Modèles de prédiction des séries temporelles avec un grand nombre de variables

Youssef Hmamouche, Alain Casali, Lotfli Lakhal

May 26, 2017

Les travaux de la thèse portent sur l'étude et le développement des modèles de prédiction pour des séries temporelles caractérisées par un large nombre de variables (STLV). Ces travaux se font dans le contexte du projet *e-Business Optimization with Big Data* (eBOB), qui vise à révolutionner la gestion des achats en entreprises en créant de nouvelles technologies d'analyse des données d'achat mêlant les données structurées et non structurées permettant la modification fonctionnelle dynamique de l'outil selon le contexte résultant des analyses. De nos jours, des quantités énormes de séries temporelles STs sont générées par l'industrie et les systèmes de recherche, pour diverses applications, comme la biologie, la médecine, les finances, l'industrie et bien d'autres. Les systèmes modernes d'analyse de données sont censés traiter et stocker des millions de STs de grande dimension, générant des téraoctets de données. En conséquence, le nombre de STs générées augmente très rapidement et relève de la catégorie Big Data. Parmi les nombreuses applications, la prédiction des STs prend une place particulière car elle est cruciale pour la prise de décision. Sans surprise, c'est souvent une partie importante des systèmes de *Business Intelligence*. L'histoire des modèles de prédiction commence dans les années 1930, avec l'apparition de l'application des premiers modèles uni-variés sur des données. L'idée de ces modèles est simple, l'historique d'une variable est utilisé pour faire des prévisions, en se basant sur des modélisations qui permettent d'adapter les données un modèle régressif, par exemple les modèles *Auto-Regressive* (AR) et *Auto-Regressive Integrated Moving Average* (ARIMA) [2]. L'approche univariée présente des inconvénients évidents, car elle ignore les informations potentiellement exploitables d'autres STs. Les modèles de prédiction multivariés supposent que la valeur d'une variable dépend de valeurs précédentes d'elle-même et des autres variables, par exemple le modèle Vector Auto-Regressive (VAR), Vector Error Correcion (VEC) [4]. Ces modèles sont largement utilisés, seuls ou combinés avec d'autres techniques comme la modélisation dynamique avec les réseaux neuronaux artificiels, par exemple le modèle Vector Auto-Regressive Neural Network (VARNN) [7].

Dans les STs à grande dimension, on peut souligner deux principales problématiques, (*ii*) l'utilisation de toutes les séries n'améliore pas forcément les performances des prédictions, parfois même impossible à cause des problème de

résolution des modèles comme VAR et VEC. (i) les relations de causalité entre les séries ne sont ni symétriques ni monotones. Ceci nous mènent automatiquement à la formulation suivante : Considérons $Y = \{y_1, y_2, \ldots, y_n\}$ une série temporelle multivariés de dimension n et une variable cible y à prédire. L'idée est de sélectionner un sous-ensemble de Y qui permet d'obtenir les meilleures prédictions de y. D'un point de vue combinatoire, il y a $\sum_{i=1}^{k} \binom{n}{i}$ partitions possible de taille $\leq k$, et de manière générale, il y a 2^n partitions possible.

De nouvelles approches de prédiction des STLV ont été proposés, basées généralement sur des systèmes à deux étapes [1, 6, 5], dont la première étape est la réduction des variables prédictives via des méthodes comme *Principal Component Analysis* et *Factor Analysis*. Ensuite, les variables générées sont utilisées pour prédire une série cible via un modèle multivarié. La plupart de ces méthodes sont basées sur la notion de corrélation, or, la causalité est plus pertinente dans le contexte de prédiction vis à vis de l'historique des données, puisqu'elle caractérise naturellement l'effet prédictif (non symétrique) entre deux variables. De ce point de vue, nous avons proposé une méthode d'extraction de variables en considérant le graphe de causalités, avec l'idée de l'analyse des dépendances entre les variables [3]. Les travaux en cours porte sur l'introduction d'un système générale pour la prédiction des STLV. Ce système porte sur une approche hybride basée sur une combinaison dynamique des prévisions générées par plusieurs modèles de prédiction.

References

[1] A. Abraham, B. Nath, and P. K. Mahanti. Hybrid Intelligent Systems for Stock Market Analysis. In *Computational Science - ICCS 2001*, pages 337–345. Springer, Berlin, Heidelberg, May 2001.

[2] G. Box. Box and Jenkins: Time Series Analysis, Forecasting and Control. In *A Very British Affair*, Palgrave Advanced Texts in Econometrics, pages 161–215. Palgrave Macmillan UK, 2013.

[3] Y. Hmamouche, A. Casali, and L. Lakhal. Causality based feature selection approach for multivariate time series forecasting. Feb. 2017.

[4] S. Johansen. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2):231–254, June 1988.

[5] I. Koprinska, M. Rana, and V. G. Agelidis. Correlation and instance based feature selection for electricity load forecasting. *Knowledge-Based Systems*, 82:29–40, July 2015.

[6] J. H. Stock and M. W. Watson. Generalized Shrinkage Methods for Forecasting Using Many Predictors. *Journal of Business & Economic Statistics*, 30(4):481–493, Oct. 2012.

[7] A. Trapletti, F. Leisch, and K. Hornik. Stationary and Integrated Autoregressive Neural Network Processes. *Neural Computation*, 12(10):2427–2450, Oct. 2000.