U4: Entrepôts de données, fouille de données M2 IBM

Sujet : Analyse de données cliniques, des patients atteints d'un cancer du Poumon Sous RStudio

Réalisé Par: CHALAH Diana

Responsable UE4: JANNOT Anne-Sophie

Université PARIS 13 et PARIS DESCARTES

Année Universitaire: 2017/2018

1. Introduction :

L'imagerie médicale est l'un des principaux facteurs qui ont éclairé la science médicale et le traitement et a un grand potentiel pour guider la thérapie, car elle peut fournir une vue plus complète de la tumeur entière et elle peut être utilisée sur une base continue pour surveiller le développement et la progression de la maladie ou sa réponse à la thérapie. En outre, l'imagerie est non invasive et est déjà souvent répétée pendant le traitement dans la pratique de routine, contrairement à la génomique ou la protéomique, qui sont encore difficiles à mettre en œuvre dans la routine clinique.

Un objectif clé de l'imagerie actuelle est la «médecine personnalisée», où le traitement est de plus en plus adapté en fonction des caractéristiques spécifiques du patient et de sa maladie.

On parle dans ce texte d'une analyse radiomics de 440 caractéristiques quantifiant l'intensité, la forme et la texture de l'image tumorale, qui sont extraites de données de tomodensitométrie de 1 019 patients atteints de cancer du poumon ou de la tête et du cou.
Les données cliniques de ce texte vont nous servir à faire notre analyse de données cliniques sur les patients atteints du cancer du poumon.

(Rf :https://www.nature.com/articles/ncomms5006#f2)

2. Matériels et Méthodes :

Notre travail consiste à faire une analyse descriptives des données cliniques des patients atteints d'un cancer du poumon.
On est appelé à faire :
- Une interface interactive pour visualiser ces données cliniques, extraites depuis l'article mentionné au dessus et depuis ce lien : *«https://*

wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-»
- Faire une analyse descriptive .
- Faire une analyse uni-variée interactive permettant de visualiser les relations entre les données d'expression et les sous)types de cancer.
- Implémenter des arbre de décisions interactifs pour prédire les différents sous-types de cancer à partir des données de l'analyse uni-variée.

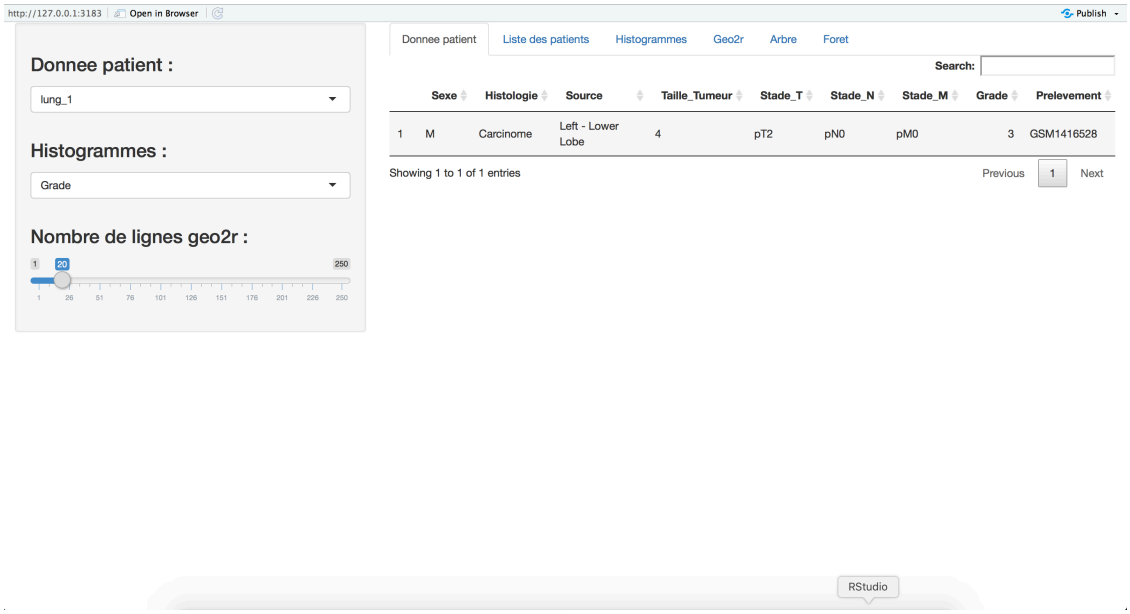
La travail se fera avec le langage **RStudio** et tous les packages associés pour faire l'analyse, ex de Shiny pour l'interface.
Geo2r pour l'analyse uni-variée (identifier les gènes qui sont exprimés différemment).

3. Résultats :

3.1 Réalisation de l'interface interactive pour la visualisation des données cliniques :

A l'aide de RStudio :
On Importe les données de Lung-cancer, faire quelques petite modifications dessus pour rendre la table plus facile à manipuler, ex des colonnes qui semblaient les plus pertinentes à l'étude, renommées, ainsi les différentes données.
J'ai choisi de faire deux groupes histologiques: **Carcinome vs Adenocarcinome.**
Récupérer les patients grâce à leur iD.
Préparer le code pour l'affichage des histogrammes(plots).
Le but est de faire un code affichant la répartition des différentes caractéristiques de l'étude.
Utilisation de l'application **Shiny** pour la partie interactive, une fois qu'on sélectionne un patient par exemple, on va appeler notre fonction avec getpatient qui nous retournera un tableau.
Cette application nous ouvre une interface interactive qui nous permet de visualiser toutes les données cliniques.
Dans cette interface on trouve un menu déroulant affichant les données patients avec la liste de tous les patients, les different histogrammes pour chaque caractéristique ex(Sexe, Garde, Histologie etc…), mais aussi la

partie geo2r pour la sélection des gènes les plus significatifs, à partir de ces gènes on va construire un arbre de décision et le random Forest . https://github.com/Diana1192/Projet_R_UE4/blob/master/Nouveau%20Script%20(Coder%2BShiny)



3.2 Analyse descriptive des données cliniques de ces patients :

Dans cette partie, il suffit d'aller sur la partie histogramme (à gauche de l'écran de l'interface), choisir quelle donnée ou caractéristique à afficher pour voir sa distribution dans l'étude, représentée en histogramme.
Si on prend l'exemple du Sexe, on voit que la hommes sont les plus atteints du cancer des poumons d'une proportion de 60% de l'étude et 29 pour les femmes.
Pour ce qui est des histologies 45% pour les carcinomes et 44 pour les adénocarcinomes.
Un autre histogramme encore intéressant, représentant les différentes valeurs des stades des tumeurs (N, M et T).

3.3 Analyse uni-variée interactive (relation entre les données d'expression et les sous types de cancer) :

Dans cette partie, j'ai utilisé **Géo2R** qui est un outil web interactif qui

permet aux utilisateurs de comparer deux ou plusieurs groupes d'échantillons dans une série GEO afin d'identifier les gènes qui sont exprimés différemment dans les conditions expérimentales. Les résultats sont présentés sous la forme d'un tableau de gènes classés par ordre de signification.
Rf : https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html#interpret

A partir de la matrice « GSE58661_series_matrix » qu'on a construit à partir de geo2r on prenant les 20 gènes les plus significatifs ayant les plus petites p valeur à laquelle s'ajoute le type histologique.

Dans l'interface shiny la partie geo2r, nous montre un tableau classé par p value ajustée, ce qui nous donne les gènes les plus significatifs en fonction de cette p valeur, on peut aussi choisir le nombre de lignes geo2r à afficher (à gauche de l'écran),.

Avec ces gènes on va pouvoir construire un arbre de décision (avec les 20 premiers gènes les plus significatifs).

3.4 Arbre de décision interactif pour prédire les différents sous-types de cancer à partir des données de l'analyse uni-variée :

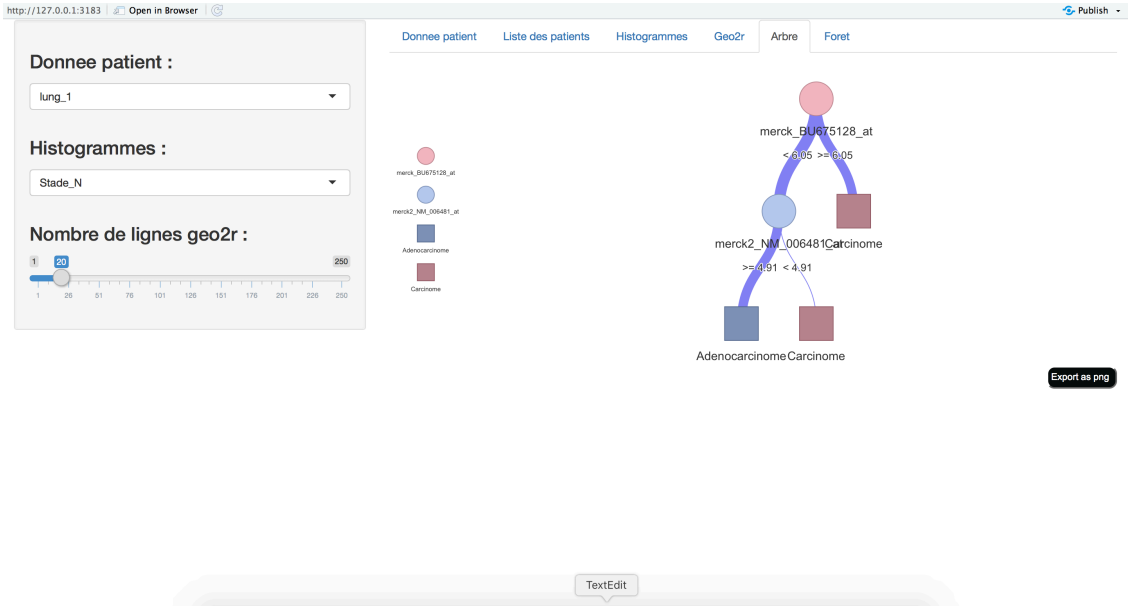
Pour réaliser les arbres de décision on doit construire 2 types d'échantillons : Apprentissage et Test.
On partant de la table ou matrice obtenu avec le code r de géo2r contenant toutes les données d'expressions génétiques la matrice à laquelle s'ajouté le type histologie pour les 2 sous types de cancers (adenocarcinome et carcinome). Le but est de définir les différents sous types de cancer à partir des données d'expression génétiques.

Libraire utilisées : rpart et randomforest.

J'ai tout d'abord commencer par charger les données, puis apporter de petite modification comme la suppression de la 1ère colonnes des identifiants.

Je suis passée en suite à la construction d'un groupes de 2 échantillons:

échantillons d'apprentissage pour l'arbre et échantillons de test pour la partie random Forest et tester la validité de l'arbre, dont 80% pour les échantillons d'apprentissage et 20% pour les échantillons test.
Le random Forest nous donne le plus petit arbre possible de tous les arbres et de rendre le modèle arbres construits plus indépendants entre eux.



L'arbre nous donne comme Racine le gène merck_BU675128_at qui représente le gène qui définit le plus les sous type de cancer avec les valeur < 6.05 et >= 6.05.
Pour les valeurs inférieures à 6.05 on trouve le gène merck_NM_006481_at à gauche il est adenocarcinome >= 4/91 Carcinome <4.91
Pour ce qui est du gène le plus représentatif il directement carcinome pour une valeur >= 6.05

Nb: L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents.