ALGORITHMES DE DÉTECTION DE VALEURS ATYPIQUES, APPLICATION AUX SÉRIES TEMPORELLES

Juin 2020

par Cheick Sanogo, Data Scientist

Contexte

La gestion efficace des données est devenue un élément clé de la stratégie d'amélioration de la performance pour les sociétés. En effet, le volume des données est en constante augmentation et elles proviennent de sources très diverses : à côté des sources plus classiques comme les données de marché et les données de transaction, des données nouvelles issues par exemple de l'activité des médias ou encore des interactions mobiles sont intégrées.

Malheureusement, cette quantité et cette diversité s'accompagnent trop souvent d'une mauvaise qualité de la data qui empêche encore la plupart des organisations de réaliser leur plein potentiel. Nous avons alors recours aux techniques de Data Quality, qui permettent d'améliorer la qualité des données et d'assurer que ces dernières soient utilisables. Mais comment gérer les types de données semi-structurées et non structurées, qui introduisent une complexité supplémentaire dans la validation des données et qui font du nettoyage manuel une solution non viable ?

Les principaux problèmes traités dans le cadre de la Data Quality sont les suivants : combler les lacunes dans les données, évaluer la pertinence, détecter les anomalies, identifier et effacer les doublons. À cela, il faut ajouter que les règles de l'entreprise concernant les données changent si fréquemment que nous avons besoin de systèmes suffisamment intelligents et agiles pour s'adapter à ces changements à un rythme plus rapide.

Le Machine Learning constitue une approche intéressante pour traiter les problèmes de Data Quality. Dans cet article, nous nous intéressons plus particulièrement à l'utilisation du Machine Learning pour la détection de valeurs atypiques. Dans ce qui suit, nous présentons donc différents algorithmes de détection de valeurs atypiques puis décrivons une application spécifique aux séries temporelles.

1 - Quelques algorithmes préconisés pour la gestion de la Data Quality

Chaque algorithme de détection de valeurs atypiques affecte un score d'anomalie aux points du set de données. Plus ce score est élevé plus le point est considéré atypique.

Le premier algorithme que nous listons est basé sur les K plus proches voisins ou **KNN** (k nearest neighbors). Le principe de cet algorithme est qu'un point est considéré atypique s'il est loin de ses plus proches voisins. Pour chaque point du set de données, le score d'anomalie est égal à la moyenne des distances entre le point et ses K plus proches voisins. Nous pouvons, bien sûr, utiliser d'autres métriques, comme la médiane, selon l'interprétation et la vision métier que nous avons des données.

Local Outlier Factor (LOF) est un autre algorithme basé sur les k plus proches voisins, qui examine la densité locale d'un point et la compare aux densités associées à ses voisins. Si la densité en un point est plus petite que les densités de ses voisins alors le point est considéré atypique. On définit pour chaque point x, $\mathrm{D}_k(x)$ sa distance par rapport à son k-ième plus proche voisin, $\mathrm{N}_k(x)$ l'ensemble de ses k plus proches voisins. La distance d'accessibilité $\mathrm{R}_k(x,y)$ de x par rapport à y se définie comme étant le maximum entre $d(x,y)$ et $\mathrm{D}_k(y)$. C'est une distance qui n'est pas symétrique. On définit la distance d'accessibilité moyenne $\mathrm{AR}_k(x)$ de x comme étant égale à la moyenne des distances d'accessibilité de x avec tous les points de son voisinage $\mathrm{N}_k(x)$ et on définit la densité d'accessibilité locale $\mathrm{f}_k(x)$ comme étant l'inverse de la distance d'accessibilité moyenne. Alors le LOF au point x est égal à la moyenne du rapport $\mathrm{f}_k(y)/\mathrm{f}_k(x)$ pour tous les y dans $\mathrm{N}_k(x)$. Le LOF mesure l'écart local d'un point par rapport à ses k voisins les plus proches.

Isolation Forest est un algorithme basé sur les arbres de décision. Il a quelques similarités avec Random Forest. C'est une combinaison d'un ensemble d'arbre d'isolement. Un arbre d'isolement

est construit en sélectionnant récursivement au hasard un attribut de la donnée, puis en choisissant aléatoirement une valeur de cet attribut comprise entre le minimum et le maximum de l'attribut, cette valeur sera le critère de split pour partitionner la donnée. Le partitionnement récursif est fait de manière à isoler les instances en nœuds avec de moins en moins d'instances jusqu'à ce que les points soient isolés en nœuds singleton contenant une seule instance. Isoler signifie séparer un point du reste des points. Les points atypiques étant différents, ils sont plus faciles à isoler. Isolation Forest se base sur le principe que les points atypiques ont des valeurs des attributs très différentes de celles des points normaux, donc plus facile à isoler. Ainsi, les branches des arbres d'isolement contenant les points atypiques sont sensiblement moins profondes, par conséquent la distance de la feuille à la racine est utilisée comme l'inverse du score d'anomalie. L'étape de combinaison finale est effectuée en faisant la moyenne des longueurs de trajet des points de données dans les différents arbres de l'Isolation Forest et l'inverse donne le score d'anomalie.

La **PCA** (analyse en composantes principales) et l'auto encoder peuvent ensuite être utilisés pour faire de la réduction de dimension sur la donnée, l'erreur de reconstruction est utilisée comme score d'anomalie. Les points atypiques sont des points qui ont de grandes erreurs de reconstruction.
Nous pouvons agréger le résultat de plusieurs algorithmes de détection de valeurs atypiques via une méthode d'ensemble, en normalisant les scores d'anomalies puis en faisant la moyenne des scores normalisés afin d'obtenir un score plus robuste.
Une fois les valeurs atypiques détectées, l'équipe métier doit intervenir dans l'identification des valeurs anormales parmi les valeurs atypiques puisque toutes les valeurs atypiques ne sont pas nécessairement anormales.

2 - Tracking de valeurs atypiques dans les séries temporelles

Intéressons-nous maintenant à la détection de valeurs atypiques dans les séries temporelles, avec comme illustration une **série temporelle financière** correspondant au cours d'une action cotée sur le marché.

La détection de valeurs atypiques dans une série temporelle peut se faire de deux façons : la détection des points de la série qui ont des valeurs atypiques et la détection des sous séquences de la série qui ont des formes atypiques comparées aux autres sous-séquences.

2.1 - Détection des points atypiques d'une série temporelle

Les valeurs atypiques d'une série temporelle correspondent à des changements soudains dans les tendances de la donnée. Dans de tels cas, les problèmes de continuité temporelle sont critiques et l'irrégularité présente un manque de continuité avec l'historique de la série temporelle. Ces valeurs atypiques définissent des irrégularités à des instants spécifiques de la donnée sur la base des relations entre les valeurs des données à des instants temporels adjacents. Une telle approche permet de détecter des changements soudains dans le processus sous-jacent.

L'objectif est d'affecter à chaque point de la série un score d'anomalie. Le score d'anomalie associé à un point de la série correspond au degré d'irrégularité affecté à ce point. Plus ce score est élevé plus le point est atypique.

Pour calculer le score d'anomalie pour les points de la série, un modèle d'auto régression tel que **ARIMA** ou un réseau de neurone tel que **LSTM** ou **GRU** est entrainé sur la série dans le but de faire de la prédiction. Une fois le modèle entrainé, le score d'anomalie en chaque point de la série est la déviation en valeur absolue entre la valeur prédite par le modèle et la valeur observée sur la série.

Les 10 premiers points atypiques détectés avec le LSTM. Données utilisées : un des composants de l'indice CAC40, données historiques de 2005 à 2012.

2.2 - Détection de sous-séquences atypiques d'une série temporelle

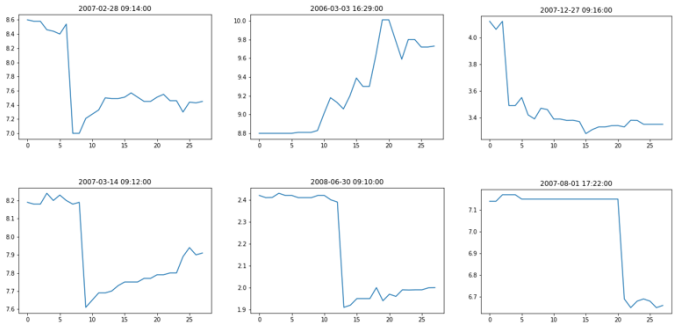
Certains types d'irrégularités sont basés non seulement sur les écarts individuels des points de la série, mais également sur les formes des parties spécifiques de la série par rapport aux autres parties de la série. Ainsi, le but est de déterminer des sous séquences dans lesquelles la série temporelle se comporte différemment par rapport à l'ensemble

des sous séquences. Chaque sous séquence est considérée comme une série temporelle indépendante, le problème revient alors à la détection de valeurs atypiques parmi plusieurs séries temporelles.

L'identification d'une irrégularité implique de comprendre qu'un point de données (sous séquence) est différent des autres. Le principe est de modéliser les points atypiques comme des points isolés du reste de la donnée en se basant sur les distances ou les similarités. Nous dirons qu'un point est atypique si sa distance par rapport aux autres est grande. Cependant les algorithmes de détection de valeurs atypiques diffèrent dans la façon dont cette distance est évaluée.

Chaque point de donnée (sous séquence) étant considéré comme une série temporelle, calculer la distance entre les séries temporelles nécessite une distance autre que la distance euclidienne, puisque celle-ci ne prend pas en compte les décalages temporels.

Pour se débarrasser de la dépendance temporelle, nous pouvons appliquer une transformation de signal à toutes les sous séquences, avec par exemple la transformation discrète des ondelettes ou **Discret Wavelet Transform** (DWT), et obtenir les coefficients. L'intérêt de cette transformation est que l'on peut désormais traiter la nouvelle représentation comme une donnée multidimensionnelle au lieu d'une donnée orientée dépendante car les dépendances à court terme et à long terme de la donnée sont codées à l'intérieur des coefficients DWT.
Ainsi, avec cette nouvelle transformation, la distance euclidienne peut être utilisée entre les coefficients DWT pour calculer la distance entre deux points (sous séquences) du set de données et n'importe quel algorithme de détection de valeurs atypiques peut être utilisé.



Les 6 premières sous-séquences atypiques détectées par une combinaison d'algorithmes de détection d'anomalies, classées par score d'anomalie croissant.
Données utilisées : un des composants de l'indice CAC40, données historiques de 2005 à 2012.

Conclusion

Différents types d'irrégularités peuvent être définis dans les séries temporelles selon qu'il est souhaitable d'identifier les points de déviation dans les séries ou d'identifier les sous séquences de formes inhabituelles.

Nous avons utilisé un apprentissage supervisé pour la détection des valeurs atypiques de la série en utilisant le LSTM pour faire de la prédiction et calculer les scores d'anomalie, puis un apprentissage non supervisé pour la détection des sous séquences de la série dans lequel nous avons utilisé une méthode d'ensemble avec différents algorithmes. Parmi les algorithmes utilisés le KNN détecteur et le LOF déterminent les valeurs atypiques grâce à l'utilisation d'informations de proximité entre les points et nécessitent beaucoup de calculs, en général le LOF détecte plus de valeurs atypiques que le KNN détecteur. Isolation Forest a un temps de calcul très faible, mais a tendance à favoriser des types spécifiques de valeurs atypiques comme les valeurs extrêmes.

Quand le set de données montre des corrélations significatives entre les différents attributs alors une modélisation linéaire avec la PCA ou une modélisation non linéaire avec l'auto-encoder permet de détecter les valeurs atypiques de façon robuste. En général l'auto-encoder détecte plus de valeurs atypiques que la PCA, mais le temps de calcul de la PCA est plus faible.

A propos de Coperneec

"From revolution to performance"

Coperneec est un cabinet de conseil cross-sectoriel spécialiste de la valorisation de la Data. Nous intervenons sur l'ensemble de la chaîne des savoir-faire autour de la Data Science, la Data Analyse et du Data Management.
Nos méthodes et techniques scientifiques éprouvées permettent de résoudre des problématiques dans tous les secteurs de l'industrie.

Notre vocation : extraire la connaissance à partir des données et pérenniser les avancées technologiques qui en découlent. La R&D est au cœur de notre ADN et les expertises de nos consultants (data scientists, data analysts, data engineers) sont en permanence challengées afin d'accompagner au plus près les révolutions technologiques et scientifiques.

Contactez-nous

Aymeric Lisbonne
Partner
alisbonne@coperneec.com
06 88 69 67 75

coperneec

est une marque de

canopee group