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

Reunan Bellec & Malo Gillard
Enseignant référent : David Rousseau

6 mai 2019

Table des matières

Première partie

Introduction

Introduction du contexte

Le dérèglement climatique, le développement de nouvelles maladies des plantes, la maîtrise des rendements, amènent les états et l'ensemble des acteurs de l'agriculture en charge de la sélection variétale, à identifier des semences performantes, résistantes aux maladies, à des périodes de sécheresse ou de brusques variations environnementales durant leur développement. Ces travaux peuvent bénéficier d'avancées technologiques récentes en matière de traitements de l'information, applicables sur de larges populations de plantes. Une échelle particulièrement importante est celle de la graine, dont la qualité germinative conditionne la suite du développement de la plante. Dans ce projet annuel, nous nous intéressons à un échantillon de graines de betterave sucrière pour laquelle la France est l'un des plus gros producteurs au monde.

L'objectif de l'expérience est d'élargir la variabilité génétique de la betterave, dans le but de la rendre plus compétitive, en doublant le rythme de croissance annuelle de son rendement en sucre. Des tests ont alors été réalisés sur différents génotypes de betteraves (200 individus) pour étudier leur germination. À partir des résultats obtenus sur les semences sélectionnées, nous observons des différences, que l'on cherche à caractériser. Parmi les 200 génotypes, diverses variables sont à expliquer, telles que la surface, la longueur, la largeur, l'imbibition, la vitesse de germination par exemple.

Introduction du sujet

Le but de ce projet est de comparer des méthodes visant à diminuer la dimension des données, à réduire le nombre de variables, pour pouvoir répondre à une problématique de manière efficace. INSÉRER PARAGRAPHE SUR LES PROBLEMES GÉNÉRER PAR LES TROP GRANDES DIMENSIONS.

La problématique au centre de ce projet est de vérifier par visualisation des données que des biais d'expérimentation ne sont pas introduit dans l'étude des génotypes. Pour se faire, nous allons donc considérer plusieurs méthodes : PCA, t-SNE, Random Forest, LDA, ISOMAP, Relief.

Nous commencerons par un rappel sur les notions d'apprentissages, supervisé et non supervisé. Nous étudierons ensuite chaque méthode en détail, pour ensuite pouvoir les appliquer à notre jeu de données principal, les données issues de l'expérience de germination. Enfin, nous comparerons les résultats obtenus.

Introduction à la réduction de dimension

Le but de ce projet est de comparer des méthodes visant à diminuer la dimension des données, à réduire le nombre de variables, pour pouvoir répondre à une problématique de manière efficace. INSÉRER PARAGRAPHE SUR LES PROBLEMES GÉNÉRER PAR LES TROP GRANDES DIMENSIONS.

La problématique au centre de ce projet est de vérifier par visualisation des données que des biais d'expérimentation ne sont pas introduit dans l'étude des génotypes. Pour se faire, nous allons donc considérer plusieurs méthodes : PCA, t-SNE, Random Forest, LDA, ISOMAP, Relief.

Nous commencerons par un rappel sur les notions d'apprentissages, supervisé et non supervisé. Nous étudierons ensuite chaque méthode en détail, pour ensuite pouvoir les appliquer à notre jeu de données principal, les données issues de l'expérience de germination. Enfin, nous comparerons les résultats obtenus.

Notions élementaires

Apprentissage supervisé

Les tableaux de données sont constitués d'individus décrits par plusieurs variables dont l'une d'entre elles est qualitative. On veut fabriquer une fonction qui prédit cette variable par un modèle mathématique.

Deuxième partie

Les méthodes

0.1 Méthodes par sélection

0.1.1 Apprentissage supervisé

Random Forest

Arbre de décision

On veut construire des sous-groupes les plus homogènes du point de vue de la variable à prédire.

Mathématiques

Informatique

Explication simplifiée .

0.1.2 Apprentissage non-supervisé

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

Relief

Méthode de filtrage pour la sélection de variables explicatives. Principe : calculer une mesure globale de la pertinence des caractéristiques des données en accumulant la différence des distances entre des exemples Trouve le plus proche voisin dans la même classe et le plus proche se trouvant dans une classe différente. Identification des différences de valeurs pour les variables

Mathématiques

Informatiques

Explication simplifiée

0.2 Méthodes par transformation

Dans cette section, les méthodes traitées sont celles qui effectuent une transformation sur les variables. Une fois, ces méthodes appliquées, les individus peuvent être représenté dans un espace de dimension réduite où les axes sont nouveaux. Un nouveau système de représentation est ainsi construit.

0.2.1 Apprentissage supervisé

LDA

Mathématiques

Informatiques

Explication simplifiée

SIMCA

Mathématiques

Informatiques

Explication simplifiée

0.2.2 Apprentissage non-supervisé

PCA

Explication simplifiée

Le but de l'analyse en composante princiaple est de partir des variables corrélées, pour obtenir des variables non corrélées (composantes principales/ axes principaux) et ainsi éliminer l'information récurrente/redondante. Les variables utilisées sont donc toutes quantitatives. Les nouvelles composantes synthétisant l'information seront rangées par ordre croissant de leur contenu en information. L'information considérée ici est l'inertie. La corrélatio, mesure de liaison entre deux variablees, est un point clé de cette méthode.

On peut savoir l'importance des axes par leur pourcentage de l'nertie associé. La notion de réduction de dimension intervient lors du choix du nombre de composante pour obtenir la réprésentation la plus fidèle possible.

Lorsque les données sont ordonnées dans une matrice X, avec les individus en lignes et les variables en colonnes, l'étude des lignes portera sur les distances entre individus et l'étude des colonnes sur les relations entre variables.

Mathématiques La solution mathématique de ces études revient à trouver une matrice de rang donné la plus proche de celle-ci. Le but de cette méthode est de projeter orhtogonalement le nuage de points dans un sous-espace de dimension réduite avec une réprésentation fidèle des distances initiales. On va alors décomposer la matrice X par SVD : $X = U\Delta V = \sum_i s_i u_i v_i^t$ Ainsi, le sous espace $Vect(v_1, ..., v_s)$ est le sous espace de dimension s optimisant l'inertie projetée. L'inertie de ce sous-espace est $s_1^2 + ... + s_s^2$, la variance de Fi est s_i^2. Les projections sur v_s, $F_s = XQu_s$ correspondent aux composantes principales. Les nouvelles variables définies par les colonnes de $F_L = XQV$ sont non corrélées. Où Q EST LA MÉTRIQUE.

Utilisation pratique/informatique Paramètres à régler

Sorties

Qualité de représentation globale : partie d'intertie expliquée par l'axe de projection. L'inertie expliquée par les k premiers axes factoriels :
$\frac{\sum_{s=1}^{k} \lambda_s}{\sum_{s=1}^{k} \lambda_s}$

Qualité de représentation ou cos^2, d'un individu ou d'une variable suivant un axe ou un plan sa part d'inertie projetée, égale au $cos^2 = \frac{F_s^2(i)}{\sum_{s=1}^{p} F_s^2(k)}$

Isomap

Mathématiques

Informatiques

Explication simplifiée

T-SNE

Schéma illustratif à ajouter. Méthode non-linéaire. Théorie de l'information. Distribution de probabilité. Divergence KL. Attraction des points vers les points qui lui sont proches dans l'espace de grande dimension, inversement les points qui sont éloignés dans l'espace d'origine se rejettent. Déplacement des points petits à petits, individuellement.

Mathématiques

Informatiques perplexity
early exaggeration
learning rate
metric : Choix de la métrique utilisée pour calculer la distance entre individus

Explication simplifiée

Troisième partie

Données

0.3 Introduction des données

Dans le but d'analyser la germination, de la semence sèche jusqu'à la percée de la radicule, on utilise 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.

0.4 Expertise

La première tâche à réaliser avant de travailler sur les données au travers des différentes méthodes de réduction de dimension est l'expertise, c'est à dire la préparation, le nettoyage des données. Nous avons réfléchis avec un point de vue autre que celui de data scientist : qu'est-il bon de garder ou d'écarter dans le jeu de données, faut-il rajouter des variables explicatives, et pourquoi.

Comme nous nous intéressons à l'évolution des germinations jour par jour, nous avons écarté les variables qui s'exprimaient en heures, c'est à dire les variables **5°C TMG (h)** et **5°C T50 (h)**, et gardé celles qui s'exprimaient en jours, **5°C TMG (j)** et **5°C T50 (j)**.
Cependant, cette dernière comprenait un nombre important d'entrées "Nan" (plus de la moitié), ce qui signifie qu'il n'a pas été possible de calculer le délai nécessaire pour obtenir 50% de germination pour plus de la moitié des individus. Nous l'avons donc également écartée car elle présentait peu d'intérêt dans notre étude.

Nous avons ensuite rajouté 6 variables **'v15-16j'**, **'v16-17j'**, **'v17-18j'**, **'v18-19j'**, **'v19-20j'**, et **'v20-21j'**, pour exprimer la vitesse de germination de chaque individu jour après jour à partir du jour 15.

Quatrième partie

Explication des paramètres

Explication des paramètres :

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.

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

Cinquième partie

Résultats

Sixième partie

Comparaison des méthodes

Les variables les plus importantes selon les méthodes :

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	
Relief	5°C TMG (j) Aire sous la courbe	
Random Forest	5°C TMG (j) Aire sous la courbe	

Note :

Pour les méthodes par sélection de variables, les variables les plus importantes sont tout simplement celles sélectionnées, tandis que pour les méthodes par transformation il est nécessaire de regarder les contributions des anciennes variables aux nouvelles.

Septième partie

Conclusion