Nouvelles méthodologies de la modélisation: Théorie de Vapnik et mise en oeuvre par KXEN

Michel Bera

KXEN, 25 Quai Gallieni
92158 Suresnes cedex, France
Michel.Bera@kxen.com,
http://www.kxen.com

Résumé. Nous présentons ici quelques éléments de la théorie de Vladimir Vapnik. Nous montrons comment les concepts de précision et de robustesse permettent de définir un «bon» modèle. Le théorème de Vapnik, que nous présentons, indique les contraintes pour qu'un modèle soit «bon »: la classe de fonctions où l'on recherche ce modèle doit avoir une VC dimension finie et cette classe de fonctions est choisie selon le principe de SRM (minimisation structurelle du risque). Nous décrivons ensuite comment KXEN a implémenté ces éléments théoriques pour réaliser un moteur de modélisation efficace et robuste

1 Introduction

En 1995, le mathématicien Vladimir Vapnik donnait une conférence remarquée au sein des laboratoires Bell Labs, présentant une nouvelle approche scientifique de la modélisation basée sur le concept qui devait le rendre célèbre : la VC Dimension (V comme Vapnik, C comme Chernovenkis, autre mathématicien russe). Nous avons déja exposé, très brièvement, les principes de la théorie de Vapnik dans un article paru dans la revue *Risques* (Bera, 2001).

Le présent exposé est composé de deux parties : une première partie rappelle les principes majeurs de la théorie de Vladimir Vapnik ; la deuxième partie est dévolue à la mise en oeuvre de ces idées par la société KXEN, selon l'esprit du mathématicien Léon Bottou et que reflète l'un de ses propos :«*Il y a un grand nombre de problèmes de modélisation, particulièrement ardus à résoudre. Cependant, il y a un plus grand nombre de problèmes qui ne sont certes pas glorieux, mais dont on sait d'ores et déjà trouver la solution, d'une manière efficace* ».

A ce stade, il convient de définir ce que nous entendons par «modèle »ou «modélisation ». Comme données, on dispose d'un ensemble de L observations d'un phénomène, décrit par des variables $X^1, \ldots, X^n$ (on parle de données structurées). A chacune des observations, on peut adosser la réponse à une question posée, «la question métier »Y. Ainsi, le modèle est une fonction permettant de prédire Y, à partir de la connaissance des variables $X^1, \ldots, X^n$. Le type de fonction est donné de façon explicite (modèle paramétrique) ou implicite (non paramétrique) par le modélisateur. Quant à la fonction exacte, elle est déterminée ou estimée grace

aux L observations constituant l'*échantillon d'apprentissage* (on parle d'estimation supervisée). Plus précisément, si Y est à deux valeurs, comme {0,1} (Par exemple, $0 \equiv$ «*j'autorise un prêt* », $1 \equiv$ «*je le refuse* »), on parle de modèle de classification ; si Y est numérique, comme un chiffre d'affaires, on parle d'un modèle de régression. Ainsi, on peut ramener le modèle à l'équation

$$Y = f(X^1, \ldots, X^n) + \varepsilon$$

Étant données des observations de $X^1, \ldots, X^n$, le modèle est de qualité d'autant meilleure que «les affectations »estimées $\widehat{Y} = \widehat{f}(X^1, \ldots, X^n)$ et les affectations réelles Y proches ou concordantes. La notion de proximité entre estimations $\widehat{Y}$ et valeurs vraies Y sera définie avec précision plus loin. L'évaluation de la qualité du modèle, ainsi définie, est réalisée à partir de l'observation d'un *échantillon test* ; il s'agit d'un échantillon renseigné pour les variables de description $X^1, \ldots, X^n$, pour la variable cible Y et distinct de l'échantillon d'apprentissage.

Posons, pour simplifier, $X = (X^1, \ldots, X^n)$; l'approche de la statistique traditionnelle consiste, et cela est plus ou moins bien formalisé depuis les années 1930, à étudier les données que recèle l'échantillon d'apprentissage et à estimer les paramètres de la loi de $X_{|Y}$ supposée connue (approche paramétrique) ou à déterminer cette loi à partir des données et selon une plus large famille de modèles (méthodes semi paramétriques). La loi de $X_{|Y}$ étant connue, on obtient aisément le modèle de prédiction de Y connaissant X. La suite des opérations est dévolue à la validation du modèle prédictif obtenu, au moyen d'une batterie de tests statistiques dont l'issue dépend des lois retenues. Il s'agit là d'un cycle minutieux qui donne de bons résultats, avec l'aide d'experts statisticiens.

La révolution de Vladimir Vapnik tient en deux résultats fondamentaux (Vapnik, 1995).

- Le premier résultat énonce la possibilité de construire des modèles prédictifs f de qualité (qualité requise en apprentissage et en test), non pas en analysant les données mais en contrôlant la nature (prédictive) de la fonction f. Cette nature est traduite par un nombre entier, la VC dimension, qui caractérise la famille $\mathcal{F}$ dans laquelle est choisie f. A partir de ce résultat, V. Vapnik propose une stratégie, la minimisation structurelle du risque, pour contrôler et optimiser la nature de $\mathcal{F}$, puis déterminer le modèle $f \in \mathcal{F}$ optimal ;
- Le second résultat a trait aux propriétés particulières des modèles linéaires du type

$$Y = f(X^1, \ldots, X^n) = w_1 X^1 + \ldots + w_n X^n + b. \quad (1)$$

  Ce second résultat rend pertinente l'approche nouvelle qui consiste à construire un espace «étendu »de variables $\widetilde{X}^1, \ldots, \widetilde{X}^n$; chacune de ces nouvelles variables est fonction des variables initiales $X^1, \ldots, X^n$, *i.e.*, $\widetilde{X}^j = g_j(X^1, \ldots, X^n)$ et où la variable cible Y s'exprimera linéairement en fonction des nouvelles variables de l'espace étendu, *i.e.*,

$$Y = w_1 \widetilde{X}^1 + \ldots + w_n \widetilde{X}^n + b. \quad (2)$$

La révolution que propose KXEN s'appuie sur ces deux résultats. Pour un grand nombre de problèmes, «*ces petits problèmes moins glorieux que l'on sait déjà résoudre* », elle propose

- de construire pour chaque variable X^j (nominale, ordinale ou continue) un encodage $\widetilde{X}^j$;
- de bâtir le modèle linéaire final $Y = w_1 \widetilde{X}^1 + \ldots + w_n \widetilde{X}^n + b$;

– d'utiliser, à chaque étape, le principe de minimisation structurelle du risque, afin de garantir la qualité durant l'apprentissage (précision) et la qualité pendant les tests (robustesse) ;
– d'effectuer ces opérations de manière automatisée, ce qui change le paradigme de la modélisation, et permet la mise en oeuvre de nouvelles «usines à modèles »dont la productivité devient un élément clef de la conduite d'entreprise.

Pour terminer cette introduction, sur une note quantitative, il faut comprendre que l'état de l'art existant en matière de modélisation dans la banque (l'exemple de l'octroi de credit), l'assurance (l'exemple de la détection des fraudes dans les déclarations de sinistres), les télécommunications (l'exemple de l'analyse de churn ou propension à la défaillance d'un client) et la grande distribution (l'exemple de l'analyse de la rentabilité des points de vente) porte sur des milliers de variables (n) pour décrire les phénomènes, des millions d'observations (L) et la mise en oeuvre de dizaines de modèles nouveaux f, par semaine, soit plusieurs centaines par mois.

Tout porte à croire que ces chiffres vont fortement augmenter dans les années, voire les mois, à venir. C'est ce qu'on appelle le «data mining extrême ».
De telles cadences dans la production de modèles de qualité ne peuvent être atteintes que par une automatisation de cette production.

2 Rappel des principes majeurs de la théorie de Vapnik

2.1 Définition de la VC dimension

Considérons L observations points de $\mathbb{R}^n$; il est possible de segmenter ces points, en points «blancs »et points «noirs »de 2^L manières différentes. Soit maintenant une famille de fonctions $\mathcal{F} = \{f : \mathbb{R}^n \to \{0, 1\}\}$.

On dira que $\mathcal{F}$ disperse complètement notre ensemble de L points, si quel que soit le découpage de ces L points en L_0 points «noirs »et L_1 points «blancs », il existe une fonction f dans $\mathcal{F}$ telle que $f(X) = 1$ pour tout X «noir »et $f(X) = 0$ pour tout X «blanc ».

On appellera dimension de Vapnik-Chernovenkis (ou VC dimension) de la famille $\mathcal{F}$, le nombre entier h tel que
– tout ensemble de h points est complètement dispersable par $\mathcal{F}$;
– il existe un jeu de $h + 1$ points non complètement dispersable par $\mathcal{F}$.

2.2 Quelques exemples de VC dimension

– la VC dimension des droites du plan est 3 ;
– la VC dimension des hyperplans de $\mathbb{R}^n$ est $n + 1$;
– la VC dimension des courbes de fonction $f : \mathbb{R} \to \mathbb{R}$ telle que $f(X) = cos(mX + b)$ est infinie ;
– la VC dimension des fonctions multilinéaires $f(X) =< w, X >_{\mathbb{R}^n} +b$ telles que $\|w\| \leq C$ dépend de C et peut prendre toute valeur entre 1 et $n + 1$; cela est donné dans le second théorème de Vapnik.

On voit que la VC dimension d'une famille de fonctions $\mathcal{F}$ n'est pas égale à la dimension de l'espace de représentation des données X.

Comme le montre le troisième exemple, une famille de fonctions à une seule variable peut avoir une VC dimension infinie ; tandis que dans le quatrième exemple la VC dimension de la famille de fonctions, peut être petite relativement à la dimension des données.

2.3 Précision et robustesse : Les deux notions de «proximité »en modélisation

Nous avons dit que lors d'une modélisation, l'on s'intéresse à deux choses, à savoir la précision et la robustesse. On suppose définie une fonction de coût, de mesure d'erreur, entre la vraie valeur Y et la valeur estimée $\widehat{Y}$ fournie par le modèle f, *i.e.*, $\widehat{Y} = f(X^1, \ldots, X^n)$. Ainsi, on associe souvent à un échantillon S (apprentissage, test ou autre) une mesure globale de l'erreur de modélisation $Err_{mod_{|S}}$ appellée également précision. Parmi les mesures utilisées, on retrouve fréquemment :

- l'erreur quadratique moyenne ou L_2 : $Err_{mod_{|S}} = \frac{1}{Card(S)} \sum_{i \in S} (Y_i - \widehat{Y}_i)^2$;
- l'erreur absolue moyenne ou L_1 : $Err_{mod_{|S}} = \frac{1}{card(S)} \sum_{i \in S} |Y_i - \widehat{Y}_i|$;
- l'erreur absolue maximum ou L_∞ : $Err_{mod_{|S}} = sup_{\ i \in S} |Y_i - \widehat{Y}_i|$.

Ce que l'on souhaite évidemment, pour avoir un «bon modèle », c'est d'une part que «la précision »sur l'ensemble d'apprentissage soit bonne, mais aussi, lorsque l'on porte le modèle sur de nouvelles données, celles d'échantillons test, que l'erreur moyenne constatée soit du même ordre de grandeur que la précision constatée sur l'ensemble d'apprentissage. On dira alors que la modélisation est robuste et bien entendu le modèle est alors exploitable, au sens industriel du terme.

Consistance d'un modèle :
On dira qu'une modélisation est consistante si l'espérance mathématique de l'erreur d'apprentissage (précision) converge, lorsque la taille de l'échantillon d'apprentissage augmente, vers l'erreur de modélisation constatée sur l'ensemble de test.
Comme le montre la figure 1, lorsque la taille de l'échantillon d'apprentissage est très faible, l'erreur d'apprentissage est nulle ; puis, au delà d'une certaine taille L^* de l'échantillon d'apprentissage, une erreur de précision apparaît et converge vers une certaine asymptote lorsque L augmente.

De la même manière, le modèle sera imparfait si l'apprentissage s'effectue sur une taille d'échantillon très petite et l'erreur de cette modélisation, calculée sur l'ensemble test, sera grande. Au fur et à mesure que L augmente, le modèle s'améliore et l'on voit une convergence de l'erreur sur l'ensemble test vers une seconde asymptote. La consistance de la modélisation revient à dire que les deux courbes convergent vers une asymptote commune.

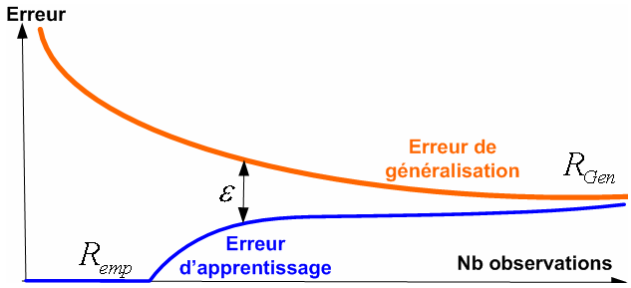

FIG. 1 – *Evolution de l'erreur d'apprentissage et l'erreur de généralisation en fonction du nombre d'observations*

2.4 Le premier grand théorème de Vapnik

Soit une modélisation effectuée sur un échantillon d'apprentissage de taille L, à partir d'une famille de fonctions $\mathcal{F}$. Soit $Q(X,Y)$ une mesure de l'erreur de modélisation. L'espérance mathématique de Q est donnée par

$$E(Q) = \int Q(X,Y)dP(X,Y)$$

avec P la mesure de probabilité du vecteur aléatoire (X,Y). Le théorème de Vapnik s'énonce ainsi :
On a presque sûrement (*i.e.*, ou avec une probabilité supérieure à q)

$$E(Q(X,Y)) \leq \frac{1}{L}\sum_{i=1}^{L} Q(X_i,Y_i) + \varphi(h,L) \tag{3}$$

avec h la VC dimension de la famille de fonctions $\mathcal{F}$ où est choisie f, et $\varphi(h,L)$ un terme qui

- ne dépend pas de la fonction coût Q ;
- ne dépend pas de la distribution de (X,Y) ;
- qui tend vers 0 avec $\frac{h}{L}$ (Vapnik parle de $\frac{h}{L} < \frac{1}{20}$ pour une modélisation de bonne qualité) ;
- ne dépend pas du nombre de variables ;
- dépend du seuil de probabilité q retenu.

Les conséquences de ce théorème, qui s'inscrit dans la lignée des grands théorèmes en statistique (Kolmogorov-Smirnov,. . .), sont considérables. En effet, le théorème permet d'envisager des modélisations à très grand nombre de variables, pour un petit nombre d'observations. Il établit le lien entre la notion de consistance et celle de VC dimension de la famille de fonctions $\mathcal{F}$ de la modélisation et ce, au moyen d'un théorème de vitesse de convergence. Il donne un sens précis et une justification à l'approche dite de la cross-validation (il n'y a pas si longtemps considérée comme simple programmation informatique *ad-hoc*) qui consiste à rapprocher la mesure de la précision d'une modélisation sur une partie de l'ensemble d'apprentissage, de l'erreur mesurée sur une autre partie de ce même ensemble ; ceci pour voir « où on en est »du

rapprochement des deux courbes de consistance. Nous avons vu que la VC dimension permettait de gérer au mieux la consistance d'une modélisation, c'est à dire la capacité du modèle à réagir « avec une qualité égale »sur des données du même univers, qu'il n'a jamais vues ; nous allons maintenant voir quelle stratégie adopter pour construire au mieux cette famille de fonctions : c'est l'approche structurelle du risque que Vapnik nous propose.

2.5 Construire une modélisation - principe de modélisation structurelle du risque

L'inéquation (3) comprend deux termes : le premier terme mesure la précision et le second la robustesse. L'idée de Vapnik est de « travailler », de manière contrôlée, la famille de fonctions modèles, en faisant varier h. Plus précisément, l'approche de modélisation structurelle du risque (d'abréviation anglophone : SRM) revient à construire une suite de familles de fonctions emboîtées $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}_m$. Plus l'indice de la classe de modèles est grand et plus cette classe est riche et donc apte à séparer les points noirs des points blancs. En notant h_i la VC dimension de $\mathcal{F}_i$, on a $h_1 \leq h_2 \leq \ldots \leq h_m$.

Une image culinaire de cet enchaînement des classes familles de fonctions, dans un exemple à une dimension, serait de prendre m paquets de spaghettis et de mettre chaque paquet, à cuire dans une casserole différente ; le premier durant 1 minute, le second 2 minutes,. . ., le dernier m minutes. Les premiers paquets donneront des spaghettis à peine cuits, trop rigides et donc peu capables de se courber pour séparer les points noirs des points blancs sur une droite ; par contre leur rigidité garantit une stabilité de l'erreur de modélisation sur de nouvelles données. Les derniers paquets, eux donneront des spaghettis trop cuits, complètement mollassons, se pliant docilement à toutes les courbures d'une classification à effectuer. Mais, trop cuits, ils ne garderont pas leur forme sur de nouvelles données et l'erreur du modèle en test sera médiocre. Dans le cas « pas assez cuits », la précision est médiocre mais la robustesse excellente ; dans le second c'est l'inverse,*i.e.*, la précision est excellente et la robustesse est médiocre.

Comme le rappelle Aristote dans l'Ethique à Nicomaque :« in medio stat virtus »,*i.e.*, il existe une « cuisson intermédiaire », celle qui réalise le meilleur compromis entre une perte en précision et un gain en robustesse ; cela donnera la familles de fonctions d'où proviendra le meilleur modèle. Ici, le paramètre « temps de cuisson »est celui qui contrôle la VC dimension et donc la robustesse de la modélisation finale. Comment trouve-t-on cette bonne dimension de Vapnik, ce temps de cuisson optimale ? Comme on dit en cuisine, on réservera une partie des données de l'ensemble initial (l'ensemble dit de *validation*) et on bâtira une modélisation sur l'ensemble des données restantes (l'ensemble dit d'*estimation* ou d'apprentissage), pour une VC dimension donnée. On testera, alors, la qualité de cette modélisation en mesurant Q sur les données mises en réserve.

La mesure d'erreur Q varie en fonction de la VC dimension ; la VC dimension optimale ou « temps de cuisson *al dente*»réalisant le minimum. Cette VC dimension optimale correspond à une « richesse »de la famille de fonctions optimale parmi laquelle on choisit la fonction de modélisation finale.

3 L'approche KXEN

3.1 Une transgression modérée de la théorie de Vapnik

L'une des principales conséquences de l'approche de Vapnik est d'admettre que « tous les chemins mènent à Rome », c'est-à-dire qu'il existe différentes manières de modéliser un problème (arbres de décision, réseaux de neurones, régression logistique, ou fonctions de KXEN) et conduisant à des modèles de qualité peut-être analogues en précision et robustesse. Ceci, à condition de savoir en contrôler la complexité et de disposer d'un algorithme d'apprentissage performant. Nous allons présenter, ici, le cas le plus simple, celui de la classification supervisée. Comme le choix des fonctions de modélisation est ouvert, KXEN a donc décidé de travailler avec ses propres familles de fonctions, avec une quadruple exigence :

- que les fonctions qui définissent le modèle soient simples à comprendre (équations, graphiques) et que leur lecture donne une information sur les données X du problème et leur relation avec la cible Y ; on retrouve une notion d'inférence ;
- que le modèle soit rapide à établir (temps de calcul) ;
- que le modèle soit de qualité (précision-robustesse) ; cela est assuré par l'approche `SRM` ;
- que le modèle soit rapide en exécution : le modèle est conçu sur des données d'apprentissage (quelques centaines de milliers de lignes), mais il peut être mis en oeuvre, par la suite, sur des données massives (des millions, voire des centaines de millions de lignes) ;
- et surtout, que l'ensemble de ces opérations puisse être automatisé avec une qualité satisfaisante, par rapport aux meilleurs modélisations « traditionnelles ».

Pour tenir compte de ces différentes exigences, `KXEN` a mis en place une transgression douce des équations de la théorie de Vapnik :

- une extension de la notion de la mesure du coût ou erreur du modèle, avec le passage de l'approche ligne à ligne de Vapnik à celle d'une fonction définie sur les blocs de lignes (le *KI* de `KXEN`) ;
- une modélisation en deux étapes, qui permet de mélanger données numériques et données nominales ou ordinales, mais aussi de règler en deux étapes distinctes le processus `SRM`, l'un pour le préscoring et l'autre pour la modélisation finale.

3.2 Une extension de la fonction « coût à la ligne »à une fonction « coût au bloc »

Nous avons déjà vu un exemple de fonction coût (cf. sous section 2.3) dont le calcul fait intervenir d'un coup l'ensemble des L lignes disponibles. Il existe d'autres fonctions coût basées sur l'usage des statistiques d'ordre utilisées par les statisticiens du Marketing direct ; il en est ainsi du *lift*. On construit, pour l'échantillon de L lignes $\{(X_i, Y_i)\}_{i=1,\ldots,L}$, le vecteur des L réponses (ou scores) $S = (S_i)$ données par le modèle.
L'idée est d'ordonner les lignes, selon l'ordre décroissant des S_i, obtenant l'échantillon ordonné $\{(X_i^*, S_i^*)\}_{i=1,\ldots,L}$. A priori, les premières lignes (ou plus grandes valeurs S_i^* correspondent aux valeurs $Y = 1$ et les dernières aux valeurs $Y = 0$.
La courbe de *lift*, comme celle donnée par la figure 2, ci-dessous, donne en ordonnée la proportion de vrais « $Y = 1$ »(ou vrais positifs) et en abscisse la population rangée par score décroissant.

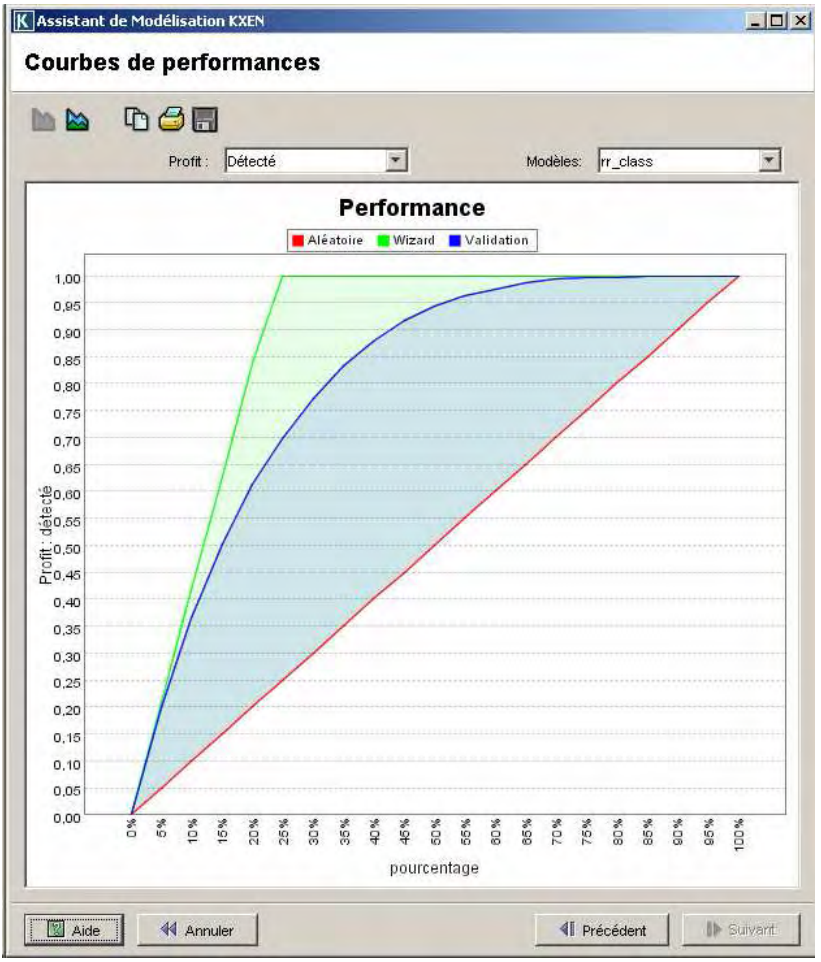

FIG. 2 – *Courbes de performances produites par KXEN*

La courbe de *lift* se trouve ainsi coincée entre deux courbes « triviales » : celle du simple modèle au hasard (pour 50% des meilleurs scores, on trouve 50% des vrais positifs) et celle du modèle « sorcier »qui connaît la réponse.
La fonction de coût global *KI*, retenue par KXEN, est définie comme le rapport de deux surfaces : la surface délimitée par la courbe de *lift* et la courbe « modèle au hasard »divisée par le maximum possible, *i.e.*, la surface délimitée par la courbe du « sorcier »et la courbe du « modèle au hasard »d'autre part. Ce rapport *KI* est par construction compris entre 0 et 1.
On montre, par ailleurs, que *KI* s'exprime simplement en fonction de la courbe *ROC* (Receiver

Operational Curve) et de l'aire *AUC* sous cette courbe ; plus précisément $KI = 2AUC - 1$.
KI s'exprime, également, en fonction du coefficient de concentration de Gini C_{Gini} : $KI = (1 - \widehat{p}_1)C_{Gini}$ avec $\widehat{p}_1$ la proportion de positifs dans l'échantillon.
Dans l'ensemble de ses composants, KXEN utilisera des méthodes de cross-validation lors des approches SRM, basées sur cette fonction coût.
Reste à démontrer une petite conjecture dont tous (y compris Vapnik) admettent la validité :

$$E(KI) \leq KI_{emp} + \Phi(h, q, L).$$

3.3 Modélisation d'un classifieur en deux étapes : préprocessing et modèle final

La modélisation de KXEN pour un classifieur s'effectue en deux étapes :
– la première étape consiste à encoder chacune des variables X^j, à partir de la cible Y, en une variable à valeurs numériques ;
– la deuxième étape bâtit un modèle multilinéaire dont la robustesse sera construite par SRM basée sur une interprétation à la Vapnik.

3.3.1 Encodage des variables X^j

L'encodage des variables X^j a un but multiple :
– transformer chaque variable X^j, qu'elle soit nominale ou numérique, en une variable $\widetilde{X}^j$, à valeurs numériques et d'échelle de valeurs comparable à celle des autres variables. Ainsi, on projette l'espace de départ sur un espace intermédiaire, sorte d'hypercube de côté voisin de 1 ;
– apporter à l'utilisateur une information précieuse, à savoir une visualisation de la réponse Y du classifieur par rapport aux valeurs possibles de X^j.

La robustesse de l'encodage est obtenue grâce à une compression qui reprend la méthode SRM, *i.e.*, si un échantillon est coupé en deux parties d'événements d'apprentissage, l'encodage de X^j sur chacune des parties est stable, ce que l'on constate clairement sur un exemple comme le test universitaire *Adult* (cf. figure 3, ci dessous.)

3.3.2 Modélisation finale : une Ridge regression régularisante

Dans la seconde phase de l'encodage, on s'intéresse à la détermination de l'importance relative des nouvelles variables $\widetilde{X}^j$.
Posons $\widetilde{X} = (\widetilde{X}^1, \ldots, \widetilde{X}^n)$ et considérons le modèle multilinéaire $Y =< w, \widetilde{X} > +b$ avec $w \in \mathbb{R}^n$ un vecteur pondérant les variables et $b \in \mathbb{R}$ un paramètre de centrage.
Nous savons que les contraintes sur w du type $\|w\| \leq C$ agissent sur la VC dimension. Plus C sera petit, plus robuste sera la modélisation.
L'approche de KXEN consiste à partir de la ridge regression classique, en minimisant la mesure d'écart *modèle-réalité*, pénalisée

$$\Psi = \Sigma_{k=1,\ldots,L}(Y_k - < w, \widetilde{X}_k > -b)^2 + \lambda\|w\|^2$$

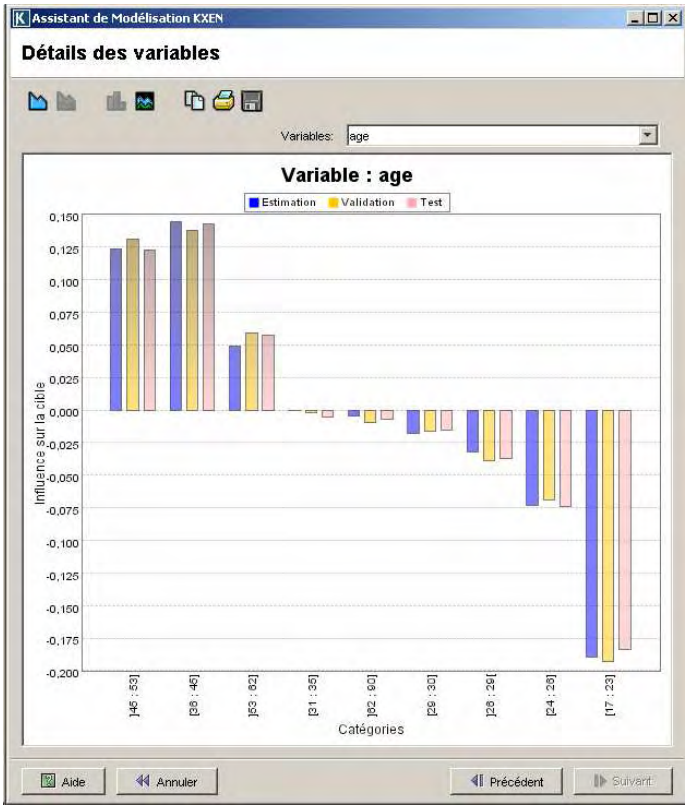

FIG. 3 – *Visualisation de l'effet sur la cible des différentes classes d'une variable.*

où $(Y_k, \widetilde{X}_k)$ est la ligne k de l'échantillon encodé ; λ est le paramètre qui va nous permettre de faire une approche SRM à la Vapnik. En effet, augmenter λ revient à pénaliser les fortes valeurs de $\|w\|^2$; pour prendre une image, un peu comme le prix d'une amende de radar qui, augmentée, oblige les automobilistes à mieux respecter les limitations de vitesse.
On reconnaît également dans l'écriture de Ψ, une approche de régularisation décrite dans (Hastie et *al*. 2001). Pour trouver le meilleur λ et donc la meilleure VC dimension, au sens d'une négociation entre *fit* et robustesse, on mesurera le *fit* par le KI du lift du classifieur obtenu à chaque itération et en procédant par cross-validation.
Le modèle final est alors recalculé sur l'ensemble de l'échantillon d'apprentissage avec cette valeur de λ optimale.
L'intérêt d'utiliser une *ridge régression* pour faire le calcul de modélisation est également lié à la forme quadratique de l'expression de Ψ (quadratique en w et en b). Le calcul du minimum se fait par simple inversion de matrice et l'on est donc assuré d'une grande rapidité de calcul.

La simplicité de l'équation d'un modèle multilinéaire permet par ailleurs, là encore, de faire une restitution descriptive de la modélisation, en représentant les poids w_i par un histogramme en bâton, ce qui permet de visualiser (comme dans la figure 4, ci dessous) le niveau de significativité des variables.

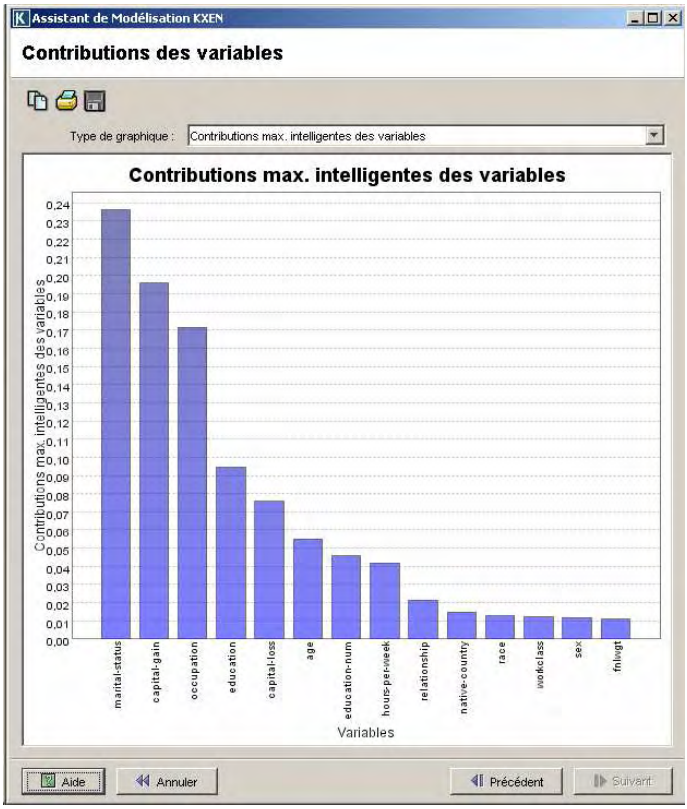

FIG. 4 – *Visualisation de la contribution des variables.*

4 Conclusion

La théorie de Vladimir Vapnik, qui vient d'être nommé à la `National Academy of Engineering` de la `National Academy of Sciences` des Etats Unis «*For insights into the fundamental complexities of learning and for inventing practical and widely applied machine learning algorithms*», ouvre la porte à une nouvelle approche de la modélisation.
A la fois descriptive par le choix de fonctions simples pour l'encodage et la modélisation finale et prédictive par l'utilisation des principes `SRM` à chacune de ces deux étapes (pour garantir

le meilleur équilibre entre *fit* et robustesse), l'approche KXEN permet surtout pour un large spectre de problèmes de modélisation prédictive l'automatisation du processus de modélisation à qualité égale, ce qui réduit considérablement le temps et donc le coût de la mise au point d'un modèle.
En 2001, où un premier article de KXEN sur le sujet est paru (Bera, 2001), on parlait couramment d'une production de modèles (modèles à plusieurs centaines de variables et quelques millions d'individus) sur un cycle de mois, avec un objectif de trois à quatre modèles l'an.
Aujourd'hui, en 2006, non seulement les nombres de colonnes et lignes des données à modéliser ont augmenté de manière considérable, plusieurs milliers et plusieurs dizaines, voire centaines de millions de lignes, mais la demande en modélisation s'est faite de plus en plus pressante : on est passé à 50 modèles au mois, pour de nombreuses applications.
Dans les applications liées à l'Internet, dans les applications liées au domaine de la grande distribution, dans les applications liées aux nouveaux Media qui succèdent à l'Internet, les volumétries de modélisation vont encore sauter un seuil, grâce à la puissance des machines et à l'automatisation intelligente, scalable et contrôlée.
Il n'est pas interdit d'imaginer, d'ici un ou deux ans, un monde de 50 modèles de l'heure, que l'on appelle dans le jargon les modèles « Kleenex ». Bien entendu, on parle de modèles dont chacun, pris au hasard dans la chaîne de production, aura les mêmes critères de qualité, *fit* et robustesse, qu'un modèle traditionnel « à la main »fait en statistique, comme une régression logistique ou un arbre de décision. Là, le coût de la modélisation devient critique.
Il est clair que dans ce nouveau contexte, le rôle de la statistique dans ces domaines se trouve bouleversé. Il ne s'agit plus de réaliser des voitures à carrosseries Charron de grand luxe, mais de réduire sur une chaîne de production moderne le temps de cycle de fabrication d'une voiture grand public de haute qualité à la Audi 4, tout en contrôlant la qualité, par des méthodes qui, pour ce qui est de la modélisation, sont elles aussi en pleine révolution, voire totalement à imaginer et construire.[1]

[1]La société anonyme des Automobiles Charron, Girardot et Voigt réalisait des voitures imposantes avec une production totale de moins de 300 véhicules par an. Les clients de ces automobiles coûteuses sont cités dans une brochure de vente datant de décembre 1904, tels le roi du Portugal, des princes, des marquises, des barons, des comtes et des figures en vue de la haute société et du monde du commerce comme Waldorf Astor, Louis Blériot, André Michelin, William Vanderbilt, James Gordon Bennett et Orly Roedere de la maison des champagnes Roederer à Reims.

Références

Bera, M., 2001. Les nouveaux énoncés de la modélisation prédictive à très grand nombre de variables. *Risques,* 45*, pp.1-7*.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning : Data mining, inference and prediction. Springer Series in statistics. *Springer Verlag*.

Vapnik, V.N., 1995. The nature of Statistical Learning Theory. *Springer Verlag*.

Summary

We present, in this paper, some elements of Statistical Learning Theory introduced by Vladimir Vapnik. We show how the two concepts of fit and robustness allow to define a « good » model. The Vapnik's theorem we present gives constraints for a model to be « good »: that the class of functions from which the model is selected is of finite VC dimension and that this class of functions is chosen using the SRM (Structural Risk Minimization) principle. We then describe how KXEN has implemented these theoretical results resulting in a modeling engine both efficient and robust.