Segmentation d'images hyperspectrales par Information Bottleneck

Noémie LAFOUGE[1, 2], Alban GOUPIL[1], Eric PERRIN[1], Valeriu VRABIE[1], Brigitte CHABBERT[2] et Sylvie RECOUS[2] *

[1]Université de Reims Champagne Ardenne, CReSTIC, EA 3804, 51097 Reims, France

[2]Université de Reims Champagne Ardenne, INRAE, FARE, UMR A 614, 51097 Reims, France

noemie.lafouge@univ-reims.fr

Résumé – Compresser une variable aléatoire X pour obtenir une variable Z qui reste pertinente en regard d'une troisième variable d'intérêt Y est l'objectif de la méthode de l'Information Bottleneck. Nous proposons une procédure permettant de construire la variable Z par une mise en cascade d'encodeurs déterministes lorsque X est discrète. Cette solution améliore une approche antérieure et permet une segmentation non supervisée d'image hyperspectrale.

Abstract – The Information Bottleneck method aims to compress a random variable X into a variable Z that keeps relevant to a third variable Y. When X is discrete, we propose an algorithm that builds such a Z variable by cascading deterministic encoders. This solution improves a previous algorithm from the literature and is applied to segment hyperspectral images.

1 Introduction

La réduction de dimension offre un cadre pour lutter contre le fléau de la dimension en traitement du signal et en apprentissage automatique et ainsi autoriser un traitement plus rapide et souvent plus efficace. Il s'agît de trouver un modèle *ad-hoc* simple avec un nombre limité de variables explicatives qui décrit correctement les données. La composante des données qui ne s'inscrit pas dans le modèle est considérée comme du bruit et s'en trouve filtrée.

Cependant, modéliser ce qui relève du signal et ce qui relève du bruit se révèle souvent complexe et demande une grande expertise pour être intégré dans une opération de réduction de dimension. La méthode de l'Information Bottleneck de Tishby *et al.* [1] propose pour séparer le signal du bruit d'utiliser une variable auxiliaire connue Y et de compresser les données X en une variable Z qui reste la plus pertinente possible vis-à-vis de Y. Ainsi, la réduction des données X se réalise par leur version compressée Z et la pertinence de celle-ci est assurée par l'information partagée avec Y.

Nous proposons dans ce papier d'utiliser la méthode de l'Information Bottleneck, pour classer de façon non supervisée des hyperspectres de résidus de culture au sol par un processus de regroupement hiérarchique ascendante. Notre méthode est similaire à celle de [2] pour les galaxies mais, hormis le domaine d'application, nous modifions le critère d'optimisation qui améliore la qualité du classifieur selon l'objectif de l'Information Bottleneck.

*Les auteurs remercient la Fondation Paris-Reims, pour l'allocation doctorale à Noémie Lafouge, INRAE pour le soutien financier via le projet "pari scientifique" RESAERO, et l'URCA pour le financement de l'équipement via la plateforme EXPERI. Pascal Thiébeau et Gonzague Alavoine pour l'appui technique aux expérimentations.

La prochaine section introduit l'Information Bottleneck et les bornes théoriques apportées par la théorie de l'information. La section suivante introduit une mise en cascade d'encodeurs et explicite le regroupement hiérarchique qui en découle. Elle se termine avec notre modification de critère. À la pénultième section, cette nouvelle méthode est appliquée aux hyperspectres issus d'une image de résidus de culture pour en faire une segmentation non supervisée et les résultats sont comparés à la méthode de l'agglomerative Information Bottleneck.

2 Information Bottleneck

Pour simplifier le discours, nous supposerons par la suite que toutes les variables considérées sont discrètes, mais l'extension au cas continu reste possible en paramétrant les probabilités idoines. Considérons donc un couple de variables aléatoires (X, Y) de probabilité $p(x, y)$. En utilisant un encodeur sous la forme d'une probabilité conditionnelle $q(z|x)$, nous pouvons introduire une variable Z qui sera une version compressée de X mais qui doit conserver au maximum l'information que X partage avec Y. Le choix de la lettre q indique que cette distribution est libre et sera optimisée par la suite.

Avec ce modèle, les trois variables X, Y et Z sont reliées par la propriété de Markov $Y - X - Z$ et la probabilité jointe du triplet est donnée par $p(x, y, z) = p(x, y)\ q(z|x)$. À partir de celle-ci, toutes les probabilités concernant les variables X, Y ou Z ou leurs couples sont calculables.

Dans un problème de classification, Y modéliserait les étiquettes, X les observations dépendantes de la classe, et Z, à travers la probabilité $q(z|x)$ serait le classifieur. Dans l'application qui nous importe ici, Y sont les hyperspectres d'une image hyperspectrale, X sont les pixels et Z regroupe les pixels en

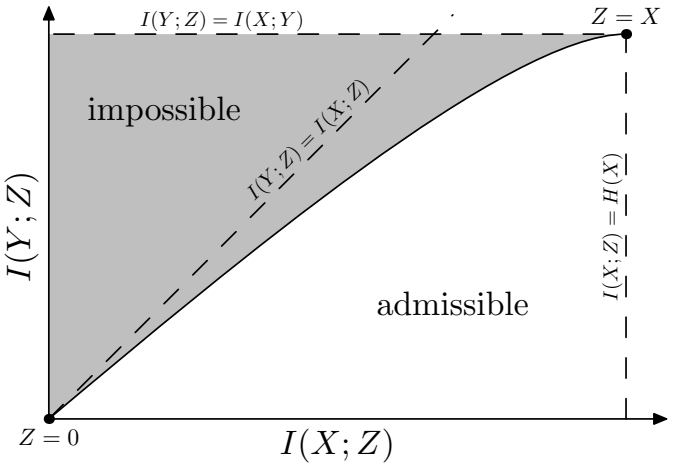

FIGURE 1 – Plan d'information

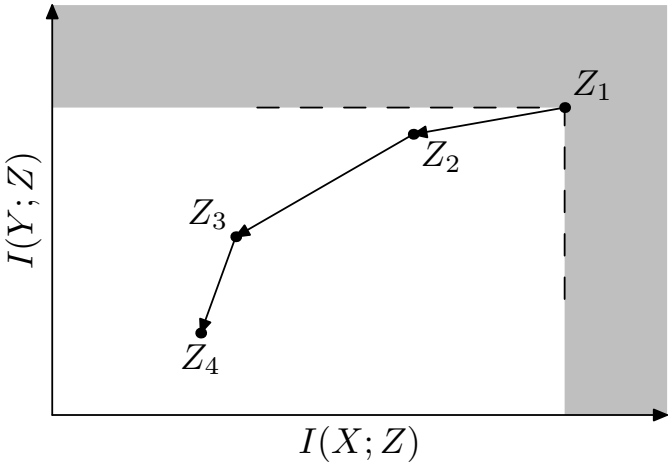

FIGURE 2 – Mise en cascade

classes dont les spectres se ressemblent. Le nombre de classes en sortie serait alors le nombre de modalités chargées de Z.

Graphiquement, à chaque variable Z est associée un point dans le plan d'information d'abscisse $I(X;Z)$ qui mesure le niveau de compression et d'ordonnée $I(Y;Z)$ qui mesure la pertinence de Z à représenter Y. La figure 1 représente un tel plan d'information.

Le caractère Markovien, $Y - X - Z$, associé à l'inégalité du traitement des données permet de limiter l'information mutuelle entre Z et X ou Y. C'est-à-dire qu'en notant I l'information mutuelle et H l'entropie de Shannon, nous avons

$$0 \leq I(Y;Z) \leq I(X;Y), \tag{1}$$
$$0 \leq I(Y;Z) \leq I(X;Z) \leq H(X). \tag{2}$$

Ces différentes inégalités qui limitent la région admissible des points associés aux variables Z dans le plan d'information sont représentées en lignes tiretées sur la figure 1. À l'instar de la théorie de la distorsion, dans laquelle l'Information Bottleneck peut s'écrire [3], il existe aussi une zone inatteignable, grisée sur la figure, limitée par la courbe des Z paramétrée par β vérifiant

$$q^*_\beta(z|x) = \underset{q(z|x)}{\arg\max}\, I(X;Z) - \beta\, I(Y;Z). \tag{3}$$

Deux solutions extrêmes, notées sur la figure, sont envisageables, la première, pour $\beta = 0$ est donnée par $Z = X$, c'est-à-dire que Z est une copie de X, et la seconde, pour $\beta \to \infty$ est $Z = 0$, c'est-à-dire que Z est constante et ne contient aucune information sur X et encore moins sur Y.

Naturellement, l'objectif est de trouver une variable Z qui s'approche au maximum du sommet en haut à gauche de compression $I(X;Z)$ maximale et de pertinence $I(Y;Z)$ maximale tout en restant dans la zone admissible. Initialement, les auteurs de [1] proposent une recherche de solution de (3) par une méthode itérative de recherche de point fixe qui n'en garantit toutefois pas l'optimalité.

3 Regroupement hiérarchique

La méthode que nous proposons ne repose pas sur une recherche de point fixe. Elle repose sur une mise en cascade d'encodeurs déterministes. Ainsi, contrairement à la méthode précédente d'optimisation de (3), la construction est déterministe, le résultat final unique, et le nombre de modalité de la version compressée constitue le seul paramètre à fixer pour l'utilisateur qui est plus naturel à manipuler que le paramètre β.

La mise en cascade consiste à créer une série de variables $Z_1, Z_2, Z_3, \ldots$ qui compressent graduellement X. Cette succession s'écrit comme une chaîne de Markov $Y - X - Z_1 - Z_2 - Z_3 - \cdots$. L'inégalité du traitement de l'information implique donc que $I(Y;Z_k) \leq I(Y;Z_{k-1})$ et $I(X;Z_k) \leq I(X;Z_{k-1})$, c'est-à-dire que l'ajout d'une variable compresse d'avantage X mais au détriment de sa pertinence vis-à-vis de Y. Dans le plan d'information, cela revient à limiter les séquences de variables possibles comme l'indique la figure 2 où Z_2 ne peut atteindre la zone grisée.

Notre méthode débute par la variable $Z_0 = X$ dont le point dans le plan d'information est en haut à droite puis construit à partir de celle-ci une séquence de variables $Z_1, Z_2, Z_3, \ldots$ Cette construction ne peut pas revenir en arrière : une fois la variable Z_k fixée, la variable Z_{k+1} est choisie selon un certain critère. Nous proposons dans cet article de rendre la pente dans le plan d'information entre Z_{k+1} et Z_k la plus horizontale possible, autrement dit

$$q^*(z_{k+1}|z_k) = \underset{q(z_{k+1}|z_k)}{\arg\min} \frac{I(Y;Z_{k+1}) - I(Y;Z_k)}{I(X;Z_{k+1}) - I(X;Z_k)}. \tag{4}$$

La recherche des encodeurs selon le critère ci-dessus est simplifiée lorsqu'ils sont déterministes et que les variables sont discrètes. En effet, un tel encodeur s'écrit $z = f(x)$ qui devient $q(z|x) = \big[z = f(x)\big]$ où $[P]$ est le crochet d'Iverson qui vaut 1 si la proposition P est vérifiée et 0 sinon. Une fonction f induit un partitionnement de l'espace de départ en cellules $f^{-1}(z_i) = \{x | f(x) = z_i\}$. Par exemple sur la figure 3, en usant abusivement des notations, la fonction f_0 regroupe x_1 et

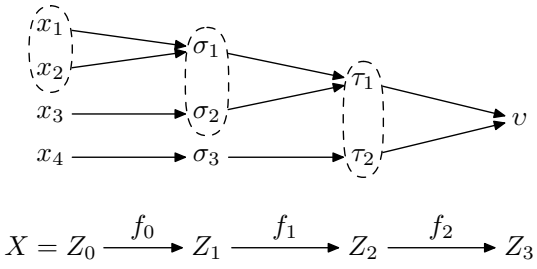

FIGURE 3 – Regroupement hiérarchique

x_2 en une cellule notée σ_1. De même f_1 regroupe les cellules σ_1 et σ_2 en une cellule τ_1.

Un regroupement hiérarchique se construit donc en enchaînant les fonctions comme sur la figure 3 : x_1 et x_2 sont regroupés ensemble puis la cellule résultante σ_1 est regroupée avec x_3 et enfin x_4 rejoint l'ensemble υ. Le regroupement se fait ici via trois encodeurs déterministes. En conclusion, le choix des variables Z_n données par les encodeurs déterministes est en correspondance avec le choix d'une suite de partitions de plus en plus grossière de l'ensemble des modalités de X.

À chaque cellule σ sont associées les probabilités $p(\sigma)$ et $p(y|\sigma)$ qui dépendent directement de $p(x, y)$ et des modalités de X regroupées dans la cellule σ. Lorsque deux cellules σ_1 et σ_2 sont regroupées en une seule cellule σ^*, ces deux probabilités se calculent selon

$$p(\sigma^*) = p(\sigma_1) + p(\sigma_2), \tag{5}$$

$$p(y|\sigma^*) = \frac{p(\sigma_1)\, p(y|\sigma_1) + p(\sigma_2)\, p(y|\sigma_2)}{p(\sigma_1) + p(\sigma_2)}. \tag{6}$$

L'algorithme que nous proposons commence par une partition initiale π_0 dans laquelle chaque modalité de X est un singleton. Tant que le nombre de cellules de la partition courante π_n est supérieur à un seuil, on regroupe deux cellules σ_1 et σ_2 bien choisies en une cellule σ^*. En pseudo-code, cette procédure s'écrit

$\pi_0 \leftarrow \big\{\{x\}$ pour chaque modalité x de $X\big\}$;
while $|\pi_n| >$ *nombre de classes visé* **do**
  Choisir 2 cellules σ_1 et σ_2 de π_n à fusionner ;
  $\sigma^* \leftarrow \sigma_1 \cup \sigma_2$;
  $\pi_{n+1} \leftarrow \pi_n \backslash \{\sigma_1, \sigma_2\} \cup \{\sigma^*\}$;
  Calculer $p(\sigma^*)$ et $p(y|\sigma^*)$ avec (5) et (6) ;

Reste maintenant à trouver les cellules σ_1 et σ_2 à fusionner. La solution proposée dans [2], appelée aIB pour agglomerative Information Bottleneck, consiste à limiter au maximum l'écart δI entre $I(Z_k; Y)$ et $I(Z_{k+1}; Y)$ où Z_k correspond à la partition où σ_1 et σ_2 ne sont pas regroupées et Z_{k+1} est la même partition mais avec ces deux cellules regroupées. Cette quantité s'exprime aussi par

$$\begin{aligned}\delta I = I(Z_k; Y) - I(Z_{k+1}; Y) &= H(Y|Z_{k+1}) - H(Y|Z_k) \\ &= p(\sigma^*)H(Y|\sigma^*) - p(\sigma_1)H(Y|\sigma_1) - p(\sigma_2)H(Y|\sigma_2),\end{aligned}$$

dont les entropies se calculent grâce $p(y|\sigma)$ avec les expressions (5) et (6).

Comme évoqué précédemment, nous proposons de modifier ce critère. En effet, ce dernier mesure la descente verticale entre Z_k et Z_{k+1} dans le plan d'information dans la figure 2. Mais l'objectif n'est pas de limiter cette descente mais plutôt de se rapprocher au maximum du point en haut à gauche de coordonnée $\big(0, I(X;Y)\big)$. Notre critère consiste donc à limiter la pente $\delta I'$ entre les points Z_k et Z_{k+1} dans le plan d'information, c'est-à-dire en réexprimant (4), en minimisant

$$\begin{aligned}\delta I' = \frac{I(Z_k; Y) - I(Z_{k+1}; Y)}{I(Z_k; X) - I(Z_{k+1}; X)} &= \frac{H(Y|Z_{k+1}) - H(Y|Z_k)}{H(Z_k) - H(Z_{k+1})} \\ &= \frac{p(\sigma^*)H(Y|\sigma^*) - p(\sigma_1)H(Y|\sigma_1) - p(\sigma_2)H(Y|\sigma_2)}{p(\sigma^*)\log p(\sigma^*) - p(\sigma_1)\log p(\sigma_1) - p(\sigma_2)\log p(\sigma_2)}.\end{aligned}$$

où le déterminisme des encodeurs rend les entropies conditionnelles $H(Z_n|X)$ nulles et justifie le dénominateur de la seconde égalité.

Intuitivement, la convexité de la région admissible dans la figure 1 motive aussi l'approche proposée car elle implique que la pente de sa frontière décroît entre les points associés à $Z = 0$ et à $Z = X$.

4 Segmentation d'image hyperspectrale

Notre méthode et celle de l'aIB ont été appliquées à une image hyperspectrale de résidus de culture sur un fond bleu de taille 100×100. À chaque pixel est associé un spectre de 300 bandes entre $380\,\mathrm{nm}$ et $1015\,\mathrm{nm}$. L'objectif est de séparer rapidement, sans supervision, le fond de la paille pour construire ensuite une base de données de spectres de résidus.

L'ensemble des hyperspectres est réorganisé en un vecteur de $10\,000$ spectres, soit un tableau $p(x, \lambda)$ de taille $10\,000 \times 300$. Ce tableau est normalisé pour que la somme de ses éléments fasse 1. Ainsi, à l'instar de [2], $p(x, \lambda)$ représente la probabilité du couple de variables (X, Λ) qu'un photon repéré par la caméra soit de longueur d'onde λ et localisé à la position x.

Nous utilisons la procédure de regroupement hiérarchique décrite auparavant pour construire la mise en cascade $\Lambda - X - Z_1 - Z_2 - \cdots$ selon les deux critères de choix de fusion de cellule δI et $\delta I'$. L'objectif est de parvenir à une variable Z_n binaire c'est-à-dire un partitionnement des spectres en deux classes, et par conséquent une segmentation de l'image en deux segments de telle sorte que les pixels soient regroupés selon leur spectre.

La figure 4 affiche les points associés aux variables Z_k lors de la construction. En bleu notre solution basée sur $\delta I'$, en vert celle de l'aIB basée sur δI. La trajectoire bleue étant au dessus de la verte, notre critère permet de regrouper de façon plus pertinente les pixels selon l'information vis-à-vis de leur spectre. Il permet d'éviter des choix de regroupement trop rapides au début de l'algorithme qui limitent les cellules à fusionner en fin d'exécution.

Les variables finales, marquées en rouge sur cette même figure diffèrent largement quant à la pertinence par rapport à la variable Λ. En effet, notre méthode obtient une valeur de $I(Z; Y)$ plus de cinq fois supérieure à la méthode de l'aIB. De

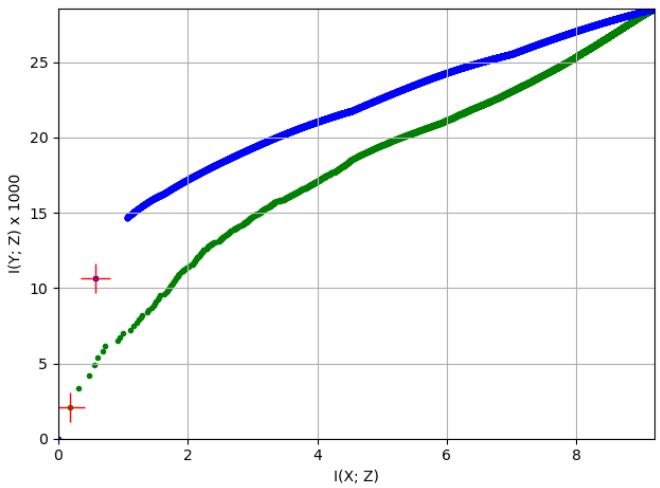

FIGURE 4 – Trajectoire dans le plan d'information. Notre solution en bleu, l'aIB en vert. Classifications binaires marquées par une croix rouge.

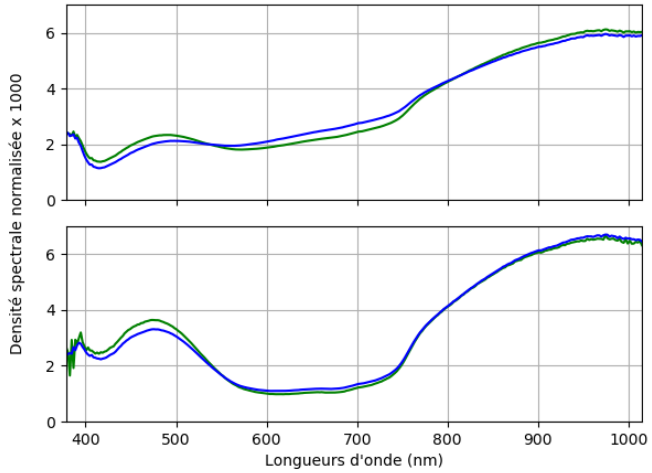

FIGURE 5 – Spectres moyens des deux classes. Notre solution en bleu, l'aIB en vert.

plus, la variable issue de notre méthode profite pleinement de son support binaire alors que la variable issue de l'aIB, avec son faible $I(Z;X)$ indique une mauvaise balance des poids entre les classes.

Au final, chacune des méthodes classe les pixels en deux groupes σ_1 et σ_2 qui définissent les probabilités $p(\lambda|\sigma_i)$ qui sont l'équivalent d'un spectre normalisé moyen. Ces spectres pour chacune des classes sont représentés sur la figure 5. Les différences entre eux se situent principalement sur la bande $420\,\mathrm{nm}$-$500\,\mathrm{nm}$ et $600\,\mathrm{nm}$-$720\,\mathrm{nm}$. Les spectres moyens de l'aIB présentent aussi des variations plus importantes en bord de bandes, avant $400\,\mathrm{nm}$ et après $960\,\mathrm{nm}$, où le bruit est plus puissant. Notre solution s'adapte donc mieux aux variations intra-classes.

Le processus de regroupement des pixels en cellule construit

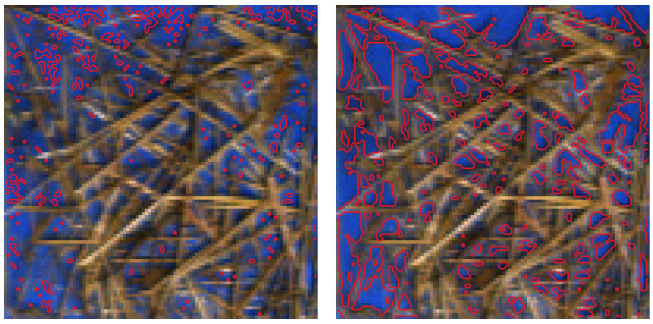

FIGURE 6 – Segmentations selon les méthodes. Notre solution à droite, l'aIB à gauche.

aussi une segmentation de l'image : à chaque cellule finale, les pixels constitutifs forment les segments. Les contours des segments ont été superposés à l'image initiale sur la figure 6. Cette image couleur est reconstruite à partir des hyperspectres en extrayant pour chaque pixel la valeurs des spectres aux longueurs d'ondes les plus proches de $611.3\,\mathrm{nm}$, $549.2\,\mathrm{nm}$ et $464.3\,\mathrm{nm}$ associées aux couleurs rouge, verte et bleue. Ces trois canaux forment une image codée en RVB.

Notre solution, en bas de la figure, indique une segmentation de bien meilleure qualité que l'aIB. Ces données sont suffisantes pour construire la base de données. Ainsi, l'utilisation de l'Information Bottleneck en cascade avec notre critère permet de rapidement faire une étude non supervisée des classes d'hyperspectres sans demander de prétraitement particulier. Naturellement, une analyse plus approfondie des spectres et de leur prétraitement améliorerait encore les performances. La prise en compte *a posteriori* des notions de voisinage dans l'image filtrerait aussi certaines erreurs de segmentation notamment à l'interface des régions.

5 Conclusion

Cet article introduit un nouveau critère permettant d'améliorer la méthode de l'agglomerative Information Bottleneck introduite dans la littérature [2]. Ce critère est justifié par la réécriture de la procédure sous forme d'une mise en cascade d'encodeurs déterministes bien choisis.

La solution ainsi développée a été testée pour classer des hyperspectres issus d'une image hyperspectrale et en autorise une première segmentation de qualité suffisante.

Références

[1] N. Tishby, F. Pereira et W. Bialek, W. *The information bottleneck method*. Proc. 37th Allerton Conference on Communication, Control, and Computing, 1999.

[2] N. Slonim, R. Somerville, N. Tishby et O. Lahav. *Objective classification of galaxy spectra using the information bottleneck method*. Oxford University Press, 2001, 323, 270-284.

[3] P. Harremoes et N. Tishby. *The Information Bottleneck Revisited or How to Choose a Good Distortion Measure* Proc. International Symposium on Information Theory, Nice, 2007.