Far Marya
M2 DLAD
11/11/2018

Machine learning : Comparaison de modèles de prédictions pour la classification de cancers

Introduction

Les études d'expression de gène du génome entier par séquençage à haut débit des ARNm fournissent beaucoup d'informations. Entre autres, elles permettent l'application de certaines approches pour comprendre la corrélation entre les profils d'expression des gènes à des pathologies. Le cancer est une pathologie qu'il est nécessaire de comprendre à des fins thérapeutiques et de dépistage. Le diagnostic précoce et le pronostic d'un type de cancer permettent de faciliter la prise en charge clinique des patients. Le *machine learning* a la capacité d'améliorer notre compréhension des cancers car il est devenu une technique prometteuse pour le traitement de données de grande dimension, avec une application croissante dans l'aide à la décision clinique.

L'étude se portera sur l'expression d'un grand nombre de gènes de patients ayant 5 types de tumeurs différents. La recherche associée à ces données a été faite par Samuele Fiorini de l'Université de Genève. Par ailleurs, ce data set a provient du data set original du *Cancer genome atlas pan-cancer analysis project* publié en octobre 2013 dans *Nature genetics*. Toutefois, le Data-Set original prenait en considération 12 tumeurs alors que celle-ci considère 5 cancers.

L'objectif de cette étude est de mettre en place un modèle de prédiction qui pourra classer correctement les patients selon leur type de cancer à partir des profils d'expression de gènes. Il pourra éventuellement être utilisé pour classer des échantillons non catégorisés d'études cliniques. Ceci est le cas classique d'un problème d'apprentissage supervisé, qui exécute un algorithme d'apprentissage sur une partie des données (training data) et applique les prédictions sur la partie test des données. Dans ce cas ci 5 algorithmes Naive Bayes, Decision Tree, K-nearest neighbor, Random Forest et Logistic Regression vont être comparés.

Cependant le principal défi pour analyser ces données d'expression de gènes ayant un grand nombre de gènes pour le peu d'échantillon, consiste à extraire les informations en lien avec les pathologies. En effet ces données contiennent beaucoup de « bruit » et de données impertinentes, et donc l'une des étapes importantes en *machine learning* consiste à effectuer le nettoyage des données ou l'extraction des attributs informatifs seulement avant le processus d'apprentissage.

Méthodes et résultats

L'analyse qui suit a été entièrement performée par un script développé avec la version 3.7 de Python.

Présentation des données

La tables de données sur laquelle le travail est effectué a été prise du UCI Machine Learning Repository du lien suivant https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq . La table nous informe sur l'expression de gènes de 801 patients atteints de diverses catégories de cancer : breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), colon adenocarcinoma (COAD), lung addenocarcinoma (LUAD), prostate adenocarcinoma (PRAD). Il y a au total 20 531 caractéristiques (gènes) associés aux patients. Les échantillons, soit les 801 patients, sont représentés par lignes et les caractéristiques génétiques sont représentées par colonnes. Les niveaux d'expressions de gène RNA-Seq ont été mesurés par la plateforme HiSeq. La table ne présente aucune valeur nulle ou non renseignée.

Table 1 : *Nombre de patients pour chaque type de cancer présent dans la table de données*

Types de cancer	Nombre d'échantillons
BRCA	300
KIRC	146
LUAD	141
PRAD	136
COAD	78

Sélections des attributs informatifs

Afin de réduire au maximum les risques d'*overfitting* il est nécessaire de retirer les gènes non informatifs pour éliminer les impertinences qui biaiserai la fiabilité de prédiction du modèle sur les données *Test* en étant trop fidèle aux données

Train. Il faut aussi empêcher l'*underfitting* en ne sélectionnant pas assez d'attribut, et donc ne pas fournir assez d'information pour entrainer le modèle.

La méthode utilisé pour sélectionner les meilleurs attributs utilisé est celle du Chi-carré, le Chi-carré entre chaque attribut et la classe cible est calculé. Ensuite le nombre desiré k, d'attributs ayant les plus grandes valeurs pour la statistique du test Chi-carré est sélectionné pour former le nouveau set de gènes les plus informatifs.

La sélection été performer sur les 5 algorithmes de prédictions choisi pour effectuer la classification, qui sont les suivant ; Naive Bayes, Decision Tree, K-nearest neighbor, Random Forest et Logistic Regression.
Pour chacun des algorithmes la démarche est la suivante :

- 8 nombres d'attributs différents ont été choisis (10, 50, 100, 200, 500, 1000, 10000, 20531).
- Pour chaque nombre d'attributs sélectionnées est formé un nouveau jeu de données à partir duquel les set Train (70% des échantillons) et Test (30% des échantillons) sont formés.
- Les valeurs d'expression de gènes sont ensuite normalisées étant donné que les algorithmes sont généralement basés sur la mesure de distance entre les échantillons.
- Les modèles sont ensuite fitté aux données Train.
- L'évaluation du modèle est faite par le calcul de la proportion de bonnes prédictions des classes de cancer des données Test

Table 2 : *Pour chaque algorithme et chaque k (nombre d'attributs utilisé pour entrainer le modèle), pourcentage de bonnes prédictions.*

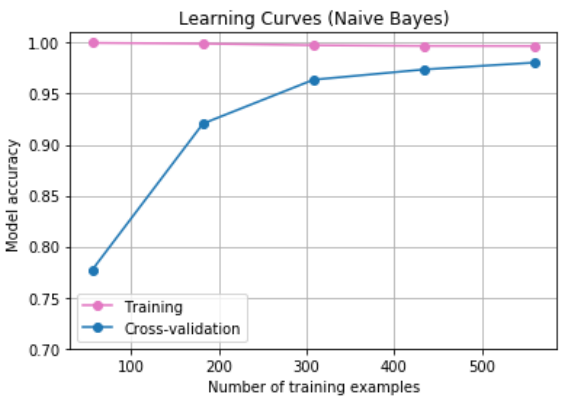
Number of features k	Naive Bayes	Decision Tree	K-nearest neighbor	Logistic regression	Random forest
10	90.46	87.14	92.53	91.29	89. 21
50	99.59	99.17	100	100	100
100	99.17	98.34	100	100	100
200	99.17	97.10	100	100	99.59
500	99.17	95.85	100	100	99.59
1000	98.76	96.68	100	100	99.59
10 000	86.31	96.68	100	100	99.59
20 531	59.75	97.51	100	100	99.59

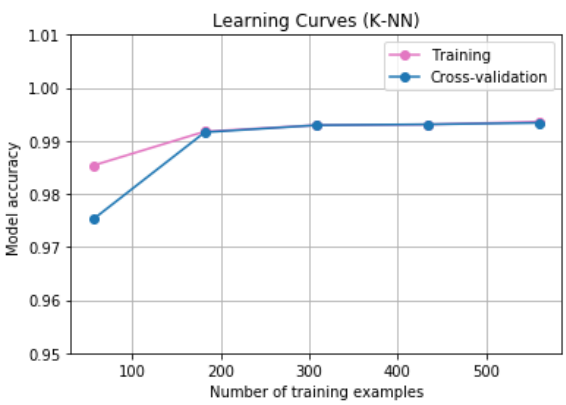
On a pu voir que certains algorithmes sont moins sensibles que d'autres au bruit causé par les gènes non informatifs notamment K-NN et la régression logistique qui ont un résultat de bonnes prédictions non parfait pour un nombre d'attributs k sélectionnés = 10, pour les autres tailles de gènes pris en comptes, les prédictions sont 100% exactes. Le modèle le plus affecté est celui de l'algorithme Naive Bayes qui ne fournit la bonne prédiction que pour 59.75% des échantillons Test lorsque la totalité (20531) des gènes est utilisée pour entrainer le modèle. Pour l'underfitting, l'algorithme le plus impacté est celui des Arbres de décision avec le plus bas taux de bonnes prédictions (87.14%). Dans les 5 cas on observe la meilleure proportion de prédictions exactes lorsque k est entre 50 et 100. C'est pourquoi pour la suite nous allons fixer ce k à 75.

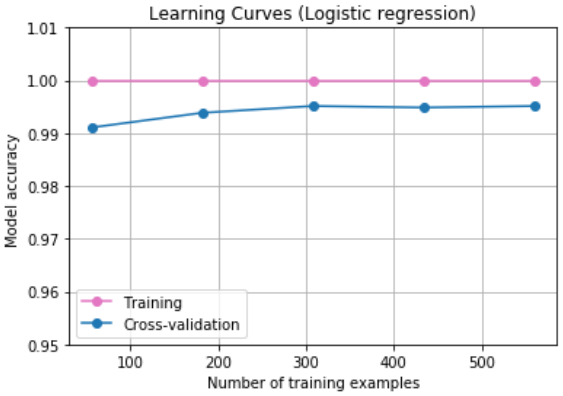
Learning curves

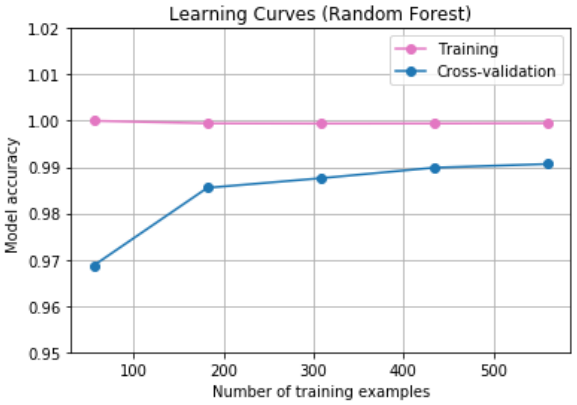
Pour évaluer les 5 modèles les courbes d'apprentissages ont été tracées. Les set Train et Test ont été formé par crossvalidation selon 100 itérations. L'apprentissage est fait en fonction de la composition et de la taille d'échantillon graduelles ; pour chaque taille la performance est évaluée en calculant la proportion de prédictions exactes.

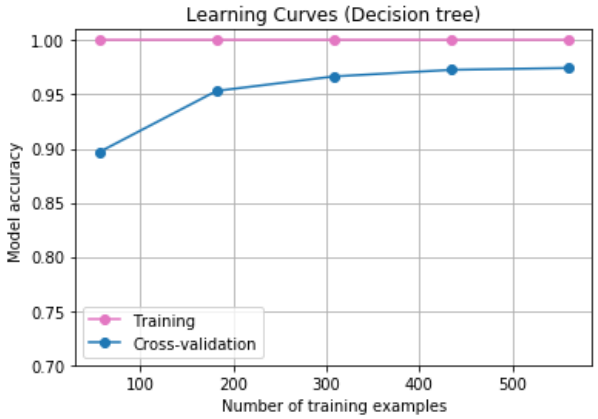
Figure 1 : *Learning curves, représentent la capacité d'apprentissage évaluer par le training score (axe des ordonnées) du modèle en fonction de l'échantillonnage (axes de abscisses).*











On peut voir pour tous les modèles que les sets de validation et d'entrainement sont relativement proche notamment pour K-NN et Logistic Regression. Pour Naive Baynes l'écart est le plus grand pour un petit échantillonnage. On voit que la prédiction est toujours au maximum (ou presque pour K-NN) pour le training set et que pour tous les modèles la qualité de prédiction du set de validation augmente avec le nombre d'échantillon d'entrainements. On note aussi que le modèle ayant la plus grande performance (0.99) sur le set de validation au minimum de la taille d'échantillon pour l'entrainement et la Regression Logistique. Basée sur les résultats précédents la création du modèle de prédiction pour les différentes classes de cancer sera basée sur un algorithme de Regression Logistique.

Création du modèle

L'algorithme de classification de Regression Logistique est utilisé pour créer le modèle de prédiction de classe de cancer à partir de 75 gènes sélectionnés précédemment. Les résultats de prédictions des données Test présentent un taux d'exactitude de 100%, soit tous les échantillons ont correctement été classés.

Table 3 : *Tableau croisé entre classe de cancer prédites et réelles*

CANCER CLASS	BRCA	COAD	KIRC	LUAD	PRAD
BRCA	100	0	0	0	0
COAD	0	25	0	0	0
KIRC	0	0	39	0	0
LUAD	0	0	0	38	0
PRAD	0	0	0	0	39

Discussion

Dans cette étude nous avons vu que les modèles de prédictions ont tous leurs avantages et limitations. Chacun des algorithmes utilisé présentent ses propres paramètres statistiques plus ou moins impacté par les données. Comme on a pu le voir leurs sensibilités aux attributs parasites et donc leur tendance à l'overfitting n'est pas la même. Sur nos données, on relève l'excellente performance de la Regression Logistique et de K-NN. De plus les courbes d'apprentissages permettent de déterminer quel modèle détient les meilleures prédictions pour le plus bas échantillonnage là encore on retrouve spécifiquement la Regression Logistique. En effet, certains algorithmes nécessitent moins d'échantillons que d'autre pour avoir une excellente performance pour les prédictions de jeu de validation. Cependant les résultats montrent qu'il n'est pas facile de déterminer quel algorithme de classification est le meilleur. La comparaison de la performance de chaque modèle à classifier les patients selon leur cancer a permis de déterminer un algorithme de classification adéquat ainsi que la redimension des données la plus informative possible. Lorsque les paramètres sont ajustés aux jeux de données les prédictions sont alors les meilleurs, c'est en effet ce que l'on a obtenu avec 100% des patients correctement classés.

En outre, les modèles de prédictions peuvent être très performants cependant une calibration de ceux-ci aux données est absolument nécessaire. Il est impératif de comprendre ses données, de les explorer afin de trouver et d'utiliser les outils susceptibles de les traiter au mieux pour fournir le plus d'informations. Cependant il faut prendre en compte que ces méthodes sont couteuses d'où la nécessité du traitement de données préalable notamment par la réduction de taille.

Dans notre cas, ces informations quantitatives sur le profil de transcription peuvent aider au diagnostic de maladies en utilisant les méthodes adéquates pour les analyser, malgré certaines limites. Grâce au *machine learning* une opportunité de développer les diagnostic et pronostic de cancer se présentent. En effet des résultats prometteurs sur les prédictions de classe de cancer à partir d'expression de gène comme ceux de cette étude révèlent le potentiel de ces méthodes.