Défi IA 2020

Détection d'anomalies dans les données des accéléromètres de AIRBUS

Joseph ASSAHOUA
Essan Armelle KADIA
Marc MFOUTOU
Youssouf SOUMAHORO

Janvier 2020

Présentation

Présentation

- Institut National Polytechnique Félix Houphouet Boigny - INP-HB
  - Enseignement Supérieur, Côte d'Ivoire.
  - Position géographique: Yamoussoukro capitale politique située à 230 kilomètres d'Abidjan la capitale économique.
  - Directeur général : M. KOFFI N'guessan
- International Data Science Institute:
  - Master DATA SCIENCE - BIG DATA
  - Directeur: M. TANOH Tanoh Lambert

Enjeux métier

- Données des capteurs des plateformes AIRBUS
- Séquences d'une minute de mesure d'accéléromètre.
- Un individu = une séquence.
- Le problème métier : expliquer si une séquence est une anomalie ou pas.

Objectif

Construire des méthodes permettant de détecter des changements anormaux, connaissant des séquences normales.

Les données

Les données

- Une base d'apprentissage $\mathcal{B}_1$: 1677 séquences.
- Une base de validation $\mathcal{B}_2$: 594 séquences.
- Une base de test $\mathcal{B}_3$: 1917 séquences.

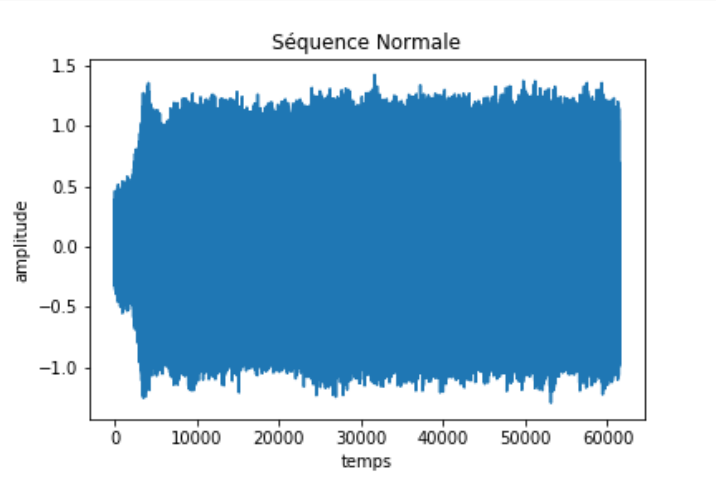
Notations :

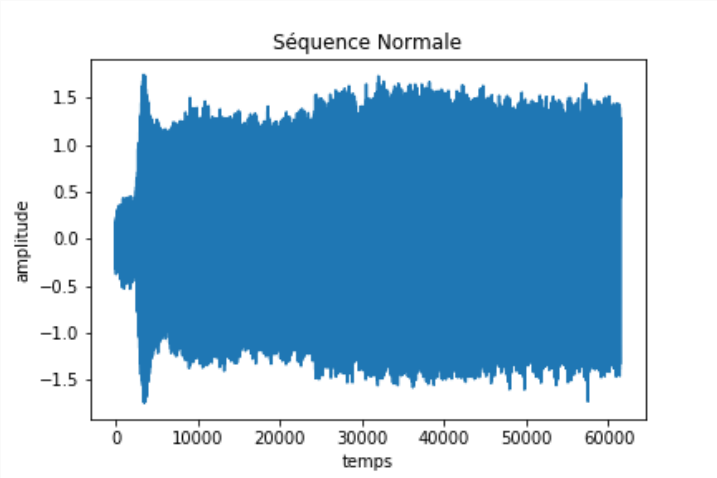
- $x_i^{(1)}$:= données de $\mathcal{B}_1$.
- $x_i^{(2)}$:= données de $\mathcal{B}_2$.
- $x_i^{(3)}$:= données de $\mathcal{B}_3$.

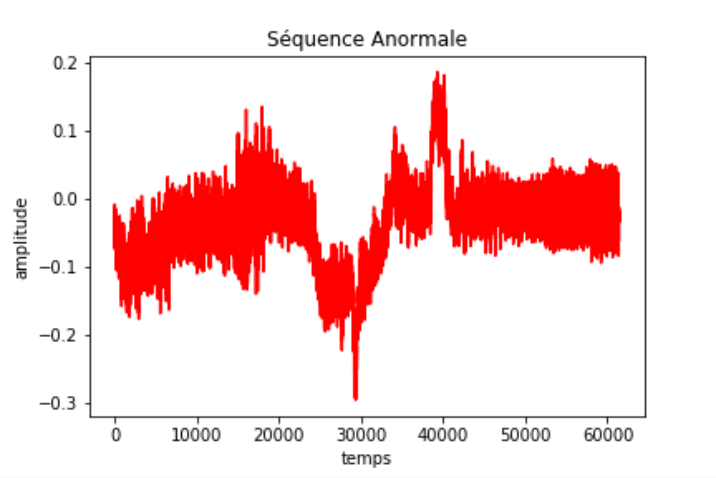
Problème statistique

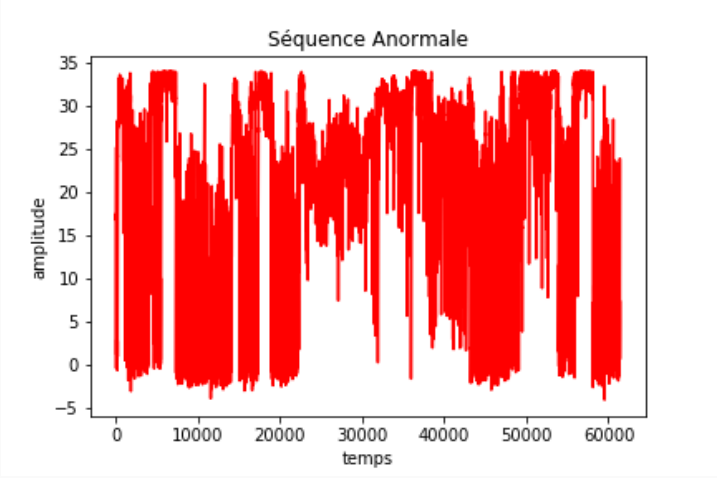
Détecter les courbes anormales dans $\mathcal{B}_2$ (puis $\mathcal{B}_3$ pour la phase finale)

Les données: Visualisation









Identification des séquences

	mean	std	min	25%	50%	75%	max
1	0.00372549	0.695547	-1.2953	-0.707911	-0.00271438	0.64168	1.42342
2	0.0127751	0.852313	-1.74512	-0.790789	0.0290494	0.794017	1.74834

(a) Résumé statistique: séquences normales

	mean	std	min	25%	50%	75%	max
918	-0.035031	0.0561867	-0.296367	-0.0629588	-0.0296148	-0.00197432	0.186683
1101	19.3072	11.1386	-4.1365	9.23659	21.3579	29.4759	34.0158

(b) Résumé statistique: séquences anormales

Méthodes et résultats

La démarche tourne autour de trois étapes essentielles.

- Etape 1 : Apprentissage non supervisé
  - IsolationForest
- Etape 2 : Apprentissage supervisé
  - RandomForest
  - Gradient Boosting
  - K-ppv
- Etape 3 : Aggrégation de méthodes

Etape 1 : Apprentissage non supervisé

Idée

Apprendre sur la base $\mathcal{B}_1$ des séquences normales $x_i^{(1)}$.

Algorithme : IsolationForest

- Algorithme permettant de détecter des anomalies dans un jeu de données.
- Calcule le score d'anomalie pour chaque donnée du jeu.
- Il isole les données atypiques, autrement dit celles qui sont trop différentes de la plupart des autres données.

1. On applique l'algorithme sur la base $\mathcal{B}_1$.
2. On prédit sur les séquences $x_i^{(2)}$ de $\mathcal{B}_2$.
3. **Résultats** :
    - 260 séquences anormales.

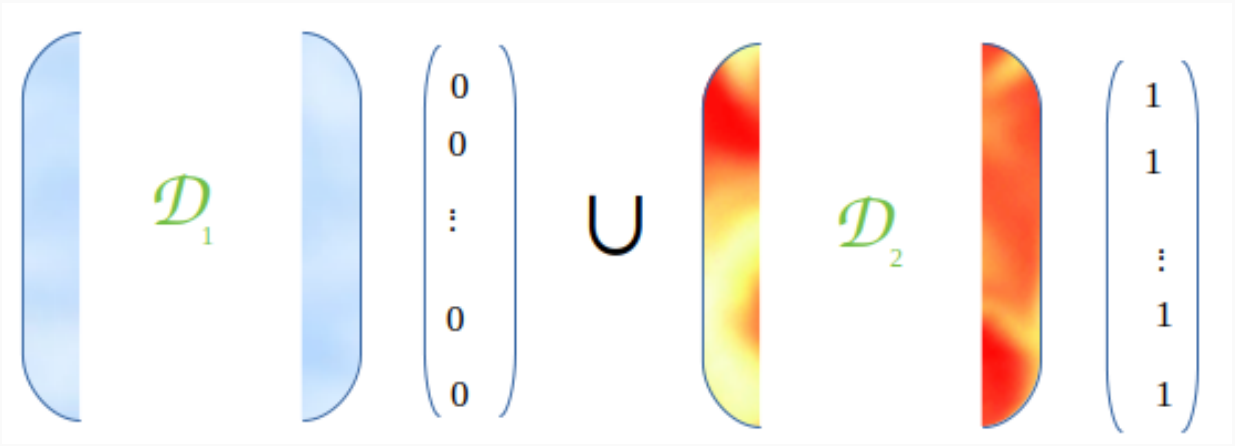
Etape 2 : Apprentissage supervisé

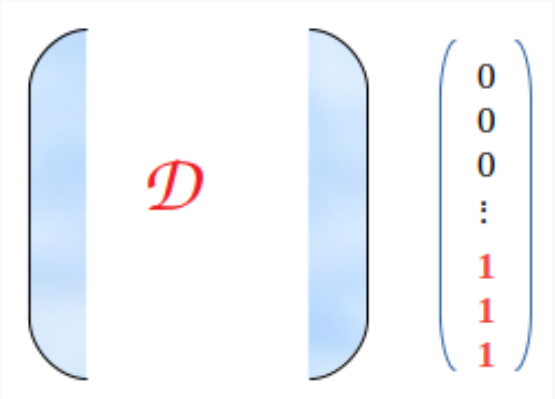
Nouvelle base d'apprentissage

- On construit un échantillon $\mathcal{D}_1 = \left\{(x_i^{(1)}, z_i)\right\}$ de taille n_1 tel que $x_i^{(1)} \in \mathcal{B}_1$ avec $z_i = 0 \ \forall i = 1, ..., n_1$.
- On construit un échantillon $\mathcal{D}_2 = \left\{(x_i^{(2)}, y_i)\right\}$ de taille n_2 avec $x_i^{(2)}$ une séquence anormale de $\mathcal{B}_2$ et $y_i = 1 \ \forall i = 1, ..., n_2$.
- La nouvelle base d'apprentissage:

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$$

Illustration





Etape 2 : démarche

- Choix du critère : on considérera l'AUC et l'erreur de classification.
- On effectue une validation croisée en 10 blocs sur $\mathcal{D}$ pour choisir l'algorithme.

Méthodes

- RandomForest
- GradientBoosting
- k-ppv

- On prédit les séquences $x_i^{(2)} \in \mathcal{B}_2$ qui n'ont pas été prédites comme étant des anomalies à l'étape 1.
- $x_i^{(2)}$ est considérée comme anomalie si:

$$\frac{1}{3}\sum_{j=1}^{3} p_{ij} \geq 0.7$$

p_{ij} := la probabilité que $x_i^{(2)}$ soit prédite comme étant une anomalie par la méthode j.

Résultats:

- 297 séquences anormales détectées dans la base de validation $\mathcal{B}_2$.
- 297 séquences normales.

Etape 3 : Aggrégation de méthodes

idée

- Reconstituer l'ensemble d'apprentissage.
- Construire des échantillons bootstrap.
- Appliquer les méthodes sur chaque échantillon bootstrap.

Etape 3 : Démarche

Soient:

- $\mathcal{B}_2^{(0)}$ l'ensemble des données normales dans $\mathcal{B}_2$.
- $\mathcal{B}_2^{(1)}$ l'ensemble des données anormales dans $\mathcal{B}_2$.
- Pour $k = 1, ..., 100$
  - On construit un échantillon $\mathcal{L}_1^{(k)}$ de taille 150 à partir de $\mathcal{B}_1$.
  - On construit un échantillon $\mathcal{L}_2^{(k)}$ de taille 150 à partir de $\mathcal{B}_2^{(0)}$.
  - On obtient l'échantillon bootstrap :

$$\mathcal{L}^k = \mathcal{L}_1^{(k)} \cup \mathcal{L}_2^{(k)} \cup \mathcal{B}_2^{(1)}$$

.

Un échantillon bootstrap ressemble à ceci:

$$\begin{pmatrix} x_{11}^{(1)} & \cdots & x_{1p}^{(1)} \\ \vdots & \ddots & \vdots \\ x_{n1}^{(1)} & \cdots & x_{np}^{(1)} \\ x_{11}^{(2)} & \cdots & x_{1p}^{(2)} \\ \vdots & \ddots & \vdots \\ x_{n1}^{(2)} & \cdots & x_{np}^{(2)} \\ x_{11}^{(2)} & \cdots & x_{1p}^{(2)} \\ \vdots & \ddots & \vdots \\ x_{m1}^{(2)} & \cdots & x_{mp}^{(2)} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Etape 3 : Algorithme

Algoritme

Soient p le nombre de méthodes, K le nombre d'échantillons bootstrap.

- Pour $j = 1, .., p$
  - Pour $k = 1, ..., K$
    - On construit l'échantillon bootstrap $\mathcal{L}^{(k)}$
    - On applique la méthode j sur $\mathcal{L}^{(k)}$.
    - On prédit sur l'ensemble test $\mathcal{B}_3$: on note $g_{jk}(x_i^{(3)})$ la prédiction de $x_i^{(3)} \in \mathcal{B}_3$.
  - La prédiction de $x_i^{(3)}$ à l'issue de $K = 100$ échantillons bootstrap est donnée par :

$$g_j(x_i^{(3)}) = \begin{cases} 1 & \text{si } \sum_{k=1}^{K} 1_{g_{jk}(x_i^{(3)})=1} > 50 \\ 0 & \text{sinon} \end{cases}$$

Suite algorithme

Enfin, pour la prédiction finale de $x_i^{(3)}$, on procède par un vote majoritaire:

$$\tilde{g}(x_i^{(3)}) = \underset{k=0,1}{\text{argmax}} \sum_{j=1}^{p} 1_{g_j(x_i^{(3)})=k}$$

Résultat:

791 anomalies détectées dans la base de test.

Merci de votre aimable attention !