Comparaison de méthodes de réduction de dimension pour des analyses de données biologiques

Reunan Bellec & Malo Gillard
Enseignant référent : David Rousseau

25 mars 2019

Table des matières

Introduction

Complément introduction des données :

On veut analyser la germination, de la semence sèche jusqu'à la percée de la radicule. La méthode utilisée est l'imagerie automatisée sur des semences semées sur des Tables de Jacobsen (bancs de germination). Les essais de germination sont menés à 5°C dans un module climatique sur 2 tables de Jacobsen. Les sondes de température de régulation et surveillance des essais ont été étalonnées avant le début de l'expérimentation. Les 2 tables de Jacobsen ont été cartographiées pour la température en 5 points. Chaque table dispose de 4 caméras sous lesquels sont semées 600 semences par caméra par zone de 5x5 semences. Il y a 24 zones / série d'image et 25 semences par zone. Les zones et semences sont codés « ligne colonne ». 3 semis on été réalisés en parallèle sur les 2 tables dans l'ordre des populations. Les images acquises pendant 21 jours ont été analysées et à partir des mesures sur images l'heure de germination a été détectée sur chaque semence ainsi que des caractéristiques complémentaires.

Première partie

Généralités

0.1 Apprentissage supervisé

→ L'algorithme est guidé par le chercheur pour apprendre, à partir d'exemples préalablement adaptés/catégorisés → à partir de ces données, une fois l'apprentissage terminé, l'algorithme pourra réaliser des prédictions.

0.2 Apprentissage non-supervisé

→ L'algorithme est autonome, il ne connaît pas les exemples de résultats attendus en sortie → Il va devoir détecter les similarités dans les données qu'il reçoit et les organiser en fonction de ces dernières

Deuxième partie

Principes des méthodes

0.3 Groupe 1

0.3.1 Principal component analysis

→ Réduction de variables : à partir des variables corrélées, on va chercher à obtenir des variables non corrélées (composantes principales/ axes principaux) pour éliminer l'information récurrente/redondante.
→ Jeu de données créé avec une quantité d'informations décroissante contrairement à précédemment où l'information était réparti de manière uniforme
→ Données n individus observés sur p variables quantitatives. L'ACP permet d'explorer les liaisons entre variables et les ressemblances entre les individus.
→ Ensuite : visualisation des individus (notion de distances entre individus) et visualisation des variables (en fonction de leurs corrélations)

0.3.2 Isomap

n_neighbors	nombre de voisins à considérer pour chaque point
n_components	nombre de dimensions auquel on veut réduire notre espace de variables
eigen_solver	méthode de décomposition en éléments propres souhaitée (par défaut = 'a
tol ; max_iter	dépendent du choix de eigen_solver si == 'dense'
path_method	méthode utilisée pour trouver le plus court chemin/distance géodésique, pa
neighbors_algorithm	algorithme utilisée pour la recherche des plus proches voisins (par défaut =
n_jobs	entier utilisé pour indiquer quelle quantité de mémoire on souhaite utiliser

0.4 Groupe 2

0.5 Groupe 3

0.5.1 Random Forest

→ « Feature selection » : sélection des variables (les plus pertinentes)
→ Notion clé : Les arbres régularisés pénalisent les variables lorsqu'elles sont similaires aux variables déjà choisies dans les nœuds précédant. → Rappel : Les arbres binaires de décision :

→ Formation d'une série de nœuds de décision : à chaque nœud, une variable explicative est sélectionnée et une "question" est posée aux n individus, ce qui sépare l'échantillon en 2 sous-groupes en fonction de la réponse.
→ Choix de la variable explicative pour que les individus soient les plus homogènes possibles au vu de la variable à expliquer.
→ Mesure d'hétérogénéité : couramment utilisées : l'indice de diversité du Gini et l'entropie de Shannon.
→ Pour chaque variable, la "question" à poser est choisie en maximisant le gain d'homogénéité.
→ Le couple (variable,question)est choisit de façon à maximiser le gain d'homogénéité au noeud.
Etape 2 : Élagage de l'arbre
→ Un nœud devient terminal lorsque toute nouvelle séparation de l'échantillon qui lui est appliqué ne peut pas améliorer l'homogénéité déjà atteinte, on parle alors de « feuille » pour désigner ce noeud terminal.
→ Un élagage est effectué afin de réduire la complexité de l'arbre et d'éviter le sur-apprentissage (utilisation du taux d'erreurs de classement sur l'échantillon de validation)
→ Sélection de variables en utilisant Random Forest
Les forêts aléatoires peuvent être utilisées avec une problématique de sélection de variables.
→ Les forêts aléatoires fournissent deux mesures de l'importance des variables : "Mean Decrease Gini" et "Mean Decrease Accuracy"
→ Plus la perturbation d'une variable aura détérioré la précision de la prédiction, plus cette variable sera jugée importante.

Explication des paramètres :

Utilisation du « grid-search » pour le choix des paramètres

0.6 Groupe 4

n_estimators	La précision a tendance à augmenter quand le nombre d'arbres augmente, néanmoins, le temps de calcul est de plus en plus long.
max_features	Nombre de features à considérer pour la meilleure séparation d'un noeud
max_depth	profondeur maximale de l'arbre
min_samples_split	Nombre minimal d'observations pour séparer un noeud
criterion	Critère de découpe des noeuds, Gini pour la précision, Entropy pour un gain d'information
random_state	si la méthode repose sur de l'aléatoire, fixe le générateur pour reproduire les résultats.

Troisième partie

Outils utilisés

Paragraphe à compléter sur la séparation des jeux de données split ...

Quatrième partie

Réduction de dimension sur des données biologiques

0.7 Expertise

Cinquième partie

Comparaison des résultats

Méthode	Dimensions retenues/ importantes	Remarque sur la méthode
PCA	Axe 1 : 5°C TMG (j) Aire sous la courbe v15j-16j Axe 2 : v17j-18j v18j-19j v16j-17j	
Random Forest	5°C TMG (j) Aire sous la courbe	