Projet energy monitoring

Alexis DAVID , Vincent CASA NOVA, Joseph PLOT, Tom CHARDON

I. Introduction:

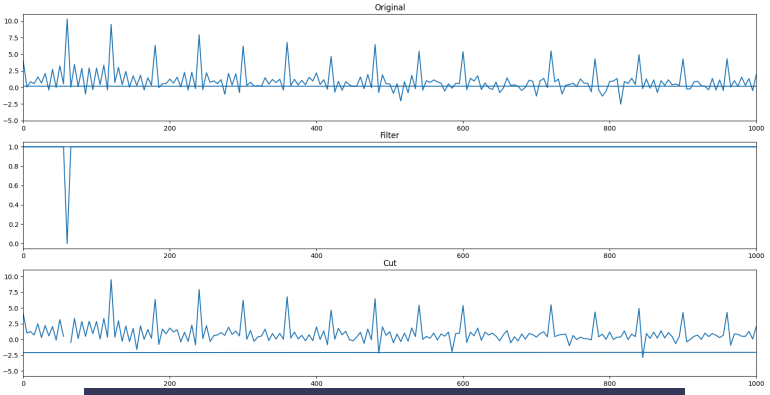
L'objectif du projet est de mettre au point une méthode d'**identification des appareils allumés dans une maison** à partir de la consommation électrique globale.
Pour cela nous avons étudié la consommation dans le **domaine spectrale**.

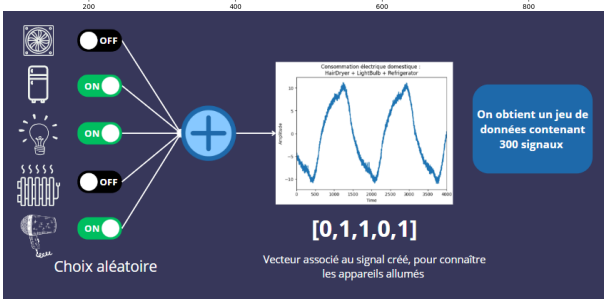
A. Jeu de donnée de départ:

A notre disposition nous avons des courbes de consommation correspondant à différents appareils domestiques.
Donc 5 appareils que nous souhaitons pouvoir identifier. A partir de cette base de données nous allons pouvoir créer une base de données plus réaliste en superposant différentes consommations.

II. Création des jeux de données:

Notre objectif est d'avoir des courbes correspondantes aux différents cas de figure possibles dans une maison.
Autrement dit, nous allons créer des courbes de consommation en superposant la consommation de différents appareils. Pour avoir un jeu de données équilibré on choisit avec une probabilité de ½ de sommer le signal de chaque appareil après en avoir retiré la composante 60Hz. Puis on ajoute à la fin le 60Hz pour ainsi avoir les signaux en phase par rapport à la porteuse. On ajoute également du bruit gaussien à nos signaux de départ pour avoir des données plus réalistes.





Suite à cette étape, on a formé 300 signaux, avec pour chaque signal un vecteur associé pour indiquer quels appareils sont allumés sur le réseau.

III. Méthodes d'identifications:

Lors de notre projet nous avons utilisé différentes méthodes d'identification.Pour chaque méthode le principe sera le suivant, on va créer des datasets supervisés d'entraînement et de validation, avec une répartition de 70% training / 30%validation.
Puis on utilisera ces datasets pour entraîner nos modèles.
Les données d'entrée ne seront pas les signaux bruts. En effet, comme les signaux ont une grande dimension (20 000) on effectuera d'abord une traduction de dimension.

A. Multi Logistic regression:

La première méthode essayée est une régression logistique multiple (généralisation d'une régression logistique simple permettant d'avoir plus de 2 catégories).
Nous avons entraîné un modèle par classe et avons calculé une accuracy et un f1 score pour chaque pour avoir une idée des performances de détection et de non détection. Puis nous en avons fait la moyenne pour avoir des métriques de performance globale.
Ce modèle est évidemment très simple et nous obtenons des résultats insuffisants. En effet, il faut que chaque classe soit linéairement séparable dans l'espace des composantes principales pour que cette méthode soit performante. C'est pour cela que nous allons considérer un modèle plus complexe.

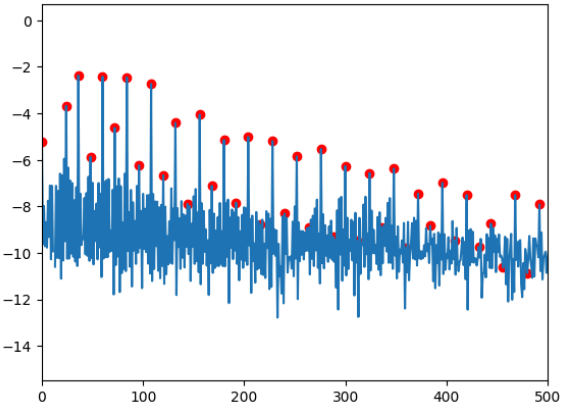
B. Convolutional Neural Network:

Nous avons voulu essayer une méthode de deep learning. Pour cela, nous avons extrait des features automatiquement avec des successions de filtres de convolution à une dimension. Cela permet un modèle plus restreint et une réduction de dimension. Ainsi les 20000 points temporels sont réduits d'un facteur 10 grâce à 3 convolutions et 3 max pooling.

La sortie du modèle est une couche dense de classification qui sort un vecteur de probabilité de présence de chaque classe.
Ce modèle est bien plus lourd à entraîner mais il offre de bonnes performances.

C. Support Vector Machine:

Nous avons essayé un modèle de SVM sur 3 types de données: Le signal temporel, fréquentiel et les harmoniques. Le signal étant quasi stationnaire nous n'avons pas utilisé de représentation temps fréquence. Voici l'extraction des harmoniques:



Ce modèle introduit des non linéarités grâce aux kernels du SVM et il permet de réduire la dimension en ne considérant que les vecteurs de support.

Le kernel choisi est radial basis function avec un gamma = 1e-3 et C = 100.
Ce modèle marche particulièrement bien pour les features harmoniques. Cela permet d'utiliser des connaissances préalables du signal pour économiser du temps de calcul.

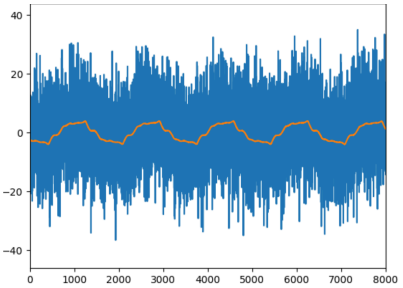
Résultats:

	feature	f1 score	accuracy
SVM	Time signal	0.85	0.87
	Harmonics	0.97	0.97
	Spectrum	0.55	0.55
CNN	Time signal	0.97	0.97
Multi Reg	Time signal + PCA (n=5)	0.65	0.74

Nous obtenons les performances suivantes pour les 3 modèles.

Nous voyons que les meilleurs modèles sont le réseau de neurones à convolution (CNN) et le SVM. Mais celui-ci est très lourd à entraîner.
En utilisant les connaissances préalables on peut utiliser les harmoniques comme feature pour entraîner un modèle moins coûteux tel que le SVM. Ainsi on a un bon compromis entre performance et complexité.

Avec un bruit beaucoup plus grand on des signaux comme le suivant :



	feature	f1 score	accuracy
SVM	Time signal	0.55	0.54
	Harmonics	0.68	0.68
CNN	Time signal	0.7	0.7
Multi Reg	Time signal + PCA (n=5)	0.	0.5

Les modèles ont les performances suivantes:

Les modèles SVM et CNN sont aussi les plus robustes avec un léger avantage pour le CNN.

Voici les pistes d'amélioration pour ce projet:

- Agrandir notre jeu de données (60 signaux par classe utilisés ici, mais beaucoup plus à disposition)
- créer des bases de données plus réalistes : entraîner sur même type de data mais tester sur des données avec des proportions plus réalistes. Par exemple le frigo souvent branché et le sèche cheveux que ponctuellement.
- Il aurait été très intéressant de considérer des combinaisons de modèles. En effet, en machine learning une combinaison de modèles peut donner une meilleure performance que la performance du meilleur modèle seul.