Expliquer la classification d'expression de gènes par la méthode des gradients intégrés

Myriam BONTONOU[1,2], Jean-Michel ARBONA[1], Benjamin AUDIT[2], Pierre BORGNAT[2]

Univ Lyon, ENS de Lyon, CNRS, LBMC[1], Laboratoire de physique[2], Lyon, France

GraphNEx Graph Neural Networks for Explainable Artificial Intelligence

CHIST-ERA-19-XAI-006
for GRAPHNEX ANR-21-CHR4-000

Contexte

- **Modèles** entraînés à classifier des états biologiques à partir de données moléculaires. *Ex : classification d'un type de tumeur à partir de l'expression des gènes.*
- Prise de décision des modèles fondée sur certaines variables et sur certaines interactions entre ces variables. *Ex : sur-expression simultanée de certains gènes.*
- **Variables classées en fonction de leur contribution aux décisions d'un modèle** par des **méthodes d'explicabilité**.
- Variables en tête des classements interprétées comme étant **informatives pour la tâche de classification**.

Question : à quel point les variables mises en avant par les méthodes d'explicabilité sont pertinentes pour le modèle ?

GitHub

Contributions

- **Métriques** quantifiant la **pertinence des variables mises en avant**.
  ⟶ Variables suffisantes / nécessaires pour le modèle ?
- Expériences sur des **jeux de données d'expression des gènes réels** et **simulés** avec des variables discriminantes connues.

PyTorch code `https://github.com/mbonto/XAI_for_genomics`.

Cadre méthodologique

- **Échantillon** $\mathbf{x} \in \mathbb{R}^F$ de classe $c \in [\![1\ ;\ C]\!]$.
  Échantillon de référence $\mathbf{x}' \in \mathbb{R}^F$.
- **Modèle supervisé** $f : \mathbb{R}^F \mapsto \mathbb{R}^C$. *Ex: réseau de neurones.*
- **Méthode d'explicabilité** attribuant à chaque variable i de l'échantillon $\mathbf{x}$ un score $\phi_i(\mathbf{x})$ représentant son importance pour le modèle.
  Ici, méthode des gradients intégrés [3].

$$\phi_i(\mathbf{x}) = (\mathbf{x}_i - \mathbf{x}'_i) \int_{\alpha=0}^{1} \left. \frac{\partial f_c(\mathbf{z})}{\partial \mathbf{x}_i} \right|_{\mathbf{z}=\mathbf{x}'+\alpha(\mathbf{x}-\mathbf{x}')} d\alpha\ .$$

- **Classement des variables** de l'échantillon $\mathbf{x}$.
  - **Local** selon les $\phi_i(\mathbf{x})$.
  - **Global** selon les ϕ_i^c. Soit M le nombre d'exemples m de classe c utilisés pour entraîner le modèle et la norme Euclidienne $\|\cdot\|$,

$$\phi_i^c = \frac{1}{M} \sum_{m=1}^{M} \frac{\phi_i(\mathbf{x}^m)}{\|[\phi_1(\mathbf{x}^m), \ldots, \phi_F(\mathbf{x}^m)]\|}$$

Métriques pour évaluer la pertinence des classements

Mesurer la pertinence des variables pour un modèle

- Masquage des variables contribuant le moins aux décisions sans perturber le modèle ?
- Nécessité des variables contribuant le plus aux décisions ?

Variation des prédictions lors du masquage des variables ?

- Masquer une variable : remplacer $\mathbf{x}_i$ par $\mathbf{x}'_i$.
- Masquage progressif des variables de $\mathbf{x}$ en suivant un ordre sur le classement ϕ ou ϕ^c.
  - **PGU** scores croissants (moins importants en premier).
  - **PGI** scores décroissants (plus importants en premier).
- Exemple $\mathbf{x}$ avec f variables masquées : $\tilde{\mathbf{x}}_f$.

$$\textbf{Prediction gap PG} = \sum_{f=1}^{F} \frac{\max(f_c(\mathbf{x}) - f_c(\tilde{\mathbf{x}}_f), 0)}{F}\ .$$

Apparition des variables discriminantes connues en tête des classements ?

- Ensemble $\mathcal{E}_r$ de variables informatives (disponible pour des données simulées).
- Ensemble $\mathcal{E}_i$ contenant les $|\mathcal{E}_r|$ variables les plus importantes identifiées.

$$\textbf{Feature agreement FA} = \frac{|\mathcal{E}_r \cap \mathcal{E}_i|}{|\mathcal{E}_r|}.$$

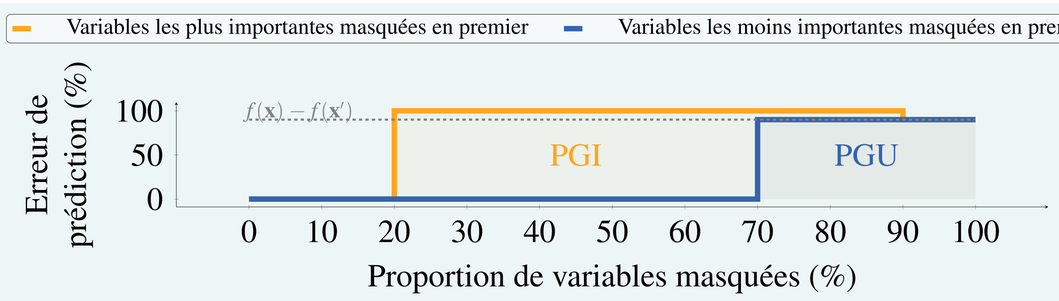


Figure 1 - Schéma décrivant les métriques PGI and PGU.

Cadre expérimental

- Données d'expression des gènes avec 1 classe saine et 1 classe tumeur [1].
  - BRCA / KIRC : 1210 / 606 exemples, 58274 / 58233 variables.

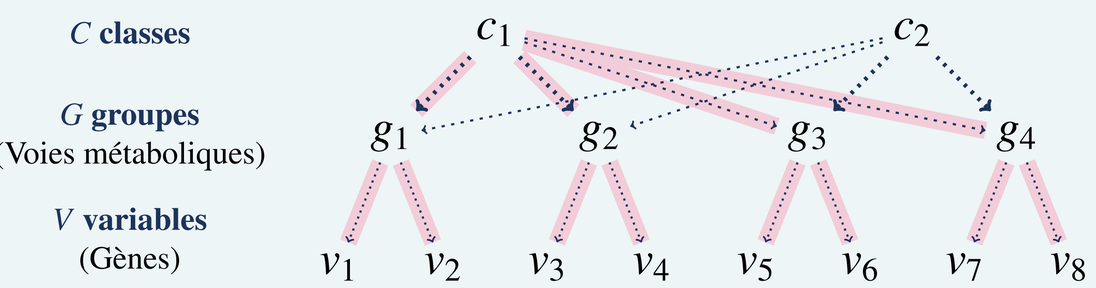
  → Référence $\mathbf{x}'$: moyenne des exemples sains utilisés pendant l'entraînement.

  → Classe à expliquer : classe tumeur.
- Simulations avec 1 classe homogène et 1 classe plus ou moins hétérogène (Fig. 2).
  - SimuA / B / C : 1200 exemples, 50000 variables, classe hétérogène avec 1 / 3 / 5 sous-classes.
  - 500 variables informatives pour la classe 1 et pour les sous-classes de la classe 2.

  → Référence $\mathbf{x}'$: moyenne des exemples d'entraînement de la classe homogène.

  → Classe à expliquer : classe hétérogène.
- Modèle supervisé : perceptron avec 1 couche cachée.

Jeux de données simulés



A priori sur la proportion des groupes au sein d'une classe c $\boldsymbol{\alpha}_c \in \mathbb{R}_+^G$.
A priori sur la proportion des variables au sein d'un groupe g $\boldsymbol{\eta}_g \in \mathbb{R}_+^V$.
Proportion des variables au sein d'un groupe g $\boldsymbol{\beta}_g \sim \text{Dirichlet}(\boldsymbol{\eta}_g)$.
Proportion des groupes dans un exemple s de classe c_1 $\boldsymbol{\theta}_s \sim \text{Dirichlet}(\boldsymbol{\alpha}_{c_1})$.

Figure 2 - Modèle LDA utilisé pour simuler des données [2]. Le séquençage d'ARN mesure l'expression des gènes dans un tissu. Chaque gène i est caractérisé par un nombre de reads séquencés v_i. Ce nombre dépend de la voie métabolique (groupe g) auquel le gène appartient et de l'utilisation du groupe dans la classe considérée c. Un exemple s de classe c_1 est généré en tirant un grand nombre de reads (variables) à partir des deux lois multinomiales associées à l'exemple : 1) $g \sim \text{Multinomiale}(\boldsymbol{\theta}_s)$, 2) $v \sim \text{Multinomiale}(\boldsymbol{\beta}_g)$.

Résultats

Table - Métriques obtenues sur des données réelles (BRCA, KIRC) et des données simulées (SimuA, SimuB, SimuC).

Dataset	BRCA	KIRC	SimuA	SimuB	SimuC
Exactitude équilibrée (↑)	99.6 ± 0.1	99.6 ± 0.7	100	100	100
Classement local ϕ					
PGI (↑)	98.7 ± 0.3	98.4 ± 0.5	94.9 ± 0.2	97.4 ± 0.2	97.4 ± 0.2
PGU (↓)	0.9 ± 0.2	1.0 ± 0.2	5.5 ± 0.2	6.7 ± 0.3	5.7 ± 0.2
FA (↑)	-	-	79.9 ± 0.3	69.1 ± 0.6	67.8 ± 0.7
Classement global ϕ^c					
PGI (↑)	98.2 ± 0.3	98.0 ± 0.5	92.1 ± 0.3	96.4 ± 0.2	96.3 ± 0.3
PGU (↓)	1.6 ± 0.3	1.3 ± 0.3	8.3 ± 0.3	9.3 ± 0.6	7.7 ± 0.2
FA (↑)	-	-	100.0 ± 0.1	94.6 ± 0.5	90.5 ± 0.7

(a) Échantillons $\mathbf{x}$.

(b) Scores $\phi(\mathbf{x})$ sur la classe tumeur.

Figure 3 - Visualisations t-SNE du jeu de données BRCA.

Résultats sur BRCA : recherche des variables pertinentes pour identifier une tumeur.

- Par échantillon, 524 variables sont suffisantes. Mais pas nécessaires.

  → *Garder les 524 variables les plus importantes permet de préserver la décision. En garder moins ne le permet pas (PGU local).*

  → *Prédictions dégradées après le masquage de 757 variables importantes (PGI local). Masquer les 524 identifiées ne perturbe donc pas le modèle.*
- Pour l'ensemble des échantillons, 932 variables sont suffisantes (PGU global). Mais pas nécessaires (PGI global).
- Hétérogénéité non détectée par les PGs globaux.

  → *Deux groupes de la classe tumeur (Fig. 3a) observés dans les scores (Fig. 3b).*

  → *PGs locaux et globaux relativement similaires.*

Conclusion

- Évaluation de la **complexité de deux jeux de données d'expression des gènes du point de vue d'un modèle d'apprentissage supervisé**.
- **Analyse** de la pertinence des classements **sur des données simulées hétérogènes** (FA).
- **Définition ambiguë** d'un ensemble de variables informatives pour un modèle donné.

[1] `https://portal.gdc.cancer.gov/`.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[3] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*. PMLR, 2017.