Nouveau modèle hiérarchique ascendant pour la sélection des bandes spectrales discriminant les maladies de la vigne

Shurong Zhang[1], Eric Perrin[1], Alban Goupil[1], Valeriu Vrabie[1], Marie-Laure Panon[2]

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

[2]Comité Champagne, 5 Rue Henri Martin, 51200 Épernay, France

[1]shurong.zhang@univ-reims.fr, [2]marie-laure.panon@civc.fr

Résumé − Nous proposons un nouveau modèle hiérarchique ascendant pour la sélection des bandes spectrales adaptées à la discrimination de plusieurs classes à partir de spectres acquis par spectroscopie infrarouge. Nous utilisons différentes mesures de similarité pour comparer des bandes adjacentes et nous utilisons deux types de critères, sélection et agrégation, pour obtenir des bandes discriminantes. Ce modèle est appliqué sur des spectres acquis sur des feuilles de vigne au cours des trois années de 2020 à 2022. Les résultats montrent qu'un nombre limité de bandes étroites suffit pour identifier les classes d'intérêt par une analyse discriminante linéaire. Notre modèle est plus simple à mettre en œuvre que les modèles de regroupement hiérarchique ascendants existants. Il ouvre la voie à l'identification de bandes spectrales discriminantes de maladies et à la spécification de caméras multispectrales adaptées.

Abstract − We propose a new bottom-up hierarchical model for the selection of spectral bands suitable for multi-class discrimination of spectra acquired by IR spectroscopy. We use different similarity measures to compare adjacent bands and two types of criteria, selection and aggregation, to obtain discriminative bands. This model is applied on spectra acquired on vine leaves during 3 years, from 2020 to 2022. The results show that a limited number of narrow bands is sufficient to identify the classes of interest through a LDA (Linear Discriminant Analysis) classifier. Our model is simpler to implement than existing bottom-up hierarchical clustering models. It opens the way to the identification of discriminating spectral bands of diseases and to the specification of adapted multispectral cameras.

1 Introduction

La flavescence dorée, maladie grave et épidémique, est l'une des deux principales jaunisses de la vigne. Pour maîtriser son risque de propagation, il est nécessaire de concevoir un outil de détection efficace des jaunisses. Dans ce contexte, nous nous intéressons à la conception de méthodes de sélection de bandes spectrales discriminantes pour les maladies de la vigne à partir de spectres acquis par spectroscopie infrarouge, en vue de spécifier des caméras multispectrales adaptées à des acquisitions à large échelle. Les bandes spectrales peuvent avoir des largeurs différentes et doivent permettre de séparer plusieurs classes.

Nous nous plaçons donc dans un contexte de réduction des dimensions de données spectrales. Cette tâche peut être effectuée en utilisant des techniques d'extraction ou de sélection des caractéristiques. Les techniques d'extraction créant de nouvelles caractéristiques à travers des combinaisons des bandes [1] ne sont pas adaptées, car elles nécessitent de disposer de toutes les informations spectrales. Les techniques de sélection, qui permettent de choisir des bandes représentatives, sont plus adaptées.

Dans [2], un modèle de regroupement hiérarchique non supervisé est proposé. L'information mutuelle et la divergence de Kullback-Leibler sont utilisées pour calculer la similarité entre bandes et deux modèles WaluMi et WaluDi ont été proposés pour fusionner ces bandes. La longueur d'onde de chaque bande présentant la plus grande similarité moyenne avec le reste est choisie comme représentant d'une bande. Un travail similaire [3], en utilisant une distance hyperbolique adaptative, permet d'éviter l'obtention de bandes contenant une unique longueur d'onde, ce qui améliore les performances. Cependant, d'un point de vue pratique, que des longueurs d'onde adjacentes peuvent être fusionnées. Nous avons toutefois comparé les performances obtenues avec les bandes identifiées par notre méthode avec les modèles WaLuDi et WaLuMi, y compris en utilisant la distance hyperbolique adaptative de [3].

Dans [4], l'algorithme non supervisé MRMR (Max Relevance and Min Redundancy) adopte une sélection statistique en mesurant indépendamment l'importance de chaque bande pour sélectionner un sous-ensemble optimal en utilisant l'information mutuelle. Nous avons retenu cette méthode pour la comparaison puisqu'elle est aussi basée sur des notions de théorie de l'information. Une approche hiérarchique a également été proposée [5]. Celle-ci calcule

une corrélation moyenne entre les informations spectrales à toutes les longueurs d'onde pour fusionner des bandes. Cette approche hiérarchique étant assez similaire, nous l'avons aussi retenue à titre de comparaison.

2 Regroupement proposé

Nous proposons une classification ascendante hiérarchique qui permet soit d'agréger des longueurs d'onde adjacentes dans des bandes selon leur similarité, soit de trouver des longueurs d'onde représentatives. Dans le premier cas, chaque bande est représentée par un intervalle regroupant plusieurs longueurs d'onde pondérées afin de simuler la réponse d'un filtre optique. Dans le deuxième cas, nous conservons aussi la notion de bande, mais celle-ci se résume à une seule longueur d'onde. Le pseudo-code de notre méthode est présenté dans l'algorithme 1.

2.1 Conservation de l'adjacence

Nous disposons de P spectres de réflectance, chacun ayant été acquis pour une classe $c \in C$, sur n longueurs d'onde $S(\lambda_i) \in \mathbb{R}^P$, où $\lambda_i \in [\lambda_1 ... \lambda_n]$. Initialement, chaque bande B_i ne contient qu'une seule longueur d'onde λ_i. Les bandes adjacentes B_i et B_{i+1} les plus similaires sont fusionnées à chaque pas tant que le nombre de bandes n'atteint pas la valeur cible. La similarité n'étant calculée qu'entre une bande et ses deux voisines, il suffit de calculer la liste des similarités entre les bandes adjacentes au départ puis de la mettre à jour à chaque fusion de bandes.

Data: Liste des longueurs d'onde $\lambda_1, \ldots, \lambda_n$
Data: Réflectance $S(\lambda_i) \in \mathbb{R}^P$, avec P le nombre de spectres acquis, chacun appartenant à une classes $c \in C$
Result: Ensemble de m bandes $\mathcal{B} = \{B_1, \ldots, B_m\}$
`/* Initialisation */`
$\mathcal{B} \leftarrow \{B_i\}$ avec $B_i = \{\lambda_i\}, i = 1, \ldots, n$;
Représentants $R_i = S(B_i)$ cf section 2.2.1;
Distances $d(R_i, R_{i+1})$ entre les représentants des bandes adjacentes, $i = 1, \ldots, n-1$ cf section 2.2.2;
`/* Regroupement hiérarchique */`
while $|\mathcal{B}| > m$ **do**
  $i \leftarrow \arg\min_i d(R_i, R_{i+1})$;
  $B_i \leftarrow B_i \cup B_{i+1}$ et retirer B_{i+1} de la liste $\mathcal{B}$;
  Calculer le nouveau représentant $R_i = R(B_i)$ de la bande B_i;
  Recalculer les distances $d(R_{i-1}, R_i)$ et $d(R_i, R_{i+1})$;
return $\mathcal{B}$;

Algorithme 1: classification ascendante hiérarchique

2.2 Recherche d'une similarité pertinente

Le regroupement est réalisé selon une similarité entre des bandes définies dans 2.2.2. Cette similarité est calculée entre les représentants R_i de chaque bande B_i estimés cf. 2.2.1. Si la similarité est maximale, alors les deux bandes respectives sont regroupées dans une nouvelle bande.

2.2.1 Représentant d'une bande

Nous proposons deux approches pour trouver le représentant R_i d'une bande B_i : soit une sélection de la bande maximisant l'information mutuelle avec la classe soit une bande moyenne pondérée par une fenêtre de Hanning.

Sélection par Information Mutuelle (SIM) La première approche consiste à trouver la longueur d'onde la plus représentative d'une bande en utilisant l'information mutuelle. Cette mesure s'appuie sur la relation entre les informations spectrales (les valeurs de réflectance acquises par le spectromètre) contenues à une longueur d'onde λ_k et les classes d'appartenance des spectres $c \in C$. Nous calculons l'information mutuelle des spectres à chaque longueur d'onde λ_k contenue dans la bande B_i, donc $\lambda \in B_i$, par rapport aux classes c :

$$I\big(S(\lambda_k); C\big) = \int p(s,c)\ \log \frac{p(s,c)}{p(s)\,p(c)}\, ds\, dc, \qquad (1)$$

où $p(s,c)$ est la densité jointe des valeurs s des spectres $S(\lambda_k)$ à la longueur d'onde λ_k et de la classe c [6]. Cette information mesure la pertinence de la longueur d'onde λ_k pour discriminer les informations spectrales selon les classes C. L'information mutuelle est estimée selon la méthode décrite dans [7].

Nous choisissons ensuite le représentant R_i de la bande B_i comme étant l'information spectrale à la longueur d'onde qui maximise l'information mutuelle :

$$\lambda^* = \arg\max_{\lambda_k \in B_i} I\big(S(\lambda_k); C\big) \qquad (2)$$

$$R_i = S\big(\lambda^*\big). \qquad (3)$$

L'avantage de cette méthode est qu'elle s'appuie sur l'information mutuelle qui existe entre les spectres acquis aux différentes longueurs d'onde et leurs classes $c \in C$.

Agrégation avec Fenêtrage de Hanning (AFH) La deuxième approche consiste à pondérer les informations spectrales aux longueurs d'onde constituant une bande B_i. Pour ce faire, nous utilisons une fenêtre de Hanning qui approche la réponse d'un filtre optique réel. Le représentant R_i de la bande B_i est calculé comme la moyenne pondérée des informations spectrales dans cette fenêtre :

$$R_i = \frac{1}{L} S(B_i) \cdot W(L) \qquad (4)$$

où $W(L) \in \mathbb{R}^L$ est la fenêtre de Hanning de taille L, taille correspondant au nombre de longueurs d'onde contenues

dans la bande B_i, et $S(B_i) \in \mathbb{R}^{P \times L}$ aux informations spectrales aux longueurs d'onde contenues dans cette bande.

L'avantage de cette méthode est de permettre l'utilisation de filtres optiques de bande passante plus larges qui laissent passer plus de flux optique et qui souvent sont moins onéreux que les filtres à bande étroite.

2.2.2 Distance entre les représentants

Nous avons retenu cinq approches pour calculer la distance entre les représentants des bandes : la distance euclidienne L_2 qui est communément utilisée ; la distance hyperbolique adaptative D_{HY} proposée par [3] ; la divergence de Jensen-Shannon qui tire profit de la distribution des informations spectrales et non pas des valeurs des spectres ; l'information mutuelle conditionnelle qui permet de prendre en compte la classe $c \in C$ d'appartenance de chaque spectre ; une des approches proposées dans [5]. Nous ne détaillons pas ici les deux premières.

Divergence de Jensen-Shannon (JS) La divergence de JS peut être interprétée comme une sorte de similarité entre deux distributions de probabilité qui symétrise la distance de Kullback-Leibler.

Nous supposons que R_i et R_j sont les représentants des bandes B_i et B_j. La distance basée sur la divergence de JS, D_{JS} est définie par :

$$D_{JS}(R_i, R_j) = \frac{1}{2} D_{KL}(R_i \,|\, M) + \frac{1}{2} D_{KL}(R_j \,|\, M), \quad (5)$$

où D_{KL} est la divergence de Kullback–Leibler entre les densités de probabilités des variables R_i, R_j et M. Cette dernière densité M est la distribution moyenne des distributions de R_i et R_j.

La divergence de Jensen-Shannon est toujours positive ou nulle. Elle s'annule lorsque R_i et R_j ont la même distribution de probabilité. Ainsi, deux bandes B_i et B_j ayant une faible valeur de D_{JS} sont combinées si les représentants de ces bandes ont des densités proches.

Information mutuelle conditionnelle Nous cherchons à mesurer la dépendance entre les informations spectrales des bandes B_i et chaque classe $c \in C$. Pour cela, nous avons partitionné les spectres selon leur classe d'appartenance c. Chaque classe a une probabilité $p(c)$ correspondant à sa proportion par rapport aux autres classes (nombre de spectres acquis dans la classe c par rapport au nombre total de spectres acquis dans toutes les classes C). Nous calculons l'information mutuelle $I(R_i; R_j \mid c)$ entre les représentants des bandes B_i et B_j conditionnellement à la classe c puis estimons la moyenne de ces informations mutuelles pour toutes les classes :

$$I(R_i; R_j \mid C) = \sum_{c \in C} I(R_i; R_j \mid c) \times p(c). \quad (6)$$

La distance proposée basée sur cette information mutuelle conditionnée prend en compte l'appartenance des informations spectrales aux classes $c \in C$ dans le calcul :

$$D_{IM}(R_i, R_j) = \frac{1}{1 + I(R_i; R_j \mid C)}. \quad (7)$$

Cela permet d'identifier des bandes discriminantes par rapport aux classes dont on dispose, en occurrence les bandes qui permettent de mieux identifier les maladies de la vigne.

Parmi les approches de calcul proposées dans [5], nous avons retenu la distance moyenne basée sur des mesures de corrélation Corr_*moy*. A noter qu'au lieu de calculer une distance entre les représentants des bandes, ces approches estiment une distance moyenne basée soit sur la corrélation, soit sur l'erreur d'approximation d'un spectre, soit sur la séparabilité entre bandes. La distance retenue ici pour comparaison consiste à moyenner les corrélations entre toutes les informations spectrales à toutes les longueurs d'onde contenues dans les bandes à fusionner hiérarchiquement.

3 Résultats

Nous avons mené plusieurs campagnes d'acquisitions de spectres sur des feuilles de vigne du cépage Chardonnay dans le domaine expérimental de Plumecoq du Comité Champagne. Les acquisitions ont été réalisées après les périodes de vendange des années 2020 à 2022. Les spectres ont été acquis avec un spectromètre portable LabSpec4 en conditions contrôlées et uniquement la bande 400 nm - 1750 nm a été retenue pour l'analyse. A noter que ce spectromètre fournit des informations spectrales tous les 1 nm, ce que nous appelons par abus de langage longueurs d'onde ($\lambda_i, i = 1, \ldots, n$ avec $n = 1351$) dans cet article.

L'objectif de cette étude a été de déterminer des bandes spectrales qui permettent d'identifier 4 classes : témoins (2187 spectres) et 3 maladies de la vigne : jaunisse (2434 spectres), enroulement (2041 spectres) et esca (1567 spectres). A noter que l'esca et surtout l'enroulement ont des symptômes confondants avec la jaunisse. Pour identifier les bandes discriminantes, l'ensemble de $P = 8229$ spectres acquis pour les $C = 4$ classes a été utilisé. Ensuite, nous avons divisé les données en deux ensembles, un d'entraînement et un de test, en utilisant la stratégie de validation croisée avec 10 tirages et 3 répétitions pour chaque tirage. L'ensemble d'entraînement contient 90% des données et celui de test les 10% restant. Sur les premiers, un classifieur Linear Discriminant Analysis (LDA) a été entraîné et les performances de classification ont été évaluées sur les jeux de tests en considérant les informations spectrales aux bandes identifiées par les méthodes de sélection.

Nous présentons dans la Figure 1 les valeurs de précision moyennes sur les divers tirages pour chaque méthode de sélection et cela en fonction du nombre de bandes rete-

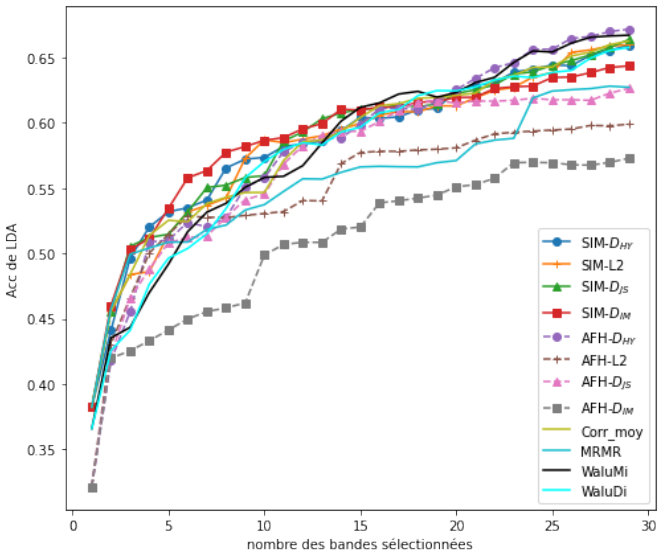

FIGURE 1 – Valeurs de précision de la LDA avec les différentes méthodes.

nues. Les courbes en trait continu, qui ont des marqueurs différents pour chacune des distances, correspondent aux méthodes de sélection par Information Mutuelle (SIM). Les courbes en trait pointillé, qui ont les mêmes marqueurs pour les mêmes distances, correspondent aux méthodes d'agrégation avec fenêtrage de Hanning (AFH). Les courbes en trait continu, sans marqueurs, correspondent aux méthodes de l'état de l'art.

D'après cette figure, nous constatons que notre méthode qui sélectionne un représentant par Information Mutuelle et la distance D_{IM} (SIM-D_{IM}) fournit les meilleures performances en choisissant un nombre limité des bandes (entre 6 et 10). Nous constatons que le choix d'un nombre limité de bandes étroites, par exemple 6 bandes pour rester dans des configurations usuelles de caméras multispectrales, offre un taux de précision correct, d'environ 55% pour une classification en 4 classes. Des gains plus importants pourront être apportés en ajoutant des informations spatiales acquises par les caméras multispectrales. Les approches avec agrégation des bandes AFH permettent d'obtenir de bonnes performances en utilisant la distance hyperbolique D_{HY} et la divergence Jensen-Shannon D_{JS}. Leurs performances sont légèrement inférieures à celles de la méthode SIM-D_{IM} en choisissant un nombre limité de bandes. Finalement, nous observons que notre méthode, qui est moins complexe que WaluDi et WaluMi, produit des résultats plus précis, en particulier lorsqu'il s'agit de sélectionner un nombre restreint de longueurs d'onde. Les conclusions sont identiques par rapport aux méthodes de l'état de l'art, MRMR et Corr_*moy*. Ces résultats ont été confirmés avec d'autres méthodes de classification telles que KNN et les forêts aléatoires.

4 Conclusion et perspectives

Nous avons proposé une méthode de sélection de bandes spectrales ascendante hiérarchique. Cette méthode permet de trouver des bandes étroites avec le critère d'information mutuelle (SIM) ou d'agréger des bandes larges par fenêtrage (AFH) en utilisant différentes distances (L2, distance hyperbolique, divergence de Jensen-Shannon, information mutuelle et corrélation moyenne) pour évaluer la similarité entre les représentants des bandes. Les bandes discriminantes, optimisant la séparation entre classes qui ne sont pas forcément bien séparées à cause de symptômes confondants, permettent de spécifier des caméras multispectrales adaptées à l'identification des maladies.

Le regroupement obtenu avec notre méthode est plus simple à implémenter que les regroupements hiérarchiques ascendants existants (complexité quadratique contre cubique). En effet, au lieu de calculer la similarité entre toutes les bandes, nous nous concentrons uniquement sur les similarités des bandes adjacentes. De plus, nos mesures de similarité se mettent à jour à chaque fusion de bandes, rendant ainsi notre algorithme efficace. Des travaux sont en cours pour rendre l'algorithme robuste en fonction des acquisitions réalisées sur plusieurs années et pour le paramétrer afin qu'il identifie des bandes compatibles avec les filtres du commerce.

Remerciements. Nous remercions la région Grand-Est pour le cofinancement de ces travaux.

Références

[1] A.K. Jain, R.P.W. Duin, and Jianchang Mao. Statistical pattern recognition : a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1) :4–37, 2000.

[2] Adolfo MartÍnez-UsÓMartinez-Uso, Filiberto Pla, JosÉ MartÍnez Sotoca, and Pedro GarcÍa-Sevilla. Clustering-based hyperspectral band selection using information measures. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12) :4158–4171, 2007.

[3] He Sun, Lei Zhang, Jinchang Ren, and Hua Huang. Novel hyperbolic clustering-based band hierarchy (HCBH) for effective unsupervised band selection of hyperspectral images. *Pattern Recognition*, 130 :108788, October 2022.

[4] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8) :1226–1238, 2005.

[5] A. Le Bris, N. Chehata, X. Briottet, and N. Paparoditis. Extraction of optimal spectral bands using hierarchical band merging out of hyperspectral data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-3/W3 :459–465, 2015.

[6] Chein-I Chang, Qian Du, Tzu-Lung Sun, and M.L.G. Althouse. A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 37(6) :2631–2641, 1999.

[7] Brian C. Ross. Mutual information between discrete and continuous data sets. *PLoS ONE*, 9(2) :e87357, feb 2014.