TECHNOLOGIE **MACHINE LEARNING**

Analyse prédictive de séries temporelles

Prédiction étendue à l'aide de l'apprentissage automatique

La recherche dans le domaine de l'exploitation des données est un sujet qui gagne de l'importance. En effet, un nombre de plus en plus considérable de données sont à disposition, et ce, notamment à cause de l'initiative des données ouvertes (open data) et de l'avènement de l'Internet des objets. La manière de traiter ces données dans un but de prédiction est esquissée dans cet article.

M. Bibimoune, S. Rigori, L. Ji, E. Rappos, S. Robert

L'apprentissage automatique (machine learning) est un domaine qui connaît actuellement un développement très important. Même si tous les problèmes ne peuvent pas être résolus avec ces méthodes, il n'en reste pas moins qu'un grand nombre de domaines ont largement évolué ces dernières années grâce à leurs apports. Il suffit de penser aux voitures autonomes, aux systèmes de recommandation en matière d'achats (par exemple sur Amazon ou Netflix), à la détection de spams, aux recherches sur le Web, au placement personnalisé de publicité, à la reconnaissance vocale ou encore à la meilleure compréhension du génome humain pour s'en convaincre.

Dans un récent rapport [1], Gartner a identifié le domaine de l'apprentissage automatique comme étant l'un des plus stratégiques au niveau technologique. Il y a, de fait, tout un marché de type « algorithmique » qui est en train de se développer. Actuellement, nous assistons à une explosion des sources de données à disposition, entre autres à cause de l'émergence de l'Internet des objets. Si auparavant les données étaient collectées de manière sporadique pour certains besoins spécifiques, à présent nous sommes en train de les regrouper pour alimenter des systèmes de plus en plus autonomes qui commencent à être capables de fournir des informations très complètes. Les réseaux de neurones « profonds » (Deep Neural Networks, DNN) en constituent un exemple. Ces derniers deviennent de plus en plus autonomes et peuvent considérer de nombreuses données complexes, ce qui leur donne un aspect « intelligent ». Les réseaux de neurones « profonds » sont en outre capables de s'adapter à leur environnement, dans les moindres détails, comme à un niveau d'abstraction plus élevé.

Cet article explique, premièrement, comment les techniques d'apprentissage automatique peuvent s'appliquer à la prédiction, sans négliger le soin qui doit être apporté au traitement des données. Un certain nombre d'algorithmes et de tests utilisés pour le domaine de la prédiction sont ensuite présentés.

Techniques d'apprentissage automatique

Il y a plusieurs techniques qui sont regroupées sous l'appellation « apprentissage automatique ». Or, il n'est pas toujours évident d'identifier la valeur et l'information au sein des données à disposition. Deux groupes d'algorithmes existent donc : l'apprentissage supervisé et non supervisé.

Avec les algorithmes supervisés, il s'agit de trouver la fonction de modélisation de la cible à partir des entrées (classes binaires simples ou multiples, valeurs de régression). Il convient dans tous les cas de définir exactement la cible, qui doit être la réponse à une question donnée. Une valeur liée au futur est souvent choisie pour « prédire » un événement particulier, mais il peut aussi s'agir d'un événement déjà réalisé dont la nature cherche à être identifiée, comme la détection de fraudes par exemple. L'apprentissage supervisé est le style de modélisation le plus courant car les risques et les opportunités sont facilement quantifiés. Le modèle doit donc être précis.

D'un autre côté, l'apprentissage non supervisé traite les données sans avoir

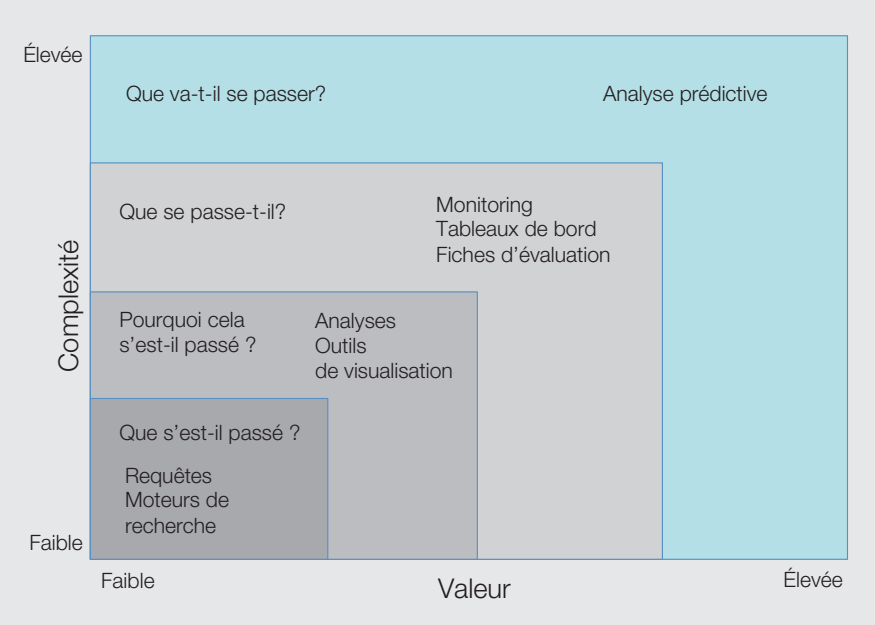


Figure 1 La majorité des entreprises les plus rentables vont gérer leurs données en faisant usage de l'analyse prédictive en temps réel [1].

TECHNOLOGIE **MACHINE LEARNING**

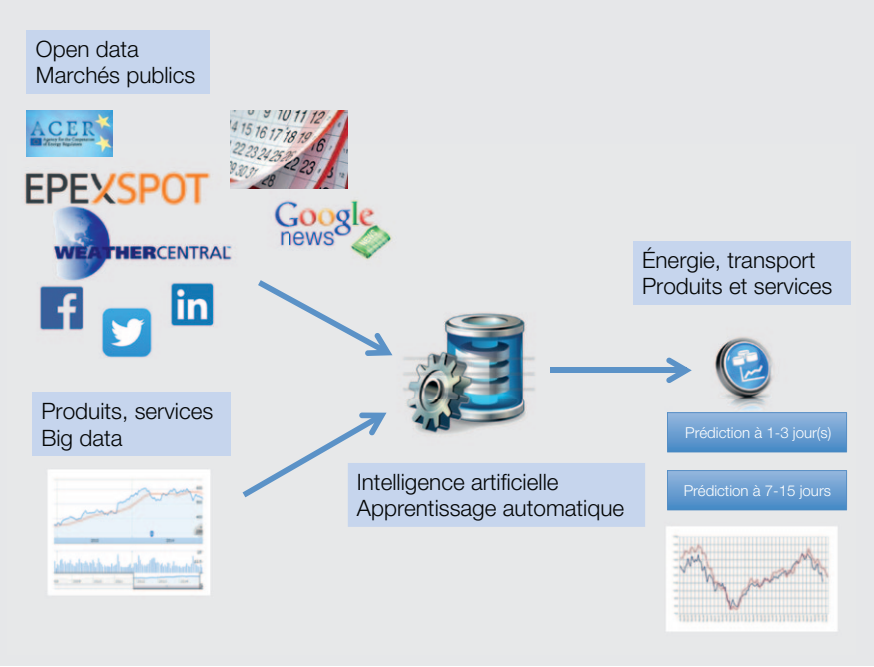


Figure 2 Sources de données pour la prédiction dans le domaine de l'énergie et du transport.

d'objectif défini les concernant. La plupart du temps, l'apprentissage non supervisé est utilisé pour faire de la segmentation (clustering) ou créer des règles d'association. Il s'agit d'essayer de trouver des structures dans les données collectées, des similitudes. Une bonne compréhension de la différence entre l'apprentissage supervisé et non supervisé permet de mieux appréhender les problèmes à résoudre.

Applications de la prévision

Dans le domaine de la prévision, il est possible de résoudre des problèmes pratiques comme la réalisation de systèmes de prévision de la demande/production (figure 1). En utilisant des données historiques concernant l'usage de certaines ressources et en faisant des corrélations avec d'autres ensembles de données, il est possible de créer un modèle capable de prédire la consommation future d'une certaine ressource.

Les applications sont multiples. Les systèmes de production et de consommation d'énergie avec différentes granularités (état, ville, quartier, bâtiment) en constituent un exemple. Il est également possible de prévoir la largeur de bande occupée d'un réseau de communication en vue d'un routage optimal ou d'optimiser une chaîne de production basée sur l'anticipation des demandes.

Dans le domaine des systèmes de recommandation d'achat, la prévision est très utilisée. Sur la base des différents comportements des clients et de leurs caractéristiques, il est possible de créer un système qui va estimer la probabilité de certaines actions des consommateurs. Il est aussi possible de faire de la détection de fraudes. Les modèles prédictifs peuvent complémenter les systèmes normalement utilisés sur un ensemble de règles. Il est possible par exemple de considérer un périmètre : pas de retrait de plus de 100 CHF par jour, pas moins de x secondes entre deux transactions avec la même carte de crédit, etc. En utilisant la prédiction analytique et sur la base des transactions historiques, il est dès lors possible de détecter plus finement celles qui sont frauduleuses.

Actuellement très en vogue dans les médias, les voitures autonomes : ces systèmes sont composés d'une grande combinaison de modèles prédictifs dont l'un, par exemple, identifie automatiquement le contexte extérieur alors qu'un autre anticipe les mouvements des autres voitures et qu'un troisième prédit quelle est la meilleure décision à appliquer dans un contexte particulier, et ainsi de suite.

Premiers pas : traiter les données

Les premiers pas de l'analyse prédictive d'un projet, en admettant que la définition des objectifs ait été faite, consistent à dessiner une carte des données à disposition.

Sélectionner les données utiles

Dans ce but, il est nécessaire d'identifier chaque source de données (figure 2) qui peut être utile pour le projet. Chacune d'elles doit être résumée et décrite précisément : technologie de stockage des données, fréquence des mises à jour des données, latence des données, temps d'exécution d'une requête.

Une fois la carte établie, il s'agit de mettre sur pied des objectifs. L'idée est de développer des fonctions pour récolter les données qui seront utiles. Il s'agira

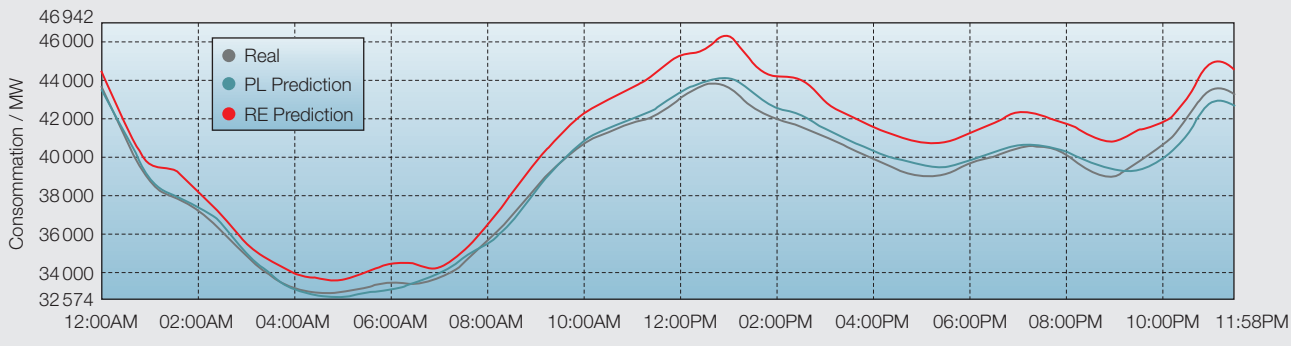
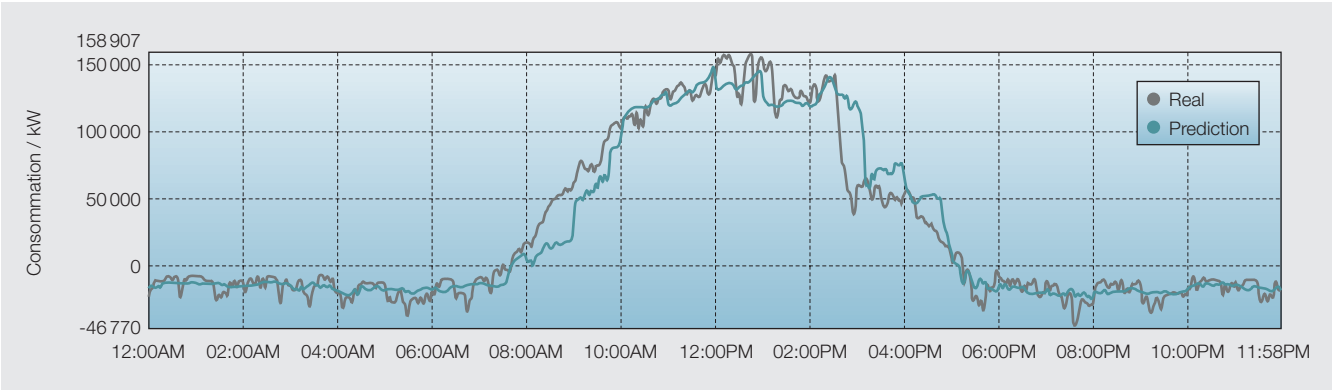


Figure 3 Modèle prévoyant la consommation d'énergie électrique de la France au jour J+1 en fonction, notamment, des données historiques et des prévisions météorologiques. La prédiction de Predictive Layer (PL) atteint une erreur moyenne de 0,82 % alors que la prédiction standard (RE) s'élève à 3,57 %.

TECHNOLOGIE **MACHINE LEARNING**



Figures: Predictive Layer

Figure 4 Modèle prévoyant la consommation d'énergie électrique d'une ville quotidiennement.

ensuite de préparer les données et de contrôler leur qualité. Il faudra vérifier qu'elles correspondent bien à ce qui est attendu et éliminer celles qui ne sont pas pertinentes, mais aussi préciser quelle est la latence de l'information de chacune des variables. Ces pas doivent être faits avant la modélisation, pour éviter une mauvaise compréhension et interprétation des futurs résultats.

Transformer les données

Le « nettoyage » et la transformation des données est un processus itératif durant toute la durée du projet. Une bonne représentation des données est fondamentale car elle permettra aux algorithmes d'être appropriés et efficaces dans leur contexte de production. Ces pas sont appelés « feature engineering » et demandent une bonne compréhension des concepts sous-jacents combinés avec une certaine dose de créativité.

Les principales compétitions en analyse prédictive (par exemple Kaggle) sont souvent remportées par des compétiteurs qui ont trouvé un nouveau moyen de représenter les données. Une question légitime consiste à se demander pourquoi ne pas étudier toutes les transformations possibles et n'utiliser que les meilleures. Le problème vient de la complexité engendrée par toutes ces transformations. Des systèmes tels que des systèmes de détection de fraude ont besoin d'intégrer des variables sur des périodes de temps assez courtes (combien de transactions passées la dernière heure, ces quatre dernières heures, etc.) alors que la modélisation médicale se concentre plutôt sur les données statiques du patient.

Échantillonnage des données

L'échantillonnage des données consiste en quelques ajustements concernant la distribution de ces dernières. On peut soit suréchantillonner les données en ajoutant des lignes générées artificiellement afin de rééquilibrer les distributions, soit les sous-échantillonner en réduisant la taille de la base pour deux raisons possibles : rééquilibrer les distributions ou bien travailler sur un ensemble de données réduit pour l'utilisation d'algorithmes plus coûteux en mémoire et temps de calcul.

Identification des variables de prédiction

Les variables de prédiction sont des variables qui peuvent être utilisées pour l'application de la modélisation. Il s'agit de variables d'entrée qui seront corrélées à la variable cible dans le cas de l'apprentissage supervisé, mais elles sont parfois complexes à identifier à cause de la latence.

Types de prédiction

Il existe deux types de prédiction : la classification et la régression.

La classification se réfère à la création d'au moins deux catégories d'éléments. Ceci implique l'utilisation d'algorithmes spécifiques pour la classification. Les questions typiques auxquelles on veut répondre sont les suivantes : « Est-ce que le client va se désabonner ? », « Est-ce que la transaction est frauduleuse ? », « Est-ce que le client va contester la facture ? », « Quel est le type de films préféré de ce client ? », « Est-ce que ce client va réagir à cette campagne de marketing ? ». Certains algorithmes, tels que les arbres de décision, k-moyennes (k-means), ou séparateurs à vaste marge (support vector machines) sont applicables aux deux types de prédiction, contrairement à d'autres : régression logistique (logistic regression) ou classification naïve bayésienne (Naive-Bayes). Suivant les objectifs de classification, les algorithmes peuvent optimiser différentes métriques.

La régression fait quant à elle référence à des cibles, à des trajectoires. Comme pour la classification, il y a des algorithmes spécifiquement élaborés pour la régression. Les questions typiques auxquelles il s'agit de répondre sont les suivantes : « Combien de personnes vont prendre le train demain ? », « Quelle sera la consommation globale d'énergie en Suisse demain ? », « Quel sera le trafic informatique d'une entreprise entre 12h00 et 13h00 ? », « Combien d'attaques informatiques est-ce que mon système va subir le mois prochain ? », « Quelle sera la consommation électrique journalière de la France **(figure 3)** ou d'une ville **(figure 4)** ? ». Chaque algorithme de régression peut en principe être utilisé pour la classification, en particulier pour les classes binaires. Cependant pour la classification multiclasses, la notion de rang va apparaître entre les classes ciblées, ce qui peut conduire à une mauvaise interprétation des résultats et donc de mauvaises performances si on ne fait pas la distinction.

Quelques algorithmes pour la prédiction

La classification naïve bayésienne [2] est un type de classification basé sur le théorème de Bayes avec une hypothèse forte qui est l'indépendance des variables explicatives. On pourrait parler de « modèles à variables statistiquement indépendantes », d'où le terme « naïf ». Intuitivement, un classificateur bayésien naïf suppose l'existence d'une caractéristique pour une classe, indépendante de l'existence d'autres caractéristiques. Suivant la nature du modèle, ces classificateurs bayésiens naïfs peuvent être entraînés efficacement dans le cadre d'un apprentissage supervisé. Dans de nombreux cas, l'estimation des paramètres se fait avec le maximum de vraisemblance. L'avantage de cet algorithme est d'être relativement facile à implémenter et rapide.

TECHNOLOGIE **MACHINE LEARNING**

La régularisation Tikhonov (connue aussi sous le nom de régression d'arête ou « ridge regression ») est une méthode de régularisation pour la résolution de problèmes mal posés. Les coefficients définis par cette méthode minimisent la somme des carrés des termes erreurs auxquels on a ajouté une pénalité. Cette méthode est rapide et explicite.

La méthode Lasso (Least Absolute Shrinkage and Selection Operator) est utile quand le nombre d'observations est légèrement supérieur ou inférieur au nombre de variables explicatives. Dans le cas normal, le nombre d'observations est largement supérieur au nombre de variables explicatives, c'est pourquoi l'estimateur usuel des moindres carrés lui est préféré.

La méthode des k plus proches voisins (k-nearest neighbor ou k-NN) est une méthode d'apprentissage supervisé [3]. On a à disposition une base d'apprentissage constituée de n couples « entrée-sortie » (x_i, y_i). Lors d'une nouvelle entrée x^{n+1}, il s'agira de parcourir la base de couples à disposition pour choisir, selon une métrique à définir, quelle est la valeur de x^k la plus proche de la nouvelle entrée x^{n+1}. Cette méthode est facile à implémenter mais est très coûteuse en temps de calcul et elle est relativement peu performante.

Les méthodes de boosting (ada-boost, etc.) ont l'avantage de réduire le biais des modèles et se distinguent particulièrement pour les problèmes complexes, mais elles ont le désavantage d'être sensibles au bruit. Finalement, les réseaux de neurones artificiels sont une autre manière de faire de la prédiction.

Tests

Les procédures de test sont essentielles pour s'assurer des performances et de la durabilité des modèles de prédiction, surtout lorsque ces derniers passent en production. Il y a plusieurs approches pour valider la qualité d'un modèle, qu'il faut adapter au contexte. La « cross-validation » est l'une des méthodes les plus utilisées dans la communauté scientifique. Elle permet d'effectuer l'apprentissage et la validation d'un modèle à l'aide d'un ensemble de données. L'idée consiste à apprendre sur une partie des données et à tester le modèle à l'aide du reste des données. Il est ainsi possible de donner un score à l'ensemble des données pour évaluer la performance et la qualité du modèle.

Dans le cas où le temps fait partie de la modélisation, il est préférable de prendre dans la mesure du possible deux ensembles de données différents, l'un pour l'apprentissage et l'autre pour la validation. En général, il est également recommandé de valider les modèles sur des ensembles de données qui ont une distribution différente de ceux qui ont été utilisés pour l'apprentissage afin de valider la robustesse du modèle. L'un des pièges les plus courants consiste à surestimer les capacités réelles du modèle. Quand un modèle est en production, les changements de la distribution des données (inévitable dans le cas de la prédiction) peuvent introduire des perturbations de certains algorithmes et affecter ainsi leurs performances quant à la prédiction.

Conclusions

Il s'agit de garder en mémoire que la plupart des entreprises les plus performantes en 2018 vont gérer leurs données en faisant usage de l'analyse prédictive. Beaucoup d'éléments favorisent en effet l'essor de ce domaine : la disponibilité d'un plus grand nombre de données, la puissance des ordinateurs, une plus grande familiarité avec le domaine du traitement des données ou encore une disponibilité plus grande de logiciels spécialisés.

Il reste encore à noter que l'analyse prédictive concerne plus l'identification des données utiles à intégrer dans le modèle que le modèle lui-même. De plus, si la plupart des premières analyses restent assez simples, elles se complexifient cependant avec le temps et l'expérience. Finalement, il ne faut pas oublier que la validation constitue une alliée de taille si elle est effectuée correctement.

Zusammenfassung **Prädiktive Zeitreihenanalyse**

Erweiterte Prognose mithilfe von Machine Learning

Eine immer grössere Anzahl an Daten steht zur Verfügung, und zwar insbesondere aufgrund der Open-Data-Initiative und dem aufkommenden Internet der Dinge. Wurden diese Daten vorher sporadisch für bestimmte Zwecke gesammelt, so werden sie jetzt zusammengefasst, um immer autonomere Systeme zu speisen, die allmählich in der Lage sind, umfassende Informationen zu liefern.

Die Anwendungsmöglichkeiten dieser maschinellen Lernverfahren (Machine Learning) sind vielfältig. Hierzu zählen beispielsweise Energieerzeugungs- und -verbrauchssysteme mit unterschiedlichen Detailebenen (Staat, Stadt, Viertel, Gebäude). Möglich ist auch die Vorhersage der von einem Kommunikationsnetz belegten Bandbreite im Hinblick auf ein optimales Routing oder die Optimierung einer Produktionskette gemäss der antizipierten Nachfrage.

In diesem Artikel werden maschinelle Lerntechniken erläutert und wie sie bei der Prognose angewendet werden können. Anschliessend werden die verschiedenen Schritte der Datenverarbeitung vorgestellt (Auswahl, Verarbeitung, Sampling usw.), gefolgt von den Prognosearten sowie einigen der am meisten verwendeten Algorithmen und schliesslich die verschiedenen Methoden zur Validierung der entwickelten Modelle. CHe

Références

[1] Gartner identifies the top 10 strategic technology trends for 2016. Gartner Symposium, ITxpo, Gartner Inc., Orlando, 4-8 October, 2015. www.gartner.com/newsroom/id/3143521

[2] Trevor Hastie, Robert Tibshirani and Jerome Friedman : The Elements of Statistical Learning : Data Mining, Inference, and Prediction. 2nd ed. Springer, 2009.

[3] Stuart Russell and Peter Norvig : Artificial Intelligence : A Modern Approach. 3rd ed. Prentice Hall, 2009.

Auteurs

Mohamed Bibimoune est chief data scientist chez Predictive Layer.
Predictive Layer, 1180 Rolle, mohamed.bibimoune@predictivelayer.com

Serge Rigori est COO et cofondateur de Predictive Layer.
serge.rigori@predictivelayer.com

Dr **Lanpeng Ji** est collaborateur scientifique à la HEIG-VD.
HEIG-VD, 1400 Yverdon-les-Bains, lanpeng.ji@heig-vd.ch

Dr **Efstratios Rappos** est collaborateur scientifique à la HEIG-VD.
efstratios.rappos@heig-vd.ch

Dr **Stephan Robert** est professeur à la HEIG-VD.
stephan.robert@heig-vd.ch