VSB Power Line Fault Detection

Leila Khellouf, B.Beaufils
OpenClassRooms
Email: lkhellouf90@yahoo.com — bbeaufils.ai@gmail.com

Problématique

Le Center recherche et développe des ressources énergétiques renouvelables (ENET) cherche a détecter les décharges partielles dans les signaux acquis des lignes avec un nouveau compteur. cet article propose une étude comparative de plusieurs méthodes supervisées appliquées sur un ensemble de données de ENET pour prédire la présence ou l'absence de décharge partielle (DP) dans le signal de la ligne électrique aérienne moyenne tension.

Key words: méthodes supervisées, signal, extraction de features.

1 Introduction

Les signaux des lignes électriques aériennes moyenne tension s'étendent sur des centaines de kilomètres, transférant l'electricité d'une région à une autre. Ces grandes distances rendent coûteuse l'inspection manuelle des lignes pour les dommages qui n'entraînent pas immédiatement une panne de courant, comme une branche d'arbre heurtant la ligne ou un défaut dans l'isolateur. Ces modes de dommages conduisent à un phénomène connu sous le nom de décharge partielle - une décharge électrique qui ne relie pas complètement les électrodes entre un système d'isolation. Dans le domaine de l'électricité, une décharge partielle (DP) est une " décharge électrique localisée qui court-circuite partiellement l'intervalle isolant séparant des conducteurs " sous l'effet d'une forte tension (HTB ou HTA). Leur présence conduit à une dégradation accélérée de l'isolation qu'elle soit liquide, par oxydation, ou solide, par érosion.(wikipedia).

Le côut de l'absence d'un modèle DP dans un signal est très élevé car cela entraînerait une panne de courant affectant ainsi les moyens de subsistance de la population ou cela pourrait même déclancher un incendie. Cela implique que les faux négatifs doivent être très faibles. La détection d'un modèle PD dans un signal sain doit également être évitée car cela augmenterait les coûts de maintenance car les ressources seront consacrées à la réparation d'une ligne électrique saine.Cela implique que les faux positifs devraient être moins nombreux. L'objectif principal du problème de Kaggle est de détecter le modèle de décharge partielle (DP) présent dans le signal de la ligne électrique aérienne moyenne tension.

1.1 Enoncé du problème de Machine Learning

L'objectif du problème est de vérifier si un motif DP est présent dans un signal donné ($x(t)$) ou non.

$$x(t) = Model = y = \begin{cases} 1 \ si \quad DP \quad present \quad dans \quad x(t) \\ 0 \qquad\qquad\qquad sinon \end{cases}$$

Il s'agit donc d'un problème de classification binaire.

1.2 Indicateur de performances

La mesure de performance clé utilisée par Kaggle pour évaluer les performances du modèle est le coefficient de corrélation de Matthews (MCC). La formule de calcul du MCC est illustrée à la figure suivante.

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Avec TN= True Negative, FP= False Positive, FN= False Negative et TP= True Positive.

1.3 Feature extraction

1.3.1 Statistical Features

L'ensemble du signal comprenant 800 000 points de données. Les 14 features statistiques suivantes sont calculées pour chaque fenêtre du signal:

Mean
Standard deviation
standard deviation top, bottom
median
skewless
kurtusis
percentile values: 0.10, 0.25, 0.75, 0.90
mode
interquartile
Root Mean Square Energy

1.3.2 Zero-Crossing Rate

Le Zero Crossing Rate (souvent utilisé sous sa forme abrégée ZCR) est le taux de changement de signe d'un signal. Le ZCR a beaucoup été utilisé en reconnaissance de la parole et en Recherche d'information musicale.

1.3.3 Entropy features

En théorie de l'information, l'entropie d'une variable aléatoire est le niveau moyen d' information, de surprise ou d'incertitude inhérente aux résultats possibles de la variable. - Wikipédia

Pour une source, qui est une variable aléatoire discrète X comportant n symboles, chaque symbole xi ayant une probabilité Pi d'apparaître, l'entropie H de la source X est définie comme :

$$Entropy, H(X) = -\sum_{i=1}^{n} P(x_i) log P(x_i)$$

Il est également connu sous le nom d'entropie de Shannon

1.3.4 welch_max_power_and_frequency

On définit la densité spectrale de puissance (DSP en abrégé, Power Spectral Density ou PSD en anglais) comme étant le carré du module de la transformée de Fourier, divisé par la largeur de bande spectrale, elle-même égale à l'inverse du temps d'intégration T (ou, plus rigoureusement, la limite quand T tend vers l'infini de l'espérance mathématique du carré du module de la transformée de Fourier du signal - on parle alors de densité spectrale de puissance moyenne). Ainsi, si x est un signal et X sa transformée de Fourier, la densité spectrale de puissance vaut

$$\gamma_x = \frac{X|_2}{T}$$

1.4 Transformation de Fourier rapide

Une transformée de Fourier rapide (FFT) est un algorithme qui calcule la transformée de Fourier discrète (DFT) d'une séquence, ou son inverse (IDFT). L'analyse de Fourier convertit un signal de son domaine d'origine (souvent le temps ou l'espace) en une représentation dans le domaine fréquentiel et vice versa. La DFT est obtenue en décomposant une séquence de valeurs en composantes de fréquences différentes. Cette opération est utile dans de nombreux domaines, mais la calculer directement à partir de la définition est souvent trop lente pour être pratique.Une FFT calcule rapidement de telles transformations en factorisant la matrice DFT en un produit de facteurs clairsemés (principalement zéro).

1.5 Machine Learning

Pour la classification de l'audio en décharge partielle ou Non décharge partielle. Nous avons utiliser les méthodes distinctes de classification supervisée suivantes :

La régression logistique
Random forest
Light GBM
XGBoost

1.5.1 La régression logistique

La régression logistique est la méthode de référence pour les problèmes de classification binaire.Elle s'apparente à la régression linéaire en ce sens que l'objectif est de trouver les valeurs des coefficients qui pondèrent chaque variable d'entrée. Contrairement à la régression linéaire, la prédiction pour la sortie est transformée à l'aide d'une fonction non linéaire appelée fonction logistique.

1.5.2 Random forest

Dans la mise en sac, la même approche est utilisée, mais plutôt pour estimer des modèles statistiques entiers, le plus souvent des arbres de décision. Plusieurs échantillons de données d'ntrainement sont prélevés, puis des modèles sont construits pour chaque échantillon de données. Losque on effectu une prévision pour les nouvelles données, chaque modèle en fait une prédiction et la moyenne des prédictions est calculée afin de fournir une meilleure estimation de la valeur de sortie réelle.

1.5.3 LightGBM

C'est une structure rapide, appropriée, de renforcement de gradient boosting, dépendant du calcul de l'arbre de choix, utilisée pour le positionnement, la caractérisation et de nombreuses autres missions d'IA. Comme il dépend des calculs de l'arbre de choix, il divise la feuille de l'arbre la mieux adaptée tandis que d'autres calculs de renforcement divisent la profondeur de l'arbre en deux parties, l'une judicieuse et l'autre perspicace, par opposition à la feuille. Ainsi, lors du développement sur une feuille similaire dans Light GBM, le calcul feuille par feuille peut réduire plus de malchance que le calcul par niveau et apporte par la suite une bien meilleure précision qui peut être obtenue de temps en temps par n'importe lequel des calculs de renforcement actuels. De même, il est extrêmement rapide, d'où le mot "Light".

1.5.4 XGBoost

XGBoost (comme eXtreme Gradient Boosting) est une implémentation open source optimisée de l'algorithme d'arbres de boosting de gradient. *Le Boosting de Gradient est un algorithme d'apprentissage supervisé dont le principe et de combiner les résultats d'un ensemble de modèles plus simple et plus faibles afin de fournir une meilleur prédiction.* On parle d'ailleurs de méthode d'agrégation de modèles. L'idée est donc simple : au lieu d'utiliser un seul modèle, l'algorithme va en utiliser plusieurs qui serons ensuite combinés pour obtenir un seul résultat.

2 Evaluation

2.1 Dataset

Nos expériences utilisent les données accessibles au public [1] - Chaque signal contient 800 000 mesures de la tension d'une ligne électrique, prises en 20 millisecondes. Comme le réseau électrique sous-jacent fonctionne à 50 Hz, cela signifie que chaque signal couvre un seul cycle de réseau complet. Le réseau lui-même fonctionne sur un schéma d'alimentation triphasé et les trois phases sont mesurées simultanément.

2.2 setup

- metadata_[train/test].csv

- id_measurement : le code d'identification d'un trio de signaux enregistrés en même temps. - signal_id : la clé étrangère pour les données du signal. Chaque ID de signal est unique à la fois pour le train et le test, donc le premier ID dans le train est '0' mais le premier ID dans le test est '8712'. - phase : 3 conducteurs sont utilisés pour transférer la puissance d'une région à l'autre. Chaque conducteur transporte un signal. La phase du signal dans chaque conducteur est différente. L'identifiant de phase fait principalement référence au conducteur dans lequel le signal est acheminé. Chaque identifiant de signal est associé à un identifiant de phase unique de 0, 1 ou 2.

- Target:

0 si la ligne électrique n'est pas endommagée,
1 s'il y a un défaut.

- [train/test].parquet - Les données du signal. Chaque colonne contient un signal ; 800 000 mesures int8 exportées avec pyarrow.parquet version 0.11. Veuillez noter que cela est différent de notre orientation habituelle des données d'une ligne par observation ; le commutateur permet de charger efficacement un sous-ensemble des signaux. Si vous n'avez jamais travaillé avec Apache Parquet auparavant, veuillez vous référer au noyau de démarrage de chargement de données Python.

2.3 Results

Les résultats obtenus pour l'expérience exprimés en termes de F-mesure, Précision, Rappel et MCC sont présenté dans les tableaux . La matrice de confusion détaillée de l'expérience est présenté dans la figure 2. Les résultats montrent que l'approche B-LightGBM obtient les meilleurs résultats pour l'expérience qui obtiennent une excellente MCC moyenne égale à 0,66 .

3 Conclusion

Nous avons comparé plusieurs méthodes supervisée pour la classification d'echarge partielle (DP) et Non décharge partielle. La méthode B-lightGBM montre une amélioration de la robustesse du modèle par rapport à plusieurs méthodes. Les futurs travaux consisteront à l'utilisation de d'autres features comme **les peaks** et d'autres nouvelles approches comme le deep learning

	CV	Model	f1	mcc	recall
0	0.0	Ub_LogisticRegression	0.484782	0.000000	0.500000
1	1.0	Ub_LogisticRegression	0.484782	0.000000	0.500000
2	2.0	Ub_LogisticRegression	0.484996	0.000000	0.500000
3	3.0	Ub_LogisticRegression	0.484768	0.000000	0.500000
4	4.0	Ub_LogisticRegression	0.484768	0.000000	0.500000
5	0.0	Ub_RandomForest	0.484782	0.000000	0.500000
6	1.0	Ub_RandomForest	0.484782	0.000000	0.500000
7	2.0	Ub_RandomForest	0.484996	0.000000	0.500000
8	3.0	Ub_RandomForest	0.484768	0.000000	0.500000
9	4.0	Ub_RandomForest	0.484768	0.000000	0.500000
10	0.0	Ub_LightGBM	0.695558	0.499300	0.630435
11	1.0	Ub_LightGBM	0.785790	0.581545	0.747254
12	2.0	Ub_LightGBM	0.756803	0.553391	0.695800
13	3.0	Ub_LightGBM	0.733967	0.540035	0.665756
14	4.0	Ub_LightGBM	0.784456	0.572640	0.759464
15	0.0	Ub_XGBoost	0.821655	0.658725	0.771723
16	1.0	Ub_XGBoost	0.827339	0.664596	0.785305
17	2.0	Ub_XGBoost	0.806330	0.626377	0.760156
18	3.0	Ub_XGBoost	0.824775	0.666808	0.772175
19	4.0	Ub_XGBoost	0.773211	0.558467	0.732755
20	0.0	B_LogisticRegression	0.056734	0.007335	0.500455
21	1.0	B_LogisticRegression	0.060535	0.016429	0.502275
22	2.0	B_LogisticRegression	0.061091	-0.056249	0.488934
23	3.0	B_LogisticRegression	0.092986	0.011473	0.504633
24	4.0	B_LogisticRegression	0.079356	0.014413	0.504593
25	0.0	B_RandomForest	0.640748	0.390568	0.832061
26	1.0	B_RandomForest	0.625060	0.366108	0.817990
27	2.0	B_RandomForest	0.630605	0.385085	0.838209
28	3.0	B_RandomForest	0.652853	0.414646	0.849667
29	4.0	B_RandomForest	0.654051	0.428559	0.867764
30	0.0	B_LightGBM	0.821491	0.654000	0.778059
31	1.0	B_LightGBM	0.837629	0.678247	0.812471
32	2.0	B_LightGBM	0.834206	0.668988	0.822693
33	3.0	B_LightGBM	0.877436	0.759616	0.844183
34	4.0	B_LightGBM	0.802069	0.604463	0.793874
35	0.0	B_XGBoost	0.821655	0.658725	0.771723
36	1.0	B_XGBoost	0.827339	0.664596	0.785305
37	2.0	B_XGBoost	0.806330	0.626377	0.760156
38	3.0	B_XGBoost	0.824775	0.666808	0.772175
39	4.0	B_XGBoost	0.773211	0.558467	0.732755

Fig. 1. Résultats

```
1. The F-1 score of the model 0.8212307639954011

2. The recall score of the model 0.7634396737496771

3. The Matthews Correlation Coefficient: 0.6647435081999267

4. Classification reporct
              precision    recall  f1-score   support

           0       0.97      0.99      0.98      5493
           1       0.87      0.53      0.66       344

    accuracy                           0.97      5837
   macro avg       0.92      0.76      0.82      5837
weighted avg       0.97      0.97      0.96      5837


5. Confusion matrix
 [[5465   28]
 [ 161  183]]
```

Fig. 2. Matrice de confusion

[1] https://www.kaggle.com/c/vsb-power-line-fault-detection/overview