaissance de toutes les erreurs possibles avant de disposer de l'information en retour. Dans le mode en ligne strict la validité est obtenue au sens le plus strict possible : pour tout niveau de confiance $1-\varepsilon$ les erreurs de tout prédicteur conforme lissé sont indépendantes et de probabilité ε. Dans le cas de formateurs faibles (nous utilisons le terme 'formateur' dans le sens très général de l'entité fournissant l'information en retour, que nous avons appelée réalité dans la section précédente), nous devons accepter une notion de validité plus faible. Supposons que le prédicteur reçoive du formateur une information en retour à la fin des étapes n_1, n_2, …, telles que $n_1 < n_2 < \ldots$; cette information retour est l'étiquette de l'un des objets que le prédicteur a déjà examiné (et traité pour prédire). Ce schéma [33] concerne autant les formateurs 'lents' que les formateurs 'paresseux' (ainsi que les formateurs à la fois lents et paresseux). Il a été prouvé dans [29] (voir aussi [51], Théorème 4.2) que les prédicteurs conformes lissés (utilisant les seuls exemples d'étiquettes connues) restent valides au sens suivant $\forall \varepsilon \in (0,1): \mathrm{Err}_n^\varepsilon / n \to \varepsilon \ (\text{si } n \to \infty)$ en probabilité si et seulement si $\lim_{k\to\infty} (n_k / n_{k-1}) = 1$
En d'autres termes, il y a validité au sens de la convergence en probabilité si et seulement si le taux de croissance de n_k est sous-exponentiel (cette condition est largement satisfaite dans notre exemple de formateur donnant un retour tous les dixièmes objets).
Le mode regroupé le plus standard pour le problème de prédiction est d'une certaine manière moins exigeant que nos scénarii de formateurs faibles. Dans ce mode, nous disposons d'un ensemble d'apprentissage (1) et notre but est de prédire les étiquettes en se donnant les objets de l'ensemble test

$$(x_{l+1}, y_{l+1}), \ldots, (x_{l+k}, y_{l+k}) \qquad (14)$$

Ceci peut être interprété comme une version à distance finie du contexte du formateur paresseux : aucune étiquette n'est donnée en retour après l'étape l. Des expériences (voir par exemple la figure 8) montrent que

la validité approximative est encore acquise ; les résultats théoriques associés sont présentés dans [51], Section 4.4.

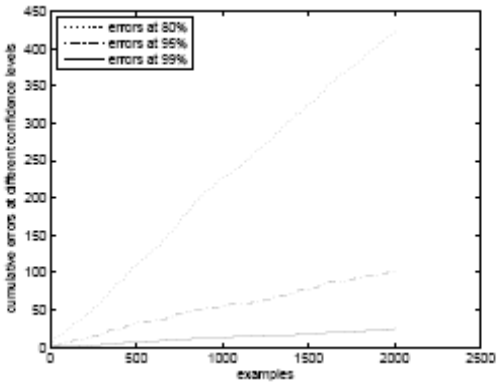

Figure 8 : Fréquences cumulées des erreurs faites sur l'ensemble test par le prédicteur conforme du plus proche voisin utilisé en mode regroupé sur les données USPS (après permutation aléatoire et partage en un ensemble d'apprentissage de taille 7291 et un ensemble test de taille 2007) aux niveaux de confiance 80%, 95% et 99%.

7 Induction et transduction

La distinction entre induction et transduction introduite par Vapnik [42, 43] , appliquée au problème de la prédiction, est décrite dans la figure 9. Dans la prédiction inductive, nous passons d'abord des exemples disponibles à une règle plus ou moins générale, que nous appelons règle de prédiction ou de décision, modèle ou théorie ; c'est l'étape inductive. Lorsqu'elle est associée à un nouvel objet, nous obtenons une prédiction à partir de la règle générale ; c'est l'étape déductive. Dans la prédiction transductive, nous faisons un raccourci en passant directement des exemples anciens à la prédiction concernant un nouvel objet.
Des exemples typiques de l'étape inductive sont fournis pas l'estimation de paramètres en statistique classique ou la recherche d'une fonction d'approximation en apprentissage statistique. Des exemples de prédiction transductive sont fournis par l'estimation d'observations futures (anticipations) en statistique classique ([9], Section 7.5, [38]) ou les algorithmes des plus proches voisins en apprentissage.

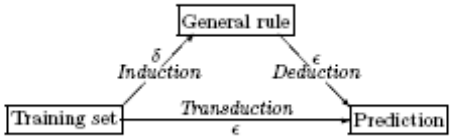

Figure 9 : Prédiction inductive et prédiction transductive

Dans le cas de prédictions simples (c'est-à-dire traditionnelles, et non encloses) la distinction entre induction et transduction est bien moins précise. Une méthode pour traiter la transduction, dans le contexte simple de la prédiction d'un label, est une méthode pour prédire y_{l+1} à partir de l'ensemble d'apprentissage (1) et de x_{l+1}. Une telle méthode donne une prédiction pour tout objet pouvant se présenter comme x_{l+1}, et ainsi elle définit, au moins implicitement, une règle qui peut être extraite de l'ensemble d'apprentissage (1) (induction), mise en réserve, et ensuite appliquée à x_{l+1} pour prédire y_{l+1} (déduction). De la sorte, toute distinction réelle n'existe qu'au plan pratique et calculatoire ; est-ce que l'on extrait et stocke la règle générale ou non ?
Pour les prédictions contrôlées, la différence entre induction et transduction est plus profonde. Nous voulons aboutir à des notions différenciées de prédiction contrôlée dans les deux cadres. Les résultats mathématiques sur l'induction mettent habituellement en jeu deux paramètres, souvent désignés ε (la précision souhaitée de la règle de prédiction) et δ (la probabilité de ne pas obtenir la précision ε), alors que les résultats sur la transduction n'impliquent qu'un seul paramètre, noté ε dans cet article (la probabilité d'erreur que nous acceptons) ; cf. figure 9. On peut se reporter à [51], Section 10.1 pour ce qui concerne la prédiction inductive de ce point de vue.

8 Prédiction conforme inductive

Notre approche de la prédiction est tout à fait transductive, et c'est ce qui donne la possibilité d'obtenir des prédictions contrôlées valides et efficientes. Dans ce paragraphe, nous allons voir cependant qu'il est aussi possible d'être inductif en prédiction conforme.

Regardons attentivement la démarche de prédiction conforme, telle que décrite dans la section 3. Supposons donnés un ensemble d'apprentissage (1) et les objets d'un ensemble test (14), notre objectif étant de prédire le label de chacun des objets de l'ensemble test. Si nous voulons utiliser le prédicteur conforme basé sur les SVM comme décrit en section 3, nous devons obtenir les multiplicateurs de Lagrange pour chaque objet test et pour chaque label *Y* qui lui soit potentiellement attribuable. Ceci implique de résoudre $k|\mathbf{Y}|$ problèmes d'optimisation indépendants. L'approche des plus proches voisins est certainement plus efficiente du point de vue des calculs, mais cependant elle s'avère plus lente que la procédure suivante, suggérée dans [30, 31].

Supposons que nous disposions d'un algorithme inductif qui, à partir d'un ensemble d'apprentissage (1) et d'un nouvel objet *x* produit une prédiction $\hat{y}$ du label *y* de *x*. Fixons alors une mesure $\Delta(y,\hat{y})$ de la différence entre *y* et $\hat{y}$. La procédure est la suivante :

1. Partager l'ensemble d'apprentissage original en deux sous-ensembles, l'ensemble d'apprentissage proprement dit $(x_1,y_1),...,(x_m,y_m)$ et l'ensemble de calibration $(x_{m+1},y_{m+1}),...,(x_l,y_l)$
2. Construire une règle de prédiction à partir de l'ensemble d'apprentissage proprement dit
3. Calculer le score de non-conformité $\alpha_i := \Delta(y_i, F(x_i)), \quad i = m+1,...,l$ pour chaque exemple de l'ensemble de calibration
4. Pour chaque objet test x_i, $i = l+1,...,l+k$, faire comme suit :

   a) pour tout label possible $Y \in \mathbf{Y}$ calculer le score de non-conformité $\alpha_i := \Delta(y_i, F(x_i))$, et la *p*-valeur

$$p_Y := \frac{\#\{j \in \{m+1,...,l,i\} : \alpha_j \geq \alpha_i\}}{l-m+1}$$

   (b) en déduire l'ensemble de prédiction $\Gamma^{\varepsilon}(x_1,y_1,...,x_l,y_l,x_i)$ donné par le membre de droite de l'équation (7).

Ceci est un cas particulier des « prédicteurs conformes inductifs » définis en [51], Section 4.1. Dans le cas de la classification supervisée, bien entendu, on peut réunir les *p*-valeurs comme une simple prédiction avec le degré de confiance (8) et la crédibilité (9).

Les prédicteurs conformes inductifs sont valides, au sens où la probabilité d'erreur $y \notin \Gamma^{\varepsilon}(x_1,y_1,...,x_l,y_l,x_i)$ ($i = l+1, \ldots, l+k, \varepsilon \in (0,1)$) ne dépasse jamais ε (cf. (12)). La version « en ligne » des prédicteurs conformes inductifs, associés à une notion plus forte de validité, est décrite dans [48] et [51] (Section 4.1).

Le principal avantage des prédicteurs conformes inductifs est leur efficience calculatoire : l'ensemble des calculs est réalisé une seule fois, et il reste uniquement, pour chaque objet test et chaque label potentiel, à appliquer la règle de prédiction trouvée à l'étape d'induction, à appliquer Δ pour trouver le score α de non-conformité pour cet objet et ce label, et à trouver la position de α au sein des scores de non-conformité des exemples de calibration. Le principal inconvénient est une possible perte d'efficience de prédiction : pour les prédicteurs conformes, nous utilisons en effet tout l'ensemble d'apprentissage comme ensemble d'apprentissage proprement dit et comme ensemble de calibration.

9 Conclusion

Cet article montre comment de nombreuses techniques d'apprentissage automatique peuvent être enrichies avec des mesures de précision et de fiabilité dont la validité peut être prouvée. Nous avons expliqué brièvement comment ceci peut être réalisé pour les machines à vecteur support, les algorithmes des plus proches voisins, et la procédure de régression ridge, mais son principe est général : virtuellement toute technique pertinente de prédiction construite pour opérer sous

l'hypothèse de randomisation peut être utilisée pour produire des prédictions contrôlées. D'autres exemples sont donnés dans le livre récent ([51]) écrit avec Glenn Shafer, où nous construisons des prédicteurs conformes et des prédicteurs conformes inductifs basés sur la régression au plus proches voisins, la régression logistique, le rééchantillonnage bootstrap, les arbres de décision, le boosting (stimulation), et les réseaux de neurones ; des schémas généraux pour la construction de prédicteurs conformes et de prédicteurs conformes inductifs sont donnés pp. 28-29 et pp. 99-100 respectivement de [51]. Le remplacement de prédictions simples par des prédictions contrôlées permet de contrôler le nombre d'erreurs au moyen d'un choix approprié de niveau de confiance.

Références

[1] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397-422, 2002.

[2] John S. Bell. *Speakable and Unspeakable in Quantum Mechanics*. Cam-bridge University Press, Cambridge, 1987. See p. 27.

[3] Tony Bellotti, Zhiyuan Luo, Alexander Gammerman, Frederick W. van Delft, and Vaskar Saha. Qualified predictions for microarray and proteomics pattern diagnostics with confidence machines. *International Journal of Neural Systems*, 15:247-258, 2005.

[4] Ulisses M. Braga-Neto and Edward R. Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20:374-380,2004.

[5] Bernard Bru. The Bernoulli code. *Electronic Journal for History of Probability and Statistics*, 2(1), June 2006. Available on-line at http://www.jehps.net.

[6] Nicolμo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, England, 2006.

[7] Joshua W. Comley and David L. Dowe. General Bayesian networks and asymmetric languages. In *Proceedings of the Hawaii International Conference on Statistics and Related Fields*, June 2003.

[8] Joshua W. Comley and David L. Dowe. Minimum message length and generalized Bayesian nets with asymmetric languages. In Peter Grünwald, Mark A. Pitt, and In Jae Myung, editors, *Advances in Minimum Description Length: Theory and Applications*, pages 265-294. MIT Press, 2005.

[9] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

[10] A. Philip Dawid. Probability forecasting. In Samuel Kotz, Norman L. Johnson, and Campbell B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210-218. Wiley, New York, 1986.

[11] A. Philip Dawid. Discussion of the papers by Rissanen and by Wallace and Dowe. *Computer Journal*, 42(4):323-326, 2000.

[12] Arthur P. Dempster. An overview of multivariate data analysis. *Journal of Multivariate Analysis*, 1:316-346, 1971.

[13] David L. Dowe and Chris S. Wallace. Kolmogorov complexity, minimum message length and inverse learning. In *Proceedings of the Fourteenth Australian Statistical Conference*, page 144, 1998.

[14] Peter Gacs. Uniform test of algorithmic randomness over a general space. *Theoretical Computer Science*, 341:91-137, 2005.

[15] Alexander Gammerman and A. R. Thatcher. Bayesian diagnostic probabilities without assuming independence of symptoms. *Methods of Information in Medicine*, 30:15-22, 1991.

[16] Murray Gell-Mann. *The Quark and the Jaguar*. W. H. Freeman, 1994. See p. 34.

[17] Hans-Martin Gutmann. A radial basis function method for global optimization. *Journal of Global Optimization*, 19:201-227, 2001.

[18] David J. Hand. Classifier technology and the illusion of progress (with discussion). *Statistical Science*, 21:1-14, 2006.

[19] Donald R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345-383, 2001.

[20] Donald R. Jones, Matthias Schonlau, and William J.Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455-492, 1998.

[21] Kevin B. Korb. Calibration and the evaluation of predictive learners. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 73-77, 1999.

[22] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1962. Third edition: 1996.

[23] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In David S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 396-404. Morgan Kaufmann, San Mateo, CA, 1990.

[24] Ming Li and Paul Vitànyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, second edition, 1997.

[25] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602-619, 1966.

[26] Thomas Melluish, Craig Saunders, Ilia Nouretdinov, and Vladimir Vovk. Comparing the Bayes and typicalness frameworks. In Luc De Raedt and Peter A. Flach, editors, *Proceedings of the Twelfth European Conference on Machine Learning*, volume 2167 of *Lecture Notes in Computer Science*, pages 360-371, Heidelberg, 2001. Springer.

[27] John S. Mill. *A System of Logic*. 1843. See p. 130.

[28] Ilia Nouretdinov, Tom Melluish, and Vladimir Vovk. Ridge Regression Confidence Machine. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 385-392, San Francisco, CA, 2001. Morgan Kaufmann.

[29] Ilia Nouretdinov and Vladimir Vovk. Criterion of calibration for transductive confidence machine with limited feedback. *Theoretical Computer Science*, 364:3-9, 2006. Special issue devoted to the ALT'2003 conference.

[30] Harris Papadopoulos, Konstantinos Proedrou, Vladimir Vovk, and Alexander Gammerman. Inductive Confidence Machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Proceedings of the Thirteenth European Conference on Machine Learning*, volume 2430 of *Lecture Notes in Computer Science*, pages 345-356, Berlin, 2002. Springer.

[31] Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Qualified predictions for large data sets in the case of pattern recognition. In *Proceedings of the International Conference on Machine Learning and Applications*, pages 159-163. CSREA Press, Las Vegas, NV, 2002.

[32] Karl R. Popper. *Logik der Forschung*. Springer, Vienna, 1934. An English translation, *The Logic of Scientific Discovery*, was published by Hutchinson, London, in 1959.

[33] Daniil Ryabko, Vladimir Vovk, and Alexander Gammerman. Online prediction with real teachers. Technical Report CS-TR-03-09, Department of Computer Science, Royal Holloway, University of London, 2003.

[34] Glenn Shafer. The unity and diversity of probability. *Statistical Science*, 5:435-444, 1990.

[35] Ilham A. Shahmuradov, Viktor V. Solovyev, and Alexander Gammerman. Plant promoter prediction with confidence estimation. *Nucleic Acids Research*, 33:1069-1076, 2005.

[36] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[37] Stephen Swift, Allan Tucker, Veronica Vinciotti, Nigel Martin, Christine Orengo, Xiaohui Liu, and Paul Kellam. Consensus clustering and functional interpretation of gene expression data. *Genome Biology*, 5:R94, 2004.

[38] Kei Takeuchi. *Statistical Prediction Theory*. Baihūkan, Tokyo, 1975.

[39] Kei Takeuchi. Non-parametric prediction regions. Hand-out for a lecture at Stanford University (3 pages), 17 July 1979.

[40] Peter J. Tan and David L. Dowe. MML inference of oblique decision trees. In *Proceedings of the Seventeenth Australian Joint Conference on Artificial Intelligence*, volume 3339 of *Lecture Notes in Artificial Intelligence*, pages 1082-1088. Springer, 2004.

[41] Vladimir N. Vapnik. Оценивание зависимостей по эмпирическим данным. Nauka, Moscow, 1979. English translation: Springer, New York, 1982. Second English edition: *Estimation of Dependences Based on Empirical Data: Empirical Inference Science*. Information Science and Statistics. Springer, New York, 2006.

[42] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995. Second edition: 2000.

[43] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[44] Vladimir N. Vapnik and Alexey Y. Chervonenkis. Теория распознавания образов *(Theory of Pattern Recognition)*. Nauka, Moscow, 1974. German translation: *Theorie der Zeichenerkennung*, Akademie, Berlin, 1979.

[45] Veronica Vinciotti, Allan Tucker, Paul Kellam, and Xiaohui Liu. The robust selection of predictive genes via a simple classifier. *Applied Bioinformatics*, 5:1-12, 2006.

[46] Vladimir Vovk. Aggregating strategies. In Mark Fulk and John Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371-383, San Mateo, CA, 1990. Morgan Kaufmann.

[47] Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213-248, 2001.

[48] Vladimir Vovk. On-line Confidence Machines are well-calibrated. In *Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science*, pages 187-196, Los Alamitos, CA, 2002. IEEE Computer Society.

[49] Vladimir Vovk. Predictions as statements and decisions. In Gabor Lugosi and Hans Ulrich Simon, editors, *Proceedings of the Nineteenth Annual Conference on Learning Theory*, volume 4005 of *Lecture Notes in Artificial Intelligence.*, page 4, Berlin, 2006. Springer. Full version: Technical Report arXiv:cs.LG/0606093, arXiv.org e-Print archive, June 2006.

[50] Vladimir Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444-453, San Francisco, CA, 1999. Morgan Kaufmann.

[51] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[52] Chris S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics. Springer, New York, 2005.

[53] Chris S. Wallace and David M. Boulton. An information measure for classification. *Computer Journal*, 11:185-195, 1968.

[54] Chris S. Wallace and David M. Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11-34, 1975.

[55] Chris S. Wallace and David L. Dowe. Minimum message length and Kolmogorov complexity. *Computer Journal*, 42:270-283, 1999.

[56] Juyang Weng. Muddy tasks and the necessity of autonomous mental development. In *Proceedings of the 2005 AAAI Spring Symposium Series, Developmental Robotics Symposium, Stanford University*, 2005.