Critique Constructive des Techniques Utilisées pour l'Estimation et la Validation des Modèles de Prédiction en Data Mining

Ioan Doré Landau, Vlad Landau, Tudor-Bogdan Airimitoaie, Bogdan Robu

Critique Constructive des Techniques Utilisées pour l'Estimation et la Validation des Modèles de Prédiction en Data Mining

I.D. Landau *[1], V. Landau[2], T.B. Airimitoaie[3], et B. Robu[1]

[1]GIPSA-Lab, Université Grenoble Alpes, UMR 5216 CNRS
[2]Free Lance Consultant
[3]Univ. Bordeaux, IMS-lab, UMR CNRS 5218, F-33405 Talence

31 mars 2017

Résumé

Cette contribution a pour premier objectif d'analyser les techniques actuelles utilisées en data mining pour l'estimation de la complexité et des paramètres des modèles de prédiction et la validation de ces modèles. Cette analyse pointe les aspects discutables de la méthodologie actuelle. Son deuxième objectif est de proposer des améliorations méthodologiques et algorithmiques pour l'estimation et la validation des modèles de prédiction en data mining. Des exemples utilisant des données simulées et réelles illustreront le potentiel des améliorations proposées.

Mots-clef :Fouille de données, Modèle de prédiction, Algorithmes d'estimation, Validation de modèles, Régression linéaire

1 Introduction

La méthodologie actuelle d'estimation des modèles de prédiction en data mining est organisée sur les principes énoncés dans le standard CRISP-DM (Cross industry standard process for data mining) ([FT13], [C.00]). Après une phase préliminaire de compréhension du problème le schéma correspondant au CRISP-DM est illustré dans la Fig. 1. On distingue plusieurs étapes :

— Acquisition des données
— Préparation (pré-traitement) des données
— Choix des attributs et de leur nombre (Estimation de la complexité du modèle)
— Estimation des paramètres du modèle de prédiction
— Validation du modèle

La phase validation du modèle de prédiction est fondamentale car elle va indiquer si le modèle peut être déployé (s'il valide) ou s'il faut revoir les choix qui ont été faits dans les différentes étapes de la procédure.

Si le schéma général donné dans la Fig. 1 est sans aucun doute le bon chemin à suivre pour l'estimation des modèles de prédiction, on peut questionner les techniques actuelles utilisées à chaque étape. Pour être concret, on se placera dans le contexte de l'utilisation des modèles de prédiction dont les paramètres sont estimés par des techniques de *régression linéaire* (algorithme d'estimation paramétrique des moindres carrés).

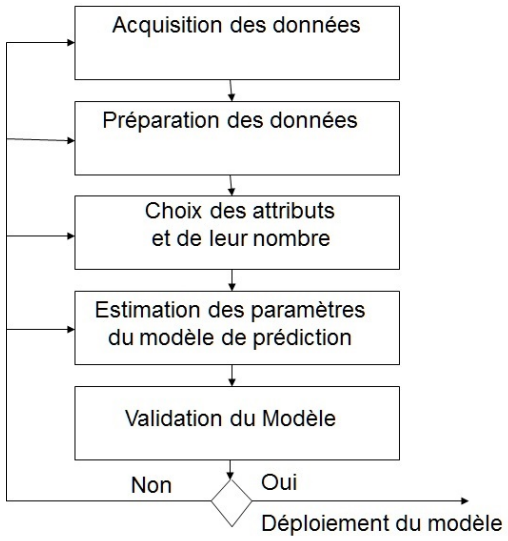

FIGURE 1 – Illustration du standard CRISP-DM.

*ioan-dore.landau@gipsa-lab.grenoble-inp.fr

Indiscutablement, le point essentiel dans la procédure d'estimation des modèles de prédiction est la phase de *validation du modèle*. Dans l'absolu, peu importe les étapes et les algorithmes qui ont été utilisés pour obtenir un modèle de prédiction à partir des données, ce qui compte, c'est que le modèle obtenu passe les *tests de validation*.

Le test de validation pose néanmoins un problème fondamental : quel est le critère d'évaluation de la validation ? Le consensus actuel est basé sur l'idée de la séparation de la base de données en une *base d'apprentissage* et une *base de test* [LP15]. De nombreux travaux ont été effectués pour trouver la meilleur façon de partitionner la base de données. En particulier la technique des n-partitions croisées (cross-validation) [FT13], [A.W01] semble donner de bons résultats. On estime les paramètres du modèle sur la base d'apprentissage et on valide le modèle sur la base de test. Le critère utilisé actuellement consiste d'une part à comparer sur la base de test en terme de somme des carrés des erreurs de prédiction résiduelles, les performances des modèles estimés de complexité croissante et d'autre part, de déterminer la complexité à partir de la quelle le phénomène de *sur-apprentissage* sur la base de test apparaît (les performances en terme d'erreur de prédiction se dégradent sur la base de test en continuant d'augmenter la complexité alors que les performances s'améliorent en général sur la base d'apprentissage avec l'accroissement de la complexité du modèle).

Une première remarque s'impose : actuellement le choix de la complexité du modèle ne se fait que dans la phase de validation. On peut se poser la question s'il n'est pas possible d'obtenir une estimation de la complexité du modèle directement à partir des données d'apprentissage avant de faire le test de validation.

Cette démarche de validation qui semble très logique à première vue fait implicitement des hypothèses qui ne sont pas nécessairement vraies en pratique. La première hypothèse est que la régression linéaire en minimisant la somme des carrés des erreurs de prédiction produit un modèle optimal non-biaisé. Il ne faut pas oublier que dans la pratique on a toujours des "bruits" (dans certains cas ce bruit correspond à des erreurs de modélisation qui ne peuvent pas être pris en compte par le type de modèle utilisé). Or cette propriété d'optimalité de la "régression linéaire" n'est vraie que si ce bruit de mesure est un "bruit blanc". Dans le cas contraire bien qu'on a l'impression qu'on minimise l'erreur de prédiction (on peut montrer que ce n'est pas le cas), le modèle sera baisé (il y aura des erreurs sur les paramètres par rapport à un modèle optimal).

Il convient donc, avant de passer à la validation sur la base de test, se poser la question : As-t-on estimé le modèle optimal sur la base d'apprentissage ? Cet aspect qui est fondamental n'est pas considéré à l'heure actuelle. Plusieurs problèmes sous-jacents se posent :

1. En supposant que la complexité du modèle est connue, quel critère de validation peut être utilisé pour tester l'optimalité du modèle estimé sur la base de d'apprentissage ?
2. Quels sont les algorithmes permettant d'obtenir une estimation optimale quand l'hypothèse *bruit blanc* n'est pas vérifiée ?
3. Y a t-il une procédure d'estimation de la complexité du modèle de prédiction à partir des données de la base d'apprentissage qui prend en compte le *principe de parcimonie* ?

Nous proposons donc de compléter la méthodologie actuelle par :

1. Des techniques d'estimation de complexité à partir des données de la base d'apprentissage.
2. Des algorithmes d'estimation dans le cas où la régression linéaire fourni un modèle de prédiction avec des paramètres biaisés
3. Des critères de validation objectifs permettant de tester l'optimalité du modèle estimé sur la base d'apprentissage

Bien sur le test final de validation se fera sur la base de test. A noter aussi que les critères de validation proposés peuvent être utilisé aussi sur la base de test en complément des procédures actuelles.

Cette contribution est organisée comme suit : Dans la Section 2 nous discuterons le problème de l'estimation optimale d'un modèle de prédiction sur la base d'apprentissage et des tests de validation associés. Des algorithmes d'estimation permettant d'améliorer les performances de la régression linéaire seront présentés. Dans la section 3 nous examinerons le problème de l'estimation de la complexité du modèle de prédiction à partir de données de la base d'apprentissage. La Section 4 présentera une comparaison entre la procédure classique d'estimation et validation des modèles de prédiction et la procédure améliorée proposée dans cette communication en utilisant des données engendrées par un modèle simulé et des données réelles dans le contexte de l'estimation du modèle de prédiction d'un dispositif électro-mécanique utilisé en contrôle actif de vibrations.

2 Estimation optimale d'un modèle de prédiction sur la base d'apprentissage

On supposera temporairement pour des raisons de clarté de l'exposé que la complexité du modèle de prédiction est connue. Supposons qu'on souhaite utiliser l'algorithme de régression linéaire (algorithme des moindres carrés) pour l'estimation des paramètres d'un modèle de prédiction dont on connaît la structure (complexité, nombre et types de variables prédictives). Il s'agit de l'estimation d'un modèle de la forme :

$$y(t+1) = \theta^T \phi(t) \tag{1}$$

où $y(t)$ est la variable prédite, θ est le vecteur des paramètres à estimer et ϕ et le vecteur des mesures (observations) réunissant toutes les variables prédictives (attributs). Si on utilise l'algorithme de moindres carrés, on fait implicitement l'hypothèse que le bruit de mesure (ou erreur de modélisation dans certains cas) est un bruit blanc. En présence de bruit, l'équation 1 est remplacée par :

$$y(t+1) = \theta^T \phi(t) + w(t+1) \tag{2}$$

où w(t+1) est un bruit. Dans ce contexte, l'estimation des moindres carrés correspondante est donnée par la formule : [LZ05]

$$\hat{\theta}(N) = \theta + \left[\frac{1}{N}\sum_{t=1}^{N}\phi(t-1)\phi(t-)^T\right]^{-1} \left[\frac{1}{N}\sum_{t=1}^{N}\phi(t-1)w(t)\right] \tag{3}$$

où N est le nombre de données.

On voit apparaître un terme d'erreur (biais) qui dépend du bruit. Ce terme est nul si et seulement si $\phi(t)$ et $w(t+1)$ sont non corrélés ou si $w(t)$ est un bruit blanc (noté $e(t)$). Comme hypothèse "bruit blanc" n'est pratiquement jamais vérifiée sur les données réelles, il est courant de considérer que $w(t+1)$ est une *moyenne mobile*.

$$w(t) = e(t) + c_1 e(t-1) + c_2 e(t-2) + ... \tag{4}$$

Revenons temporairement au cas idéal $w(t) = e(t)$. Dans ce contexte tenant compte des équations 2 et 3 il n'y a pas de biais ($\hat{\theta} = \theta$) et l'erreur de prédiction résiduelle est un bruit blanc. On peut montrer que dans ce contexte, minimiser l'erreur de prédiction ou obtenir une erreur de prédiction blanche est strictement équivalent ([LZ05], [AW84]). Cette propriété nous permet de définir un test de validation. En effet, si l'erreur de prédiction résiduelle est un bruit blanc quand on utilise la régression linéaire alors nous sommes dans les hypothèses d'une estimation non-biaisée des paramètres et le modèle obtenu est une estimation optimale. A contrario, si l'erreur de prédiction résiduelle n'est pas un bruit blanc, on n'a pas une estimation optimale et il faut considérer une autre modélisation du bruit plus réaliste et un algorithme associé qui minimise l'erreur de prédiction résiduelle et qui donc produira une erreur de prédiction blanche.

Pour être concret considérons l'exemple simple :

$$y(t+1) = -a_1 y(t) + b_1 u(t) + c_1 e(t) + e(t+1) \tag{5}$$
$$= \theta^T \phi(t) + c_1 e(t) + e(t+1) \tag{6}$$

On montre ([LZ05], [AW84]) que le prédicteur optimal permettant de minimiser l'erreur de prédiction et rendre l'erreur de prédiction blanche est :

$$\hat{y}(t+1) = -a_1 y(t) + b_1 u(t) + c_1 e(t) + e(t+1)$$
$$= \hat{\theta}^T \phi(t) + c_1 e(t) \tag{7}$$

Dans ce cas l'erreur de prédiction est un bruit blanc

$$\epsilon(t+1) = y(t+1) - \hat{y}(t+1) = e(t+1) \tag{8}$$

et le prédicteur optimal peut être ré-écrit sous la forme :

$$\hat{y}(t+1) = -a_1 y(t) + b_1 u(t) + c_1 \epsilon(t) = \theta^T \phi(t) + c_1 \epsilon(t) \tag{9}$$

Cette équation montre que dans le cas d'un bruit type "moyenne mobile" nous sommes obligés d'estimer non seulement les paramètres du modèle de prédiction caractérisé par le vecteur de paramètres θ mais aussi les paramètres du modèle de bruit.

On définit alors un modèle de prédiction étendu :

$$y(t+1) = -a_1 y(t) + b_1 u(t) + c_1 e(t) + e(t+1) \tag{10}$$
$$= \theta_e^T \psi_e(t) + e(t+1) \tag{11}$$

où

$$\theta_e^T = [a_1, b_1, c_1] \tag{12}$$
$$\psi_e(t)^T = [-y(t), u(t), e(t)] \tag{13}$$

qu'on estime par :

$$\hat{y}(t+1) = \hat{\theta}_e^T \phi_e(t) \tag{14}$$

$$\hat{\theta}_e^T = [a_1, b_1, c_1] \tag{15}$$
$$\phi_e(t)^T = [-y(t), u(t), \epsilon(t)] \tag{16}$$

Nous sommes ainsi de nouveau dans un contexte correct de l'estimation par une régression linéaire et l'algorithme correspondant porte le nom de *moindres carrés étendus* [LZ05]. On obtient une estimation optimale qui d'une part minimise réellement la somme des carrés des erreurs de prédiction et d'autre part conduit à une erreur de prédiction blanche.

3 Validation du modèle sur la base d'apprentissage

Il s'agit en fait d'un test d'hypothèses *a posteriori*. Si l'hypothèse "bruit blanc" est correcte alors en utilisant la régression linéaire classique, l'erreur de prédiction résiduelle doit être proche d'un "bruit blanc". Si ce n'est pas le cas, ceci veut dire qu'un autre modèle de bruit doit être considéré (par exemple la "moyenne mobile") et un algorithme d'estimation paramétrique adéquat qui permettra d'obtenir une erreur de prédiction résiduelle proche d'un "bruit blanc" devra être utilisé. Il faudra bien entendu en vertu du même principe de vérification d'hypothèses de faire un "test de blancheur" sur l'erreur de prédiction résiduelle.

Pour résumer : si l'erreur résiduelle est un bruit blanc nous avons une estimation paramétrique optimale (donc qui minimise effectivement le carré des erreurs de prédiction sur la base d'apprentissage) et l'estimation des paramètres est non biaisée. Le test de validation de l'optimalité du modèle de prédiction estimé sur la base d'apprentissage sera donc un test de blancheur de l'erreur de prédiction résiduelle.

Le principe du test de validation est le suivant :

- si les structures du modèle de prédiction et du modèle de bruit sont correctes (en d'autres mots ils sont représentatives de la réalité) ;
- si un algorithme approprié pour la structure modèle de prédiction + bruit a été utilisé ;

alors l'erreur de prédiction $\epsilon(t)$ tend asymptotiquement vers un bruit blanc caractérisé par :

$$\lim_{t\to\infty} E\{\epsilon(t)\epsilon(t-i)\} = 0; \quad i = \pm 1, \pm 2, \pm 3 \dots$$

La méthode de validation met en œuvre ce principe en testant la blancheur de l'erreur résiduelle [1].

1. Des routines correspondant à cette méthode de validation en Matlab and Scilab peuvent être téléchargées à partir des sites : http://www.landau-adaptivecontrol.org et http://landau-bookic.lag.ensieg.inpg.fr.

3.1 Test de blancheur

Soit $\{\epsilon(t)\}$ la séquence centrée des erreurs de prédiction résiduelles (centrée : valeur mesurée - valeur moyenne). On calcule les estimations des auto-corrélations $R(i)$ et des auto-corrélations normalisées $RN(i)$:

$$R(0) = \frac{1}{N}\sum_{t=1}^{N} \epsilon^2(t) \ , \ RN(0) = \frac{R(0)}{R(0)} = 1, \tag{17}$$

$$R(i) = \frac{1}{N}\sum_{t=1}^{N} \epsilon(t)\epsilon(t-i) \ , \tag{18}$$

$$RN(i) = \frac{R(i)}{R(0)}, \quad i = 1, 2, 3, \dots i_{max}, \dots \tag{19}$$

où N est le nombre de points d'estimation (longueur de la séquence de données).

Si la séquence des erreurs résiduelles est parfaitement blanche (situation théorique) et le nombre des échantillons est très grand ($N \to \infty$) alors $RN(0) = 1$, $RN(i) = 0$, $i \geq 1$ ce qui implique l'indépendance entre $\epsilon(t), \epsilon(t-1) \dots$, c'est à dire que la séquence des erreurs de prédiction résiduelles $\{\epsilon(t)\}$ est un bruit blanc gaussien. Dans les situations réelles , cela n'est jamais le cas, c'est à dire que les $RN(i)$ pour $i \geq 1$ ne sont pas tout à fait nuls, car d'une part , $\epsilon(t)$ contient des erreurs résiduelles de structure (erreur sur la complexité du modèle, effets non linéaires, bruits non gaussiens) et d'autre part le nombre d'échantillons (N) est fini.

On considère alors comme critère pratique de validation (testé sur de très nombreuses applications)

$$\mid RN(i) \mid \leq \frac{2.17}{\sqrt{N}} \ ; \ i \geq 1 \tag{20}$$

où N est le nombre de échantillons de la séquence ;

Ce test a été défini en tenant compte que pour une séquence *bruit blanc* gaussien $RN(i), i \neq 0$ tend asymptotiquement vers une distribution gaussienne avec une valeur moyenne nulle et un écart type : $\sigma = \frac{1}{\sqrt{N}}$

L'intervalle de confiance considéré dans l'éq. 20 correspond à un niveau de signification de 3% pour le test d'hypothèse d'une distribution gaussienne. On utilise aussi en pratique le critère ([LZ05]) :

$$\mid RN(i) \mid \leq 0.15 \ ; \ i \geq 1. \tag{21}$$

Il s'agit néanmoins d'une validation intermédiaire avant de faire une validation sur la base de test.

3.2 Test de dé-corrélation

L'éq. 3 permet de définir aussi un autre test de validation sans faire référence au bruit blanc. En effet il suffit que les séquences $\phi(t)$ et $w(t)$ soit décorrélées pour ne pas avoir de biais sur les paramètres estimés. On montre que dans l'hypothèse où le bruit de mesure est indépendant des variables prédictives ceci est équivalent à ce que l'erreur de prédiction résiduelle $\epsilon(t)$ et la variable cible $y(t)$ soient décorrélées ([LZ05]). Ceci correspond d'une part au propriétés d'une prédiction optimale et d'autre part ceci s'interprète comme le fait que l'erreur de prédiction résiduelle ne contient aucune information qui dépend des variables prédictives. L'intérêt de ce critère c'est qu'il ne fait pas des hypothèses sur la structure du bruit et par conséquent peut être utilisé tant pour la validation sur la base d'apprentissage que sur la base de test (on ne peut pas supposer en général que la réalisation du bruit est la même sur la base d'apprentissage et la base de test).

On calcule :

$$\begin{aligned} R(i) &= \frac{1}{N}\sum_{t=1}^{N}\epsilon(t)\hat{y}(t-i)\ ;\ i=0,1,2,\ldots,n_A \quad (22)\\ RN(i) &= \frac{R(i)}{\left[\left(\frac{1}{N}\sum_{t=1}^{N}\hat{y}^2(t)\right)\left(\frac{1}{N}\sum_{t=1}^{N}\epsilon^2(t)\right)\right]^{1/2}} \\ i &= 0,1,2,\ldots,n_{max},\ldots \quad (23)\end{aligned}$$

Si $\epsilon(t)$ et $\hat{y}(t-i), i \geq 1$ sont parfaitement dé-corrélées (situation théorique)

$$RN(i) = 0\ ;\ i = 1,2,\cdots,n_A\cdots$$

On considère comme critère pratique de validation le critère 20. On utilise aussi en pratique le critère 21.

4 Estimation de la compléxité du modèle de prédiction à partir des données

Il est possible d'estimer la complexité d'un modèle de prédiction à partir des données disponibles. Il s'agit en effet de trouver les variables prédictives qui permettent d'obtenir un modèle de prédiction significatif. Pour introduire le problème, on considère que le système qui nous intéresse peut être décrit par :

$$y(t) = -a_1 y(t-1) + b_1 u(t-1) \quad (24)$$

Dans cette équation $y(t)$ est la variable cible (sortie du système) et $y(t-1)$ et $u(t-1)$ constituent les variables prédictives qui formeront le vecteur des observations et a_1 et b_1 sont les paramètres du modèle (inconnus). On suppose que les données ne sont pas bruitées. L'ordre du modèle est $n = max(n_A, n_B+d)$ ou n_A est le nombre de variables cible antérieures utilisées comme variables prédictives et n_B est le nombre de variables prédictives externes (entrées) éventuellement décalées de d pas. Le nombre des paramètres à estimer est :

$$n_p = n_A + n_B \quad (25)$$

dans ce cas précis $n = n_A = n_B = 1$ et $n_p = 2$

Question : Est-il possible de tester à partir des données si cette hypothèse sur la structure du modèle est correcte ? Pour effectuer ce test , construisons la matrice suivante :

$$\begin{bmatrix} y(t) & \vdots & y(t-1) & u(t-1) \\ y(t-1) & \vdots & y(t-2) & u(t-2) \\ y(t-2) & \vdots & y(t-3) & u(t-3) \end{bmatrix} = \begin{bmatrix} Y(t) & \vdots & R(1) \end{bmatrix} \quad (26)$$

Si l'ordre du modèle donné dans Eq. (24) est correct, alors le vecteur $Y(t)$ va être une combinaison linéaire des colonnes de $R(1)$ ($Y(t) = R(1)\theta$ avec $\theta^T = [-a_1, b_1]$) et le rang de la matrice va être 2 (au lieu de 3). Si le modèle réel est d'ordre 2 ou supérieur la matrice dans (26) va être de rang plein. Dans le cas général on teste le rang de la matrice $[Y(t), R(\hat{n})]$ où :

$$\begin{aligned} R(\hat{n}) &= [Y(t-1), U(t-1), Y(t-2), U(t-2)\cdots \\ &\quad Y(t-\hat{n}), U(t-\hat{n})], \quad (27)\\ Y^T(t) &= [y(t), y(t-1)\cdots], \quad (28)\\ U^T(t) &= [u(t), u(t-1)\cdots]. \quad (29)\end{aligned}$$

A cause du bruit, cette procédure ne peut pas être directement utilisée dans les situations réelles. Une première approche pratique part de l'observation que le problème de test de rang peut être interprété comme la recherche d'un vecteur $\hat{\theta}$ qui minimise le critère suivant pour une valeur de l'ordre estimé $\hat{n}$.

$$V_{LS}(\hat{n}, N) = \min_{\hat{\theta}} \frac{1}{N}\|Y(t) - R(\hat{n})\hat{\theta}\|^2 \quad (30)$$

où N est le nombres de mesures. Ce critère est équivalent au critère des moindres carrés [SS89]. Si les conditions pour une estimation non biaisée avec les moindres carrés sont satisfaites, (30) est une façon efficace pour estimer l'ordre du modèle car $V_{LS}(\hat{n}) - V_{LS}(\hat{n}+1) \to 0$ quand $\hat{n} \geq n$. Par ailleurs, le principe de parcimonie conduit à l'idée de rajouter à (30) un terme de pénalisation de la complexité du modèle.

Donc la formulation du critère pour estimer la complexité du modèle devient :

$$J_{LS}(\hat{n}, N) = V_{LS}(\hat{n}, N) + S(\hat{n}, N) \tag{31}$$

où typiquement

$$S(\hat{n}, N) = 2\hat{n}X(N) \tag{32}$$

V_{LS} représente le critère non pénalisé. $X(N)$ dans (32) est une fonction qui décroît avec N. Par example, dans le critère nommé $BIC_{LS}(\hat{n}, N)$, $X(N) = \frac{\log N}{N}$ (d'autre choix sont possibles [Lju99], [SS89], [DL96]) et l'ordre $\hat{n}$ correspond à la valeur qui minimise J_{LS} donné par (31). Néanmoins la procédure doit être modifiée si les conditions d'estimation non biaisée ne sont pas remplies. Une solution qui donne de très bons résultats en pratique ([DL94] et [DL96]), consiste à remplacer la matrice $R(\hat{n})$ par une matrice de *variables instrumentales* $Z(\hat{n})$ dont les éléments ne sont pas corrélés avec le bruit de mesure mais sont corrélés avec les variables prédictives non bruitées. Une telle *matrice instrumentale* $Z(\hat{n})$ peut être obtenue en remplaçant dans la matrice $R(\hat{n})$, les colonnes $Y(t-1),\ Y(t-2),\ Y(t-3)$ par des versions retardées de v $U(t-L-i)$, avec $L > n$:

$$Z(\hat{n}) = [U(t-L-1), U(t-1), U(t-L-2), U(t-2)\cdots] \tag{33}$$

qui conduit au critère d'estimation d'ordre

$$J_{IV}(\hat{n}, N) = \min_{\hat{\theta}} \frac{1}{N}\|Y(t) - Z(\hat{n})\hat{\theta}\|^2 + \frac{2\hat{n}logN}{N} \tag{34}$$

et l'ordre estimé est :

$$\hat{n} = \min_{\hat{n}} J_{IV}(\hat{n}). \tag{35}$$

Une courbe typique illustrant l'évolution du critère (34) en fonction de $\hat{n}$ est présentée dans la Fig.2 .

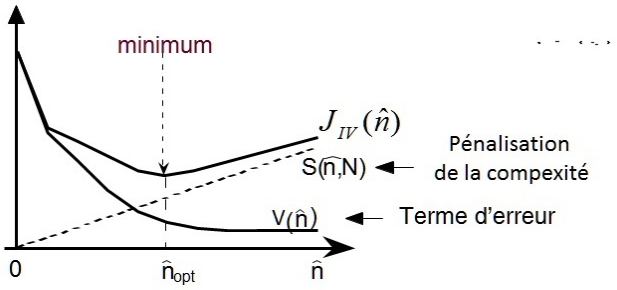

FIGURE 2 – Évolution du critère pour l'estimation de la complexité du modèle de prédiction.

Une fois l'ordre $\hat{n}$ estimé, on estime par une procédure similaire n_A, n_B et d. [2]

2. La fonction estrorderiv.(mat,sci) qui met en œuvre cette procédure est disponible sur : http://www.landau-adaptivecontrol.org

5 Résultats

L'objectif de cette section est de comparer les résultats d'estimation d'un modèle de prédiction en utilisant la procédure classique (utilisation de la régression linéaire et choix de la complexité par le test de sur apprentissage sur la base de test) et en utilisant la procédure proposée dans cet article. Un exemple simulé et un exemple réel (modèle d'un système de contrôle actif de vibrations) vont illustrer la comparaison des deux procédures.

5.1 Exemple simulé

On considère que le système pour lequel on veut trouver le modèle de prédiction à partir de la connaissance des mesures des variables prédictives et des variables cibles est représenté par :

$$\begin{aligned} y(t+1) = &-a_1 y(t) - a_2 y(t-1) + b_1 u(t-1) \\ &+ b_2 u(t-2) + c_1 e(t) + c_2 e(t-1) + e(t+1) \end{aligned} \tag{36}$$

On a

$$y(t+1) = \theta^T \phi(t) + c_1 e(t) + c_2 e(t_1) + e(t+1) \tag{37}$$

où ϕ est le vecteur des variables prédictives (attributs) et $y(t+1)$ est la variable cible. avec

$$\theta^T = [a_1, a_2, b_1, b_2] \tag{38}$$

$$\phi^T(t) = [-y(t), -y(-1), u(t-1), u(t-2)] \tag{39}$$

Les paramètres du modèle simulé sont

$$\theta^T = [-1.5\ 0.7\ 1\ 0.5]. \tag{40}$$

Il s'agit d'un modèle dont la sortie est perturbée par un bruit non-blanc (moyenne mobile) ce qui conduit à des estimations biaisées des paramètres en utilisant la régression linéaire. L'excitation du système $u(t)$ est une SBPA (séquence binaire pseudo-aléatoire) engendrée avec un registre de décalage avec $N = 9$. Une base d'apprentissage de 1024 échantillons et une base de test de 1024 échantillons sont disponibles. Les réalisations de la SBPA et du bruit de mesure sont différentes pour les deux bases de données.

L'estimation classique se fait avec un ordre $n = n_A = n_B + d$ croissant. Le meilleur modèle obtenu résulte de la courbe des sommes des carrés des erreurs de prédiction résiduelles sur la base de test donnée dans la Fig. 3. On choisit $n = n_A = n_B + d = 12$, c'est à

dire que le modèle estimé aura 24 paramètres au lieu des 4 paramètres du modèle simulé. Les résultats des tests de validation sont montrés dans les Figures 4 (test de blancheur sur les données d'apprentissage), 5 (test de dé-corrélation sur les données d'apprentissage), et 6 (test de dé-corrélation sur les données de test). Mais même avec cet ordre élevé, le test de blancheur sur la base d'apprentissage n'est pas tout à fait satisfaisant (Fig. 4).

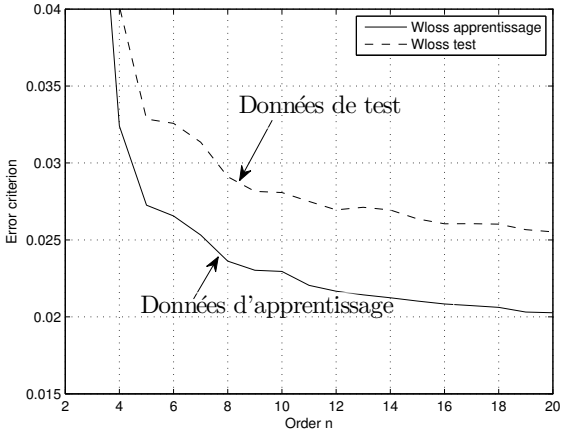

FIGURE 3 – Évolution de la somme des carré des erreurs de prédiction résiduelles.

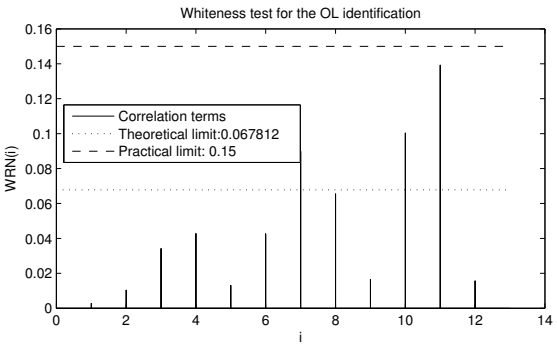

FIGURE 4 – Test de blancheur obtenu en utilisant les données d'apprentissage (approche classique).

En utilisant la méthodologie d'estimation de complexité décrite dans la Section 4 qui permet d'estimer $n = max(n_A, n_B + d)$ mais aussi n_A, n_B, d, les ordres obtenus sont : $n = 3, n_A = 2, n_B = 2, d = 1$ et le nombre de paramètres à estimer est 4. La Fig.7 illustre l'évolution du critère 34 pour l'estimation de n. On voit bien qu'il y a un minimum pour $\hat{n} = 3$.

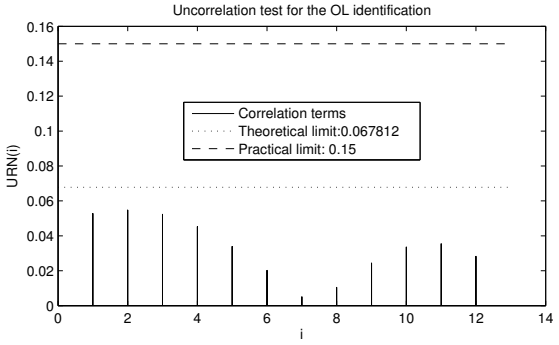

FIGURE 5 – Test de dé-corrélation obtenu en utilisant les données d'apprentissage (approche classique).

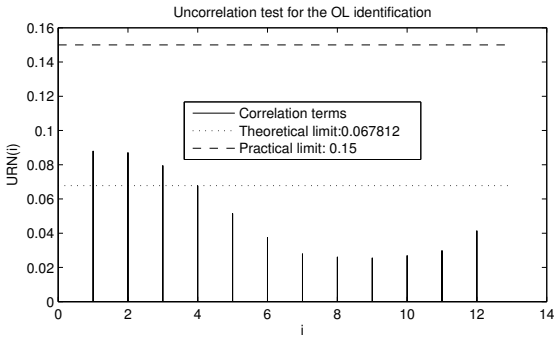

FIGURE 6 – Test de dé-corrélation obtenu en utilisant les données de test (approche classique).

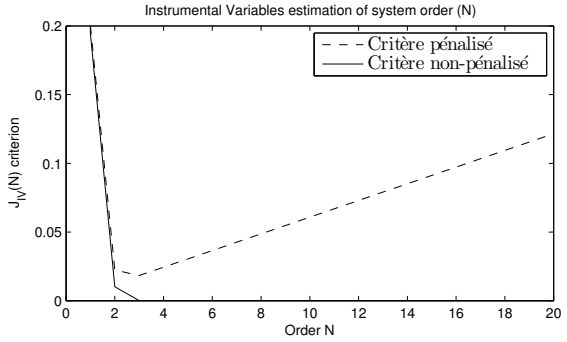

FIGURE 7 – Estimation de l'ordre du modèle.

En utilisant pour l'estimation des paramètres l'algorithme des *moindres carrés étendus* on obtient un modèle de prédiction qui conduit presque à la satisfaction du test de blancheur sur l'erreur de prédiction résiduelle tel que le montre la Figure 8 (avec les données d'apprentissage) (on déduit que le modèle du bruit qui affecte les mesures est une *moyenne mobile*). Les résultats des tests de dé-corrélation sur la base d'ap-

prentissage et de test sont montrés dans les Figures 9 et 10.

La somme des carrés des erreurs de prédiction sur la base d'apprentissage est de 0.0217 pour le modèle estimé classiquement et de 0.0219 pour le modèle estimé avec la nouvelle procédure. Ces résultats sont confirmés sur la base de test. On obtient 0.0269 pour l'estimation classique et 0.0263 pour le modèle estimé avec la nouvelle procédure.

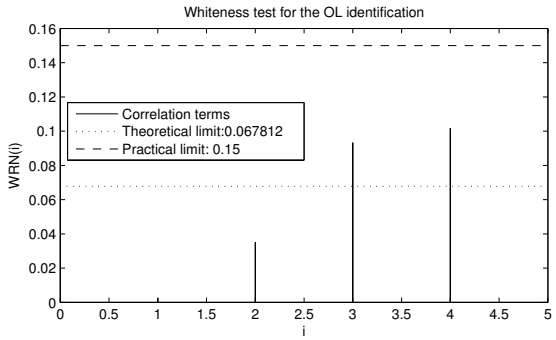

FIGURE 8 – Test de blancheur obtenu en utilisant les données d'apprentissage (approche proposée).

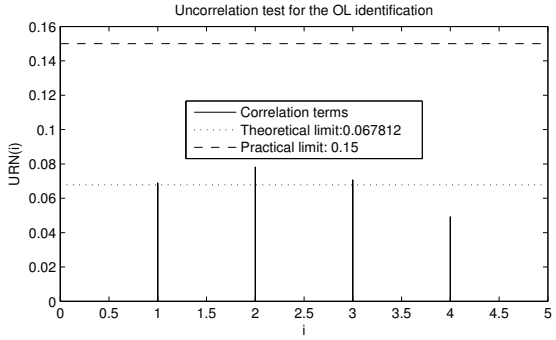

FIGURE 9 – Test de dé-corrélation obtenu en utilisant les données d'apprentissage (approche proposée).

Pour le modèle dont l'ordre a été estimé avec 34, l'identification par l'algorithme des *moindres carrés* ne permet pas d'obtenir un modèle de prédiction valide (voir dans la Fig. 11 la validation par test de blancheur sur les données d'apprentissage). Les paramètres identifiés par la méthode des moindres carrés sont

$$\theta_{MCR}^T = [-1.4863\ 0.6872\ 1.0027\ 0.5129] \qquad (41)$$

et il sont assez loin des valeurs des paramètres du modèle simulé (voir Eq. 40) Les paramètres obtenus avec la méthode méthode des *moindres carrés étendus*

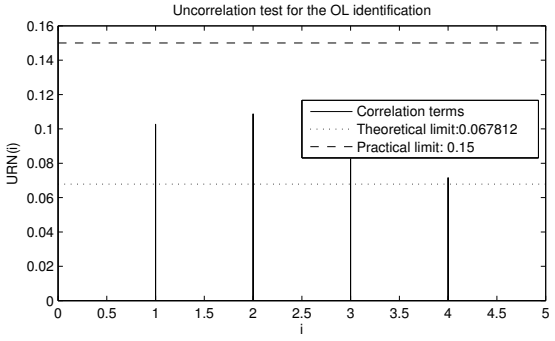

FIGURE 10 – Test de dé-corrélation obtenu en utilisant les données de test (approche proposée).

sont

$$\theta_{MCE}^T = [-1.4966\ 0.6976\ 1.0002\ 0.5072] \qquad (42)$$

et ils sont très proches des valeurs des paramètres du modèle simulé.

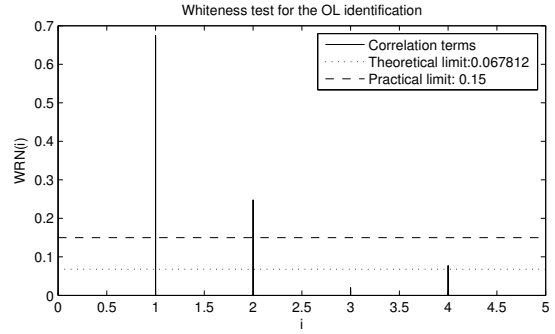

FIGURE 11 – Test de blancheur obtenu en utilisant les données d'apprentissage et les moindres carrés ($n_A = 2, n_B = 2, d = 1$).

5.2 Exemple temps réel

Ce deuxième exemple concerne l'estimation d'un modèle de prédiction dans un système de contrôle actif de vibrations.

La photo du banc de test pour le contrôle actif des vibrations est donnée dans la Fig. 12 et le schéma bloc du système est donné dans la Fig. 13. Les vibrations produites dans ce cas précis par un pot vibrant se propagent jusqu'au niveau du châssis. Pour contrecarrer l'effet de ces perturbations vibratoires, on utilise un moteur inertiel (même principe que le haut parleur) qui va produire des forces contraires aux vibrations et donc annuler leur effet. Le système de commande est un

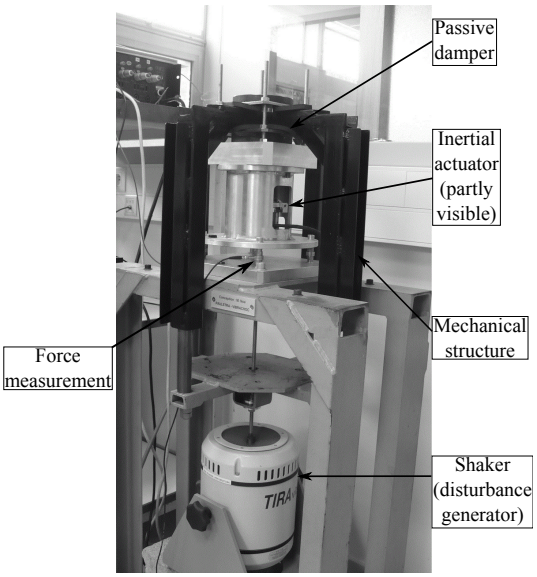

FIGURE 12 – Le banc d'essai (photo).

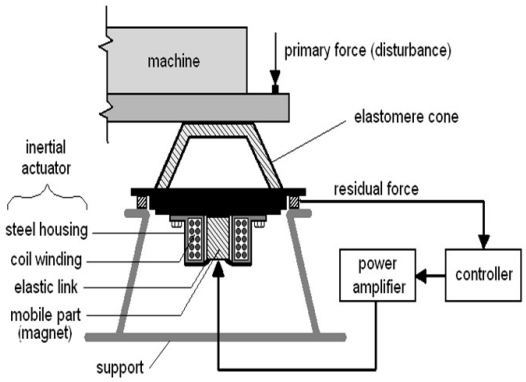

FIGURE 13 – Le banc d'essai (schéma).

système à contre réaction, qui à partir de la mesure de la force résiduelle va calculer une commande qui va être appliquée à l'amplificateur qui alimente le moteur inertiel. Pour déterminer l'algorithme de commande (mis en œuvre par le contrôleur) on a besoin d'un modèle de prédiction de la voie de compensation (ou voie secondaire). Ce modèle doit permettre de prédire la variable cible (sortie) $y(t)$ (force résiduelle) à partir de la commande $u(t)$. La variable cible à l'instant t, $y(t)$, dépend non seulement de $u(t-1)$ mais aussi de ses valeurs antérieures sur un certain horizon et des valeurs antérieures de la variable cible sur un certain horizon. Le système peut être décrit par un modèle de la forme (2) où

$$\theta^T = [a_1, a_2, ...a_{n_A}, b_1, b_2, ...b_{n_B}] \quad (43)$$

$$\phi^T(t) = [-y(t), -y(t-1)... - y(t-n_A), u(t-d), u(t-d-1), ...u(t-d-n_B)] \quad (44)$$

avec des ordres n,n_A,n_B et d à déterminer.

Le signal d'excitation utilisé est une SBPA. Les données ont été séparées en une base d'apprentissage (10000 données) et une base de test (2047 données). Pour l'approche classique l'estimation de modèles avec un ordre croissant a été faite à l'aide de l'algorithme des moindres carrés. Les résultats sont résumés dans la Fig. 14. On constate que la courbe de la somme des carrés des erreurs ne s'infléchit pas sur la base de test. Il est donc difficile de choisir un ordre car il n'y a pas de phénomène de sur-apprentissage même pour des valeurs très élevées de n. La Fig. 15 montre l'évolution du critère 34 pour l'estimation de l'ordre à partir des données. On voit un minimum qui se situe autour de $n = 13$. A partir de ce choix on obtient d'une façon similaire : $n_A = 10$, $n_B = 13$ et $d = 0$. Des tests de blancheur sur des modèles estimés avec $n = 13$, $n_A = 10$, $n_B = 13$ et $d = 0$ et l'algorithme des *moindres carrés étendus* (où ses variantes) permettent de valider ce modèle. Voir Fig. 16 pour un test de blancheur sur la base d'apprentissage et la Fig. 17 pour un test de blancheur sur la base de test. Le modèle est validé sur les deux bases. Des résultats similaires s'obtiennent avec le test de dé-corrélation.

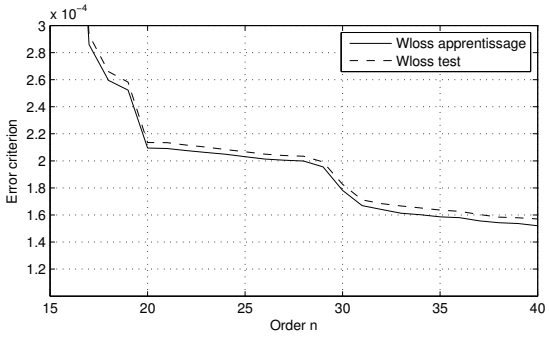

FIGURE 14 – Évolution de la somme des carré des erreurs de prédiction résiduelles.

6 Conclusion

Cette contribution pose le problème de l'estimation optimale d'un modèle de prédiction sur la base d'apprentissage avant de passer à la validation du modèle de prédiction sur la base de test.

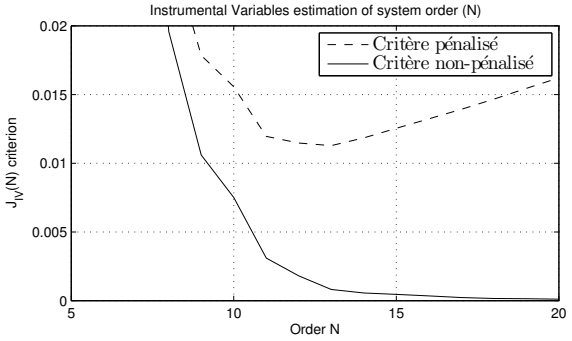

FIGURE 15 – Estimation de l'ordre du modèle.

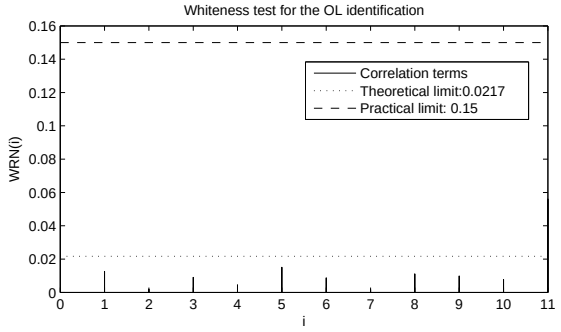

FIGURE 16 – Test de blancheur obtenu en utilisant les données d'apprentissage (approche proposée).

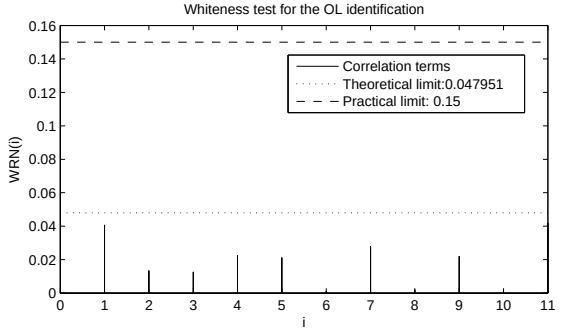

FIGURE 17 – Test de blancheur obtenu en utilisant les données de test (approche proposée).

Le potentiel des techniques proposées pour l'estimation de la complexité et des paramètres des modèles de prédiction ainsi que pour la validation sur la base d'apprentissage a été illustré par des exemples simulés et réels. Sans remettre en cause les principes fondamentaux de l'estimation et de la validation des modèles de prédiction, ces techniques constituent une alternative aux techniques utilisées actuellement et permettent d'obtenir en général des modèles plus simples et plus performants.

Références

[AW84] Karl Johan Astrom and B. Wittenmark. *Computer Controlled Systems, Theory and Design*. Prentice-Hall, Englewood Cliffs, N.J., 1984.

[A.W01] Moore A.W. Cross validation for detecting and preventing overfitting. *Carnegie Mellon University,http ://www.cs.cmu.edu/ cga/ai-course/overfit.pdf*, 2001.

[C.00] Shearer C. The crisp-dm model : The new blue print for data mining. *Journal of Data Warehousing*, 5 :13–22, 2000.

[DL94] Hoai Nghia Duong and I.D. Landau. On statistical properties of a test for model structure selection using the extended instrumental variable approach. *Automatic Control, IEEE Transactions on*, 39(1) :211–215, Jan 1994.

[DL96] Hoai Nghia Duong and Ioan Doré Landau. An IV Based Criterion for Model Order Selection. *Automatica*, 32(6) :909–914, 1996.

[FT13] Provost F. and Fawcett T. *Data Science for Business*. O'Reilly, 2013.

[Lju99] L. Ljung. *System Identification - Theory for the User*. Prentice Hall, Englewood Cliffs, second edition, 1999.

[LP15] Morel M. Raffaelli J.L. Lemberger P., Batty M. *Big Data et Machine Learning*. Dunod,Paris, 2015.

[LZ05] Ioan Dore Landau and G. Zito. *Digital Control Systems - Design, Identification and Implementation*. Springer, London, 2005.

[SS89] Torsten Soderstrom and Petre Stoica. *System Identification*. Prentice Hall, 1989.