CentraleSupélec

Metz

Challenge Data

Algorithme en Science des données

Compte-rendu

Challenge ENS : Learning biological properties of molecules from their structure *by Simulations Plus*

BARRÉE Guillaume
CARNEIRO BARBOSA ROCHA João Felipe
ECH-CHOUINI Mehdi

Table des matières

Enseignants : Frederic Pennerath

14 Février 2022

1 Contexte

1.1 Contexte du challenge

Simulations Plus mène des recherches scientifiques dans les domaines de la **prédiction** et de la **simulation** des propriétés pertinentes pour la R&D pharmaceutique. Ils sont spécialisés dans ce qu'on appelle la **prédiction ADMET** (Absorption, Distribution, Métabolisme, Élimination et Toxicité des composés chimiques dans les organismes biologiques), PBPK (Physiologically-Based Pharmacokinetics), pharmacométrie et pharmacologie/toxicologie quantitative des systèmes. Les résultats de leurs recherches sont transformés en outils logiciels utilisés par les scientifiques du secteur pharmaceutique dans l'industrie, les universités et les agences gouvernementales, ainsi qu'en services de conseil.

1.2 Objectifs du challenge

L'objectif est de **découvrir les propriétés biologiques de nouveaux composés chimiques** en utilisant des données expérimentales déjà existantes.

Les coûts actuels de la mise sur le marché d'un nouveau médicament sont énormes, atteignant **2.0 milliards de dollars** US et **10 à 15 ans de recherche continue**. Le désir d'éliminer un grand nombre de ces coûts inutiles a accéléré l'émergence et l'acceptation de la science de la chimiométrie. En se basant sur le concept selon lequel *"les produits chimiques similaires ont des propriétés similaires"*, on prend les données expérimentales existantes Y et on **construit des modèles statistiques corrélatifs** pour créer une carte entre les structures des composés chimiques et les valeurs Y observées. Ainsi, il ne serait pas nécessaire de mesurer la propriété Y des nouveaux composés chimiques. Il suffirait de dessiner la structure d'une molécule totalement nouvelle sur l'écran de l'ordinateur et de la soumettre au modèle corrélatif pour la **prédire**.

Les ordinateurs ne peuvent pas percevoir les **structures chimiques** (atomes et connectivité interatomique) comme le font les **chimistes humains**. Un logiciel est capable de calculer **un même ensemble de N descripteurs moléculaires** par composé, où N est de l'ordre de plusieurs centaines. Comme les valeurs brutes des différents descripteurs moléculaires sont calculées sur des échelles différentes, **une normalisation à une échelle commune est nécessaire** avant la modélisation (par exemple, l'échelle -1 à 1). Tous les **descripteurs ne fournissent pas une entrée significative** dans un modèle Y = f(X) réussi. Par conséquent, **le choix des bons descripteurs** pour la modélisation est la première étape critique de la construction du modèle. Une fois que le sous-ensemble approprié de N colonnes est choisi, la matrice d'apprentissage réduite correspondante ainsi que le vecteur Y à M dimensions sont soumis à un algorithme d'apprentissage de modèle.

1.3 Description des données

Pour ce défi, nous disposons d'une matrice d'entraînement en entrée de 1087 composés chimiques, chacun avec 295 descripteurs moléculaires au format CSV (*input_training.csv*). Les colonnes $2 - 296$ contiennent des descripteurs moléculaires $X1 - X339$ étiquetés de manière non consécutive. Les valeurs de toutes les colonnes ont été normalisées à l'intervalle $[-1, 1]$. Un autre fichier (*output_training.csv*) contient une colonne de 1087 valeurs observées d'une propriété biologique Y.

1.4 Problématique étudiée dans le rapport

La dimension de notre vecteur d'entrée semble être un point important du challenge.
Ce rapport se concentrera sur l'analyse des différentes méthodes de réduction de dimension, leurs impacts et leurs résultats.

2 Analyse des données

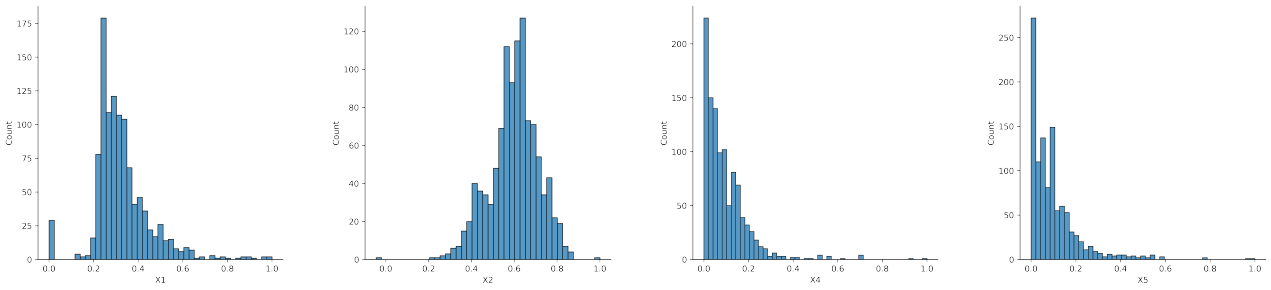
2.1 Analyse rapide des features les plus importantes

Dans ce défi, nous n'avons pas d'informations sur les caractéristiques ni sur le Y.

Afin d'analyser plus en profondeur les caractéristiques et leur influence, nous avons donc décidé d'utiliser des algorithmes d'apprentissage automatique et de tracer l'importance des caractéristiques. Nous pouvons voir des exemples de tracé de l'importance des caractéristiques dans la figure 2.

Nous avons effectué l'analyse de l'importance des caractéristiques pour les algorithmes d'apprentissage automatique où cela a un sens (Random Tress, Extra Randomized Trees and Gradient Boosting). Nous avons tracé la distribution des caractéristiques qui ont tendance à apparaître le plus souvent.

Dans la figure 1, nous avons les quatre caractéristiques principales (X1, X2, X4, X5).



(a) Feature 0 (X1) Distribution

(b) Feature 1 (X2) Distribution

(c) Feature 3 (X4) Distribution

(d) Feature 4 (X5) Distribution

FIGURE 1 – Most important features distributions

À l'avenir, ce type d'analyse, s'il est associé à une connaissance des caractéristiques et de la sortie à prédire, peut conduire à des découvertes ou à des pistes d'investigation intéressantes.

2.2 Étude des différentes méthodes de réduction de dimension

Avant de se lancer dans une étude des différentes méthodes de réduction de dimension, il est important de comprendre pourquoi ceci est nécessaire.

- visualisation des données ;
- interprétabilité des prédicteurs ;
- accélérer les algorithmes dont la complexité dépend de n ;
- les données peuvent occuper un espace de dimension inférieure à n ;
- malédiction de la dimensionnalité : les données deviennent rapidement sparses, les modèles peuvent sûr apprendre.

Il existe deux types de méthodes pour réduire la dimension de nos données à savoir

- sélectionner un sous-ensemble des caractéristiques originales : sélection de caractéristiques ;
- calculer de nouvelles caractéristiques à partir des caractéristiques originales : extraction de caractéristiques ;

2.2.1 La sélection de caractéristiques

En ce qui concerne la sélection de caractéristiques, nous allons nous concentré sur trois méthodes

Filtres : les dimensions sont sélectionnées sur la base d'une heuristique.

Wrappers : les dimensions sont sélectionnées sur la base d'une estimation du risque réel.

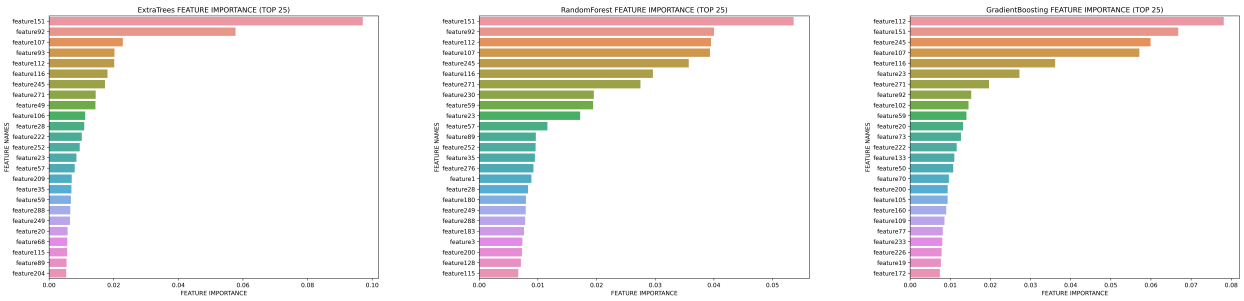
Embedded : l'algorithme ML est conçu pour sélectionner un sous-ensemble de caractéristiques.

Sélection de caractéristiques : Embedded

Les méthodes d'ensemble de régression sont un exemple de méthodes "embedded". Pour les arbres de décision, chaque nœud englobe une décision sur une variable et si l'on ne conserve que les variables utilisées pour la décision dans les nœuds, on ne conserve en effet qu'une fraction des dimensions d'entrée.

Les méthodes d'ensemble utilisées lors de notre étude sont :

- Random Forest
- Extra Trees
- Gradient Boosting



(a) ExtraTrees (b) RandomForest (c) GradientBoosting

FIGURE 2 – Most important features

Remarques : Nous pouvons déjà remarquer que les caractéristiques les plus utilisées dans les différents algorithmes sont quasiment les mêmes avec des poids légèrement différents.
Pour ExtraTrees, on remarque que quasiment toute l'information est récupérée avec la feature 151 contrairement aux deux autres ou plusieurs features impactent fortement les décisions.

Sélection de caractéristiques : Filtres univariés

Les filtres univariés classent chaque caractéristique indépendamment des autres, leur relation avec la cible. Ils sont utiles dans les grands espaces où nous ne pouvons pas nous permettre des filtres multivariés plus élaborés.

Some feature importance evaluation for feature/target types

		Target	
		Categorical	Numerical
Feature	Categorical	Chi-2 , Mutual information	ANOVA
	Numerical	ANOVA	Correlation (Pearson, Spearman)

FIGURE 3 – Différents types de filtre

Notre cible contient des valeurs numériques. Dans nos features, cela dépend. En analysant plus précisément les données, nous nous sommes rendu compte que 48 d'entre elles étaient binaires. Nous avons donc utilisé la dernière ligne du tableau 3.

Une fois les coefficients de corrélation calculé, nous pouvons extraire les caractéristiques les plus importantes.

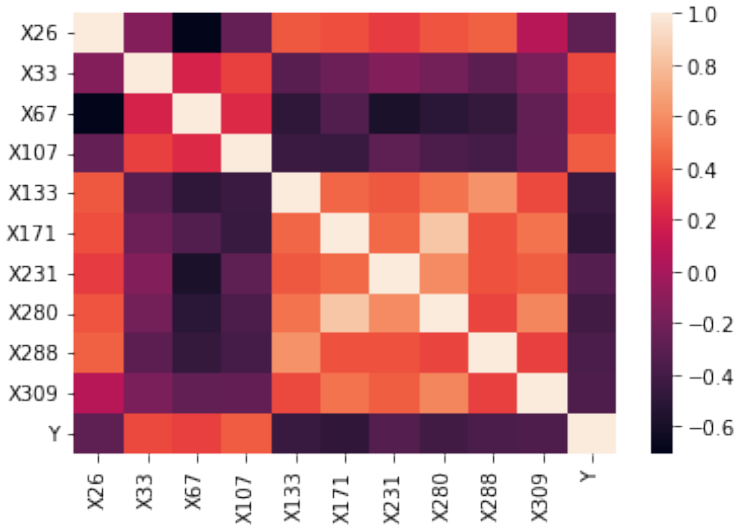
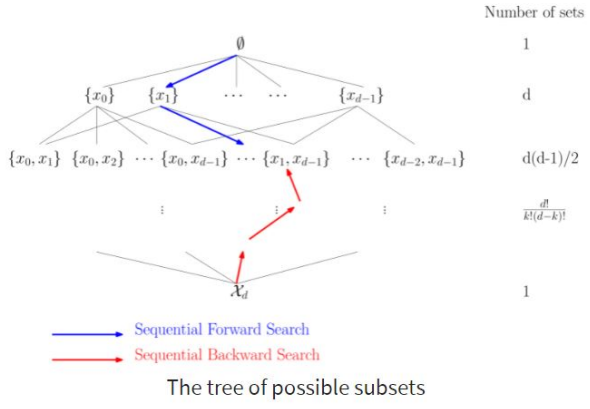


FIGURE 4 – Matrice de corrélation des caractéristiques les plus corrélées à la sortie

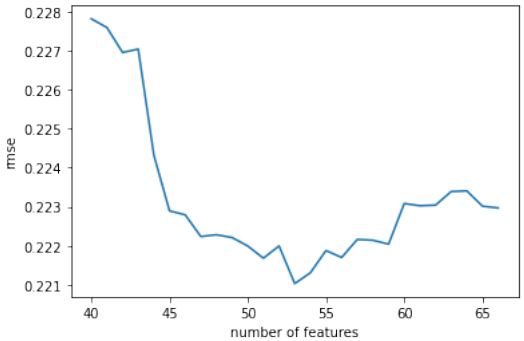
En choisissant ces entrées dans nos différents algorithmes, nous avons obtenu des résultats bien moins bon qu'en utilisant les méthodes "embedded" des algorithmes. Sur la MSE, les résultats étaient de l'ordre de 10 fois moins bons. Cela peut s'expliquer par le fait que l'on ne considère les features que par rapport à la cible. C'est à dire que l'on suppose qu'il y a indépendances des features conditionnellement à notre cible ce qui est surement faux. Nous allons voir d'autres méthodes qui permettent de limiter les hypothèses d'indépendance des features conditionnellement à notre cible.

Sélection de caractéristiques : Filtres multivariés & Wrappers

Dans les filtres multivariés, un sous-ensemble de caractéristiques est évalué par une heuristique, par exemple les statistiques entre les caractéristiques et les caractéristiques et la cible.
Par exemple, l'évaluation d'un sous-ensemble de caractéristiques basée sur la corrélation : sélectionner les caractéristiques qui sont corrélées avec la cible, mais qui ne sont pas corrélées entre elles.



(a) Les sous ensembles possibles



(b) Effet de la Forward Selection sur la rmse pour une νSVR .

2.2.2 Extraction de caractéristiques

Principal Component Analysis (PCA)

La PCA est une méthode de réduction de dimension, tout en préservant autant que possible l'information contenue dans les données d'origine. La PCA atteint cet objectif en projetant les données sur un sous-espace de dimension inférieure qui conserve la majeure partie de la variance entre les points de données.

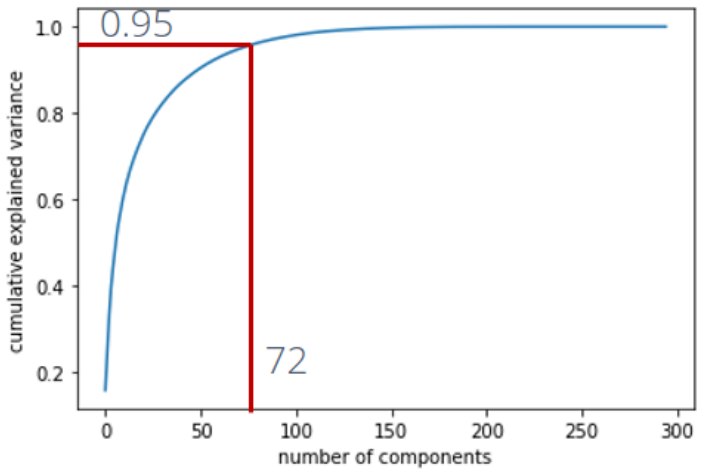


FIGURE 6 – Variance cumulatif de la PCA. On obtient une explication de 95% de la variance avec 72 composantes

3 Conclusion

Il existe beaucoup de méthodes différentes pour effectuer de la réduction de dimension. Avec une connaissance sur les données d'entrées et la sortie, il est possible d'ajouter de la connaissance pour diminuer la taille du vecteur d'entrée.

En ce qui concerne les résultats, les méthodes "embedded" semblent donner les meilleurs résultats. C'est-à-dire qu'elles s'occupent elles-même de trouver les meilleurs features.

Nous avons également utilisé des algorithmes de prédiction à base de Support Vector Machine (νSVM). Dans ce cas, bien que l'on puisse jouer sur les paramètres ν et C pour ajouter de la régularisation, appliquer une PCA avant semble avoir un impact bénéfique.

Méthode	MSE
Sans transformation	0.060
Regularisation	0.054
Selection de feature	0.049
PCA	0.051
t-SNE	0.052

TABLE 1 – Performance d'une ν-SVR en fonction de la méthode de réduction de dimensions utilisée