Brève introduction à la fouille de grandes bases de données océaniques

Guillaume Maze[1], Herlé Mercier[2],
Ronan Fablet[3], Philippe Lenca[3]
et Jean-François Piollé[4]

[1]*Ifremer, UMR 6523, Laboratoire de Physique des Océans, F-29280 Plouzané, France*
[2]*CNRS, UMR 6523, Laboratoire de Physique des Océans, F-29280 Plouzané, France*
[3]*Institut Mines-Telecom, Télécom Bretagne, UMR CNRS 6285 Lab-STICC Brest, France*
[4]*CERSAT, Ifremer, Spatial Oceanog. Lab, Brest, France*

RÉSUMÉ

Les bases de données marines, alimentées par les satellites et les robots autonomes sous-marins comme les flotteurs du réseau Argo, sont de plus en plus grandes (plusieurs dizaines de gigaoctets et teraoctets) et rapidement évolutives (elles changent d'heure en heure). Cette augmentation spectaculaire de la dimension et de la complexité des données rend difficile leur exploitation avec les outils standards. Or, c'est à partir de l'analyse des données que les chercheurs pourront réaliser de nouvelles découvertes scientifiques sur la dynamique des océans, à grande et petite échelles, et les changements climatiques régionaux et globaux. L'école d'été OBIDAM14 visait à contribuer à lever ces verrous d'analyse en introduisant les méthodes de fouille de données aux scientifiques de la communauté de recherche en océanographie physique. Pour un des premiers évènements académiques sur ce thème en France, OBIDAM14 a permis de donner un aperçu de ces méthodes. Le comité scientifique de l'école présente ici un compte-rendu des interventions.

L'école d'été OBIDAM14 a été financée par l'Ifremer, le LaBeX Mer et la région Bretagne. L'école a été organisée avec la soutient de l'IRD, de l'UBO et de Télécom Bretagne. Le comité scientifique remercie les intervenants V. Kumar, P. Naveau, P. Gançarski et S. Canu pour leur contributions. Les intervenants ont le crédit des figures utilisées dans ce document.

Table des matières

Document Ifremer ODE-LPO-14-101 enregistré sur archimer le 16 octobre 2014.
Date du dernier rendu: 16 octobre 2014

1 Opportunités et défis dans l'analyse des données du système Terre

Présentation donnée par Vipin Kumar (Université du Minesota)
Disponible en ligne : http://goo.gl/uPnKCe

V. Kumar a commencé son intervention par une brève introduction aux méthodes de fouille de données. Les méthodes se classent en deux catégories, les prédictives et les descriptives. Les méthodes prédictives utilisent un ensemble de variables pour prédire des valeurs de ces variables dans le futur ou des valeurs de variables inconnues. Par exemple, les algorithmes de classification supervisée (attribuer un attribut de classe à une donnée) sont des méthodes prédictives qui cherchent à modéliser la relation entre des valeurs d'attributs et celle de la classe à prédire. Déterminer si une transaction bancaire est légitime ou frauduleuse est un exemple de classification, tout comme la classification des types de sols (étendue d'eau, espace urbain, forêt, etc . . .) à partir des données satellites. Les méthodes descriptives quant à elles, visent à trouver dans les données des structures ("patterns") interprétables par l'homme. Le "clustering" est une méthode descriptive qui permet d'identifier des groupes d'objets de telle manière que les objets d'un groupe seront similaires (ou liés) tout en étant différents (ou non-liés) aux objets des autres groupes. L'analyse des clusters permet de comprendre la structure d'un jeu de données et par voix de conséquence d'en réduire la taille en résumant l'information qu'il contient. D'autres exemples suivront comme la détection automatique de règles associatives et la détection d'écarts significatifs à un comportement normal.

Puis V. Kumar a parlé plus spécifiquement des opportunités offertes par ces méthodes pour les sciences de l'environnement. En effet, comme les organisateurs de l'école d'été l'ont déjà souligné, les bases de données des sciences de l'environnement explosent en complexité et tailles. Cette évolution est à rapprocher de l'explosion des bases de données liées à l'usage social d'internet et auquel se confrontent les géants du web comme Google, Twitter ou Facebook pour générer des revenus. Pour générer de la connaissance et de nouvelles découvertes, notre communauté doit elle aussi se confronter au "Big Data". Avant de passer à la présentation précise de trois applications, V. Kumar nous illustre le champs des possibles comme une meilleure compréhension de la dynamique océanique globale à toutes les échelles ou l'identification de téléconnections pertinentes pour la prévision climatique mais jusqu'ici sous-évaluées ou non-détectées. Vipin n'oubliera pas de parler également les difficultés scientifiques et techniques auxquelles il faudra faire face pour développer ces nouvelles applications : l'incertitude et le bruit dans les données, le manque de validation indépendante, la perte d'information dans l'assemblage de base de données homogènes, les non-linéarités et bien sûr le caractère dynamique de l'objet d'étude, l'océan, qui change de forme, dimensions, propriétés à toutes les échelles de temps et d'espace.

V. Kumar nous détaille ensuite 3 exemples d'application des méthodes de fouilles aux données environnementales : le suivi des changements des écosystèmes, l'identification automatique de dipôles atmosphériques et la caractérisation des tourbillons océaniques.

1.1 Suivi des écosystèmes

Le suivi des écosystèmes peut se faire de manière traditionnelle en comparant deux images d'un même système mais prises à deux dates différentes. Bien sûr, cette méthode est manuelle, donc chère en ressource et expertise humaine. Dans le cas où l'écosystème à surveiller est la surface terrestre, l'analyse des données satellites d'instrument type MODIS fournit une alternative unique. L'instrument donne accès à une base de données contenant des images journalière de toute la surface du globale avec une résolution de 30 à 250m depuis plus de 10 ans. L'analyse des séries temporelles des index de couverture végétales permet la détection des changements d'origines humaines comme la déforestation ou naturelle comme les feux. V. Kumar nous présentera en détails le système "ALERTS" : Automated Land change Evaluation, Reporting and Tracking System, basée sur l'analyse automatique de ces alertes.

1.2 Identification automatique de téléconnections

Le deuxième exemple d'application est celui de l'identification automatique de dipôles atmosphériques. Ces dipôles, parmi lesquelles on connaît bien la NAO (North Atlantic Oscillation) ou la SOI (Southern Oscillation Index, qui manifeste les El Nino/La Nina), jouent un rôle très important dans le système climatique en provoquant des anomalies de température et précipitation à travers les continents et la planète.

Ces dipôles sont principalement mis au jour par l'oeil humain et l'analyse en composantes principales. Mais ces méthodes rendent difficile l'identification de dipôles de moindre importance, imposent aux dipôles des centres d'actions fixe dans l'espace et le temps et leur interprétation est parfois difficile car elle ne prend pas en compte le caractère dynamique des patterns. V. Kumar nous présente comment utiliser la théorie des graphes pour identifier automatiquement ces dipôles sans les inconvénients des méthodes classiques. La théorie des graphes (à ne pas confondre avec la représentation graphique des fonctions mathématiques) est le champ d'étude mathématique des ensembles de points ("node" ou sommet) dont certaines paires sont directement reliées par un ou plusieurs liens ("edge" ou arête). Pour

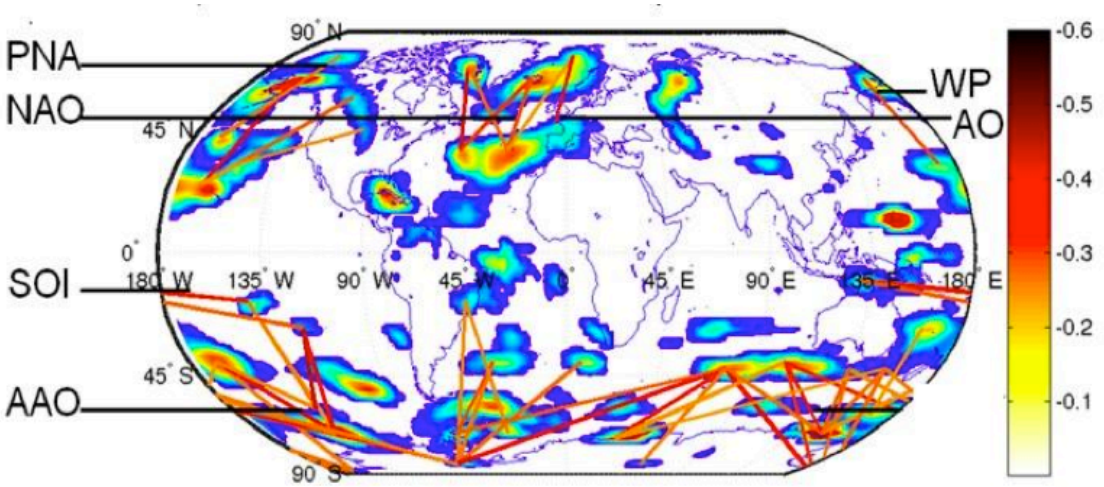

FIGURE 1: Application de la méthode automatique de détection de dipôles atmosphériques (pression de surface)

cette application, les sommets sont les points d'une grille d'une série temporelle de cartes géolocalisées - de pression atmosphérique par exemple - et les arêtes sont les corrélations des séries temporelles entre les sommets. Chaque point de la grille à donc N-1 arêtes le reliant aux autres points de la grille. Les régions pouvant être les centres d'actions d'un dipôle sont identifiées automatiquement dans cet espace à N x N-1 éléments en cherchant de manière itérative les pairs de sommets distants ayant tout à la fois (i) des arêtes, ou corrélations, négatives (car les dipôles sont des oscillations) et (ii) des voisins avec des corrélations positives. Cette approche a été développée par Jawa Kawale et co-auteurs dans une série d'articles entre 2011 et 2012 (Kawale et al., 2011a,b, 2013), elle est illustrée sur la Figure 1.

1.3 Suivi des tourbillons océaniques

Le troisième et dernier exemple donné par V. Kumar est celui de la détection et suivi des tourbillons océaniques à partir des données satellites altimétriques de hauteur de mer AVISO. V. Kumar nous présente une nouvelle méthode développée par James Faghmous (Faghmous et al., 2013). A l'inverse des méthodes classiques basées sur des valeurs seuils de détection des anomalies de hauteur de mer (dites top-down : TD) qui ont l'inconvénient de devoir paramétrer de manière plus ou moins complexes ces seuils, cette nouvelle méthode est dite bottom-up (BU). En effet elle part d'un minimum locale et augmente automatiquement son étendue jusqu'à trouver 2 extremums locaux. Les deux méthodes sont illustrées sur la Figure 2 (reprise de la fig. 4 de Faghmous et al., 2013) pour 4 types de tourbillons. Le lecteur remarquera comment la méthode BU identifie plus de détails dans la situation d/h et n'ajoute pas de bruit aux résultats comme dans la situation b/f. L'autre avantage de la méthode est de pouvoir suivre le processus de fusion des tourbillons (comme dans le cas h ci-dessus), une analyse qui reste à faire selon leurs auteurs. V. Kumar nous présentera ensuite en détail les résultats statistiques de la méthode appliquée aux 20 ans de données AVISO.

2 Méthodes statistiques pour la détection et l'attribution des changements climatiques

Présentation donnée par Philippe Naveau (LSCE, Paris)
Disponible en ligne : `http://goo.gl/rmzGvG`

Après une présentation des caractéristiques de la variabilité naturelle du climat, P. Naveau a définit les notions de **détection** et **attribution** selon l'IPCC. La détection est la démonstration que le climat, ou un système affecté par le climat, a changé au sens statistique du terme sans fournir une raison de ce changement. Ici, le sens statistique signifie au delà ce qui peut être expliqué par la variabilité intrinsèque, ou naturelle, du système. En revanche l'attribution est l'évaluation des contributions relatives de multiples facteurs de causes à un changement, ou évènement, avec un indice de confiance statistique. Ici, les facteurs se rapportent essentiellement aux influences externes au système, donc ils peuvent être anthropogéniques mais également naturelles (les éruptions volcaniques par exemple). Par voix de conséquences, pour pouvoir "faire de l'attribution", nous devons pouvoir déterminer si les changements détectés (voir définition ci-dessus) sont (i) consistants avec les réponses attendues à des forçages externes et (ii) inconsistants avec des explications alternatives. P. Naveau nous a ensuite décrit en détail deux approches statistiques classiques pour la

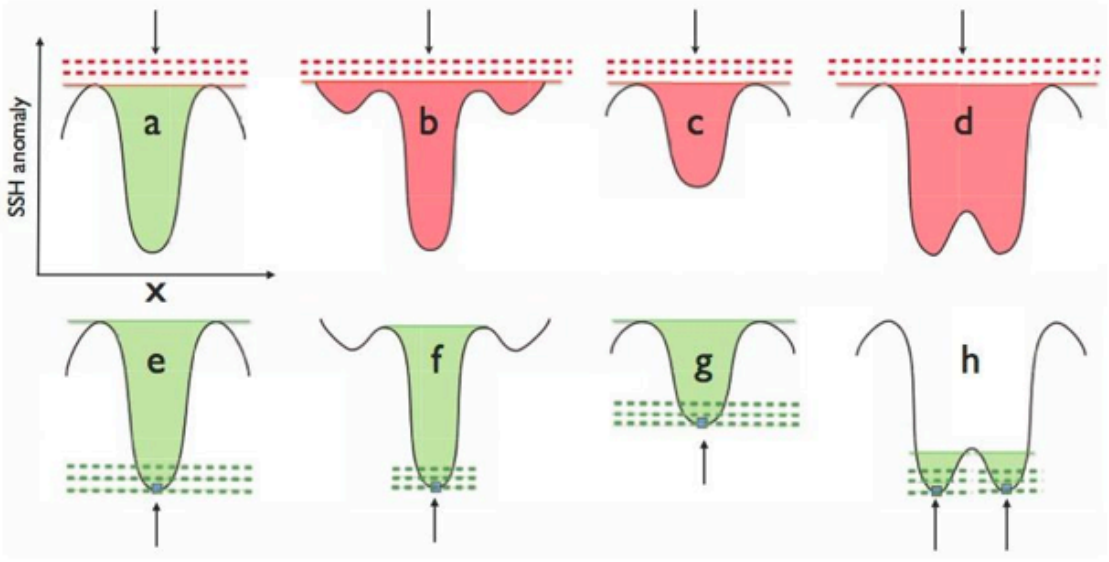

Fig. 4. A two-dimensional cross section of SSH anomalies. The arrows and dashed lines represent the direction of the iterative thresholding method. The color of the feature (red or green) represents whether each method was able to accurately recover each features boundaries. TD starts from a very high threshold and gradually decreases and stops as soon as it finds a close contour that meets its expert-criteria. Alternatively, BU (bottom row) thresholds locally starting from each local minima and gradually grows to reconstruct the features body and stops once a feature contains two extrema. Unlike TD, BU is able to avoid adding noise to the features contour (panel f), does not discard features due to arbitrary parameters (panel g), and is able to separate features in close proximity (panel h) when TD effectively merges them (panel d)

FIGURE 2: Schémas représentant comment se comporte la méthode classique top-down (a-b-c-d) et la nouvelle méthode bottom-up (e-f-g-h) de suivi des tourbillons dans 4 quatre de figures (extrait de Faghmous et al, IEEE, 2013)

détection/attribution (D&A) : la régression linéaire (problème de détection) et l'estimation de la fraction de risque attribuable ou FAR (problème d'attribution).

2.1 La détection

D'un point de vue statistique la difficulté majeure vient du fait qu'il n'existe qu'une seule Terre et donc que nous ne disposons que d'une seule et unique observation du système, un seul long vecteur temps x espace. La stratégie clé pour contourner cette difficulté intrinsèque est d'utiliser des modèles de climat pour générer des avatars - réalisations - de notre Terre (eg : une simulation avec forçages réelles - ANT - sera un avatar/une réalisation, une autre simulation avec seulement les forçages naturelles - NAT - sera un autre avatar de notre Terre). Ainsi le schéma de **régression linéaire** consistera à estimer les paramètres β et ϵ dans l'équation $Y = \beta X + \epsilon$ où dans notre cas, Y est le climat réel observé, X le vecteur des avatars de climat (ANT et NAT par exemple), β le modèle et ϵ les erreurs.

En utilisant l'approximation gaussienne, P. Naveau nous a ensuite présenté les solutions pour un système généralisé où les corrélations entre les erreurs du modèle sont supposées connues et prises en compte. Il nous présentera quelques exemples issus du rapport IPCC avant de souligner les limitations et difficultés de cette approche soulevées par (i) la dimension du système, (ii) l'estimation de la matrice de covariance des erreurs Σ (la variabilité interne) et (iii) le modèle en lui même, ici numérique.

En effet, pour un jeu de données classique, la base Y a une dimension de l'ordre de 105 points (une grille de 5x5 degrés et 50 à 100 ans de simulation), donc la matrice de covariance, ou la variabilité interne, est décrite par une matrice carrée de grande dimension 105x105 qui n'est absolument pas sparse à cause des téléconnections. La solution consiste alors à tenter de diminuer la dimension de Y (moyennes temporelles, projection sur les composantes principales, utilisation d'indicateurs climatiques) ou de calculer des estimateurs de la matrice de covariance (à partir de simulations d'ensemble mono ou multi modèles, comme CMIP5). Mais cette dernière tâche est particulièrement difficile souligne P. Naveau. En effet, il faut faire appel aux méthodes de régularisation de matrice et calcul de probabilités conditionnelles pour y parvenir et P. Naveau nous en montrera toute la difficulté en dérivant une méthode d'estimation conjointe de β et Σ.

Il conviendra enfin de noter que l'usage de simulations numériques de modèles de circulation générale introduit une source d'incertitude dans ce problème de régression linéaire et le vecteur X des avatars du systèmes est de fait un vecteur bruité. Malheureusement, il n'existe pas de solution à ce problème (connu sous le nom de EIV : Error-In-Variable model) qui est le sujet de nombreuses recherches en cours.

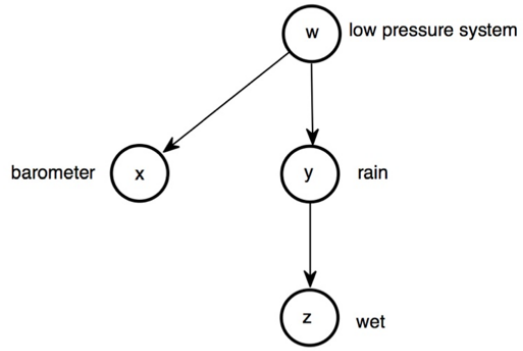

FIGURE 3: Graphe orienté du système x,y,z,w utilisé en exemple. Un arc allant d'un sommet A vers un sommet B signifie une forte probabilité que B soit la cause de A.

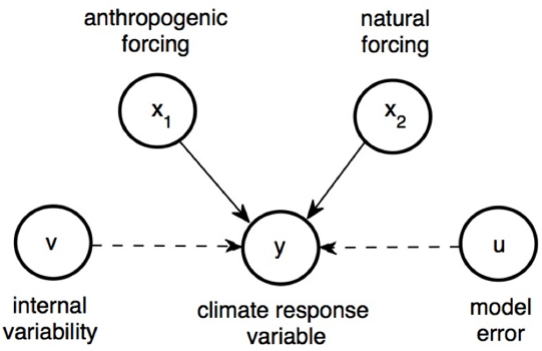

FIGURE 4: Graphe orienté du système climatique terrestre

2.2 L'attribution

P. Naveau revient ensuite sur la notion d'attribution en tentant de définir plus précisément la notion de **causalité**. Historiquement elle signifie que l'objet Y est la cause de X si Y suit X et que Y n'aurait pas été sans X (Hume, 1748). Tout le corpus théorique qui s'en suit est présenté dans le livre de J. Pearl (Pearl, 2000). P. Naveau illustre ces concepts à partir d'un exemple simple où les objets X, Y, etc, sont : la pluie, l'état d'une route (mouillée/sèche), la mesure d'un baromètre local et la présence d'un système dépressionnaire régional. Cet exemple permet d'introduire les notions de probabilité conditionnelle (eg : quelle est la probabilité que la route soit mouillée s'il a plut ? ou si la pression a diminué ?) ainsi que leur représentation en graphe orienté. On retrouve ici la théorie des graphes utilisée dans la méthode d'identification automatique de dipôles atmosphériques présenté par V. Kumar le matin. Mais ici les arêtes entre les sommets ne sont pas symétriques (comme l'était les corrélations entre séries temporelles de 2 points de grille) mais représentent la probabilité qu'un objet (sommet) soit la cause d'un autre, d'où l'adjectif "orienté". Dans ce cas, on parle d'ailleurs d'arcs, plutôt que d'arêtes. La Figure 3 donne un graphe orienté de notre exemple.

L'usage de ces derniers est cependant limité car plusieurs représentations peuvent être compatibles avec le même jeux de probabilités (et donc jeux d'observations). Ici nous aurions pu placer l'objet Z après l'objet X car la route est mouillée quand il a plut, mais également quand la pression a baissé. Est-ce la pluie ou la baisse de pression qui est la cause de la route mouillée ? Deux graphes orientés sont donc possibles pour un même ensemble d'observations.

Cette ambiguïté vient du fait que nous avons traité l'exemple avec des probabilités conditionnelles **observées**. Nous avons observé la pluie, la baisse de pression, puis la route mouillée. Pour déterminer si c'est la pluie ou la pression qui est la cause de la route mouillée, il faut pouvoir faire l'un sans l'autre. Conceptuellement, il faut séparer les liens observés entre les objets, c'est à dire rompre la chaîne de causalité en **intervenant**. Cela se fait par l'expérimentation et se formalise dans la notion de probabilité interventionnelle Hagmayer et al. (2007). Qu'elle est la probabilité que la route soit mouillée si la pression baisse sans que la pluie tombe (c'est à dire en cassant le lien de Y vers Z) ? Ces notions mènent à la différence fondamentale entre les causes nécessaires et suffisantes que P. Naveau présentera analytiquement.

Avec le formalisme probabilistique et sous certaines hypothèses, la probabilité qu'une cause soit nécessaire définit le FAR (fraction de risque attribuable) et est donnée par la différence relative de deux probabilités (p1-p0)/p1 avec p0 la probabilité d'excéder un seuil dans un monde qui n'a pas été (sans forçage anthropique par exemple, c'est la probabilité contre-factuelle : l'évènement Y a lieu mais pas X) et p1 la probabilité d'excéder ce seuil dans le monde qui est (c'est la probabilité factuelle : Y et X ont lieu). On voit naître ici toute la difficulté du processus d'attribution quand le système est le climat terrestre (Figure 4).

P. Naveau nous a ensuite fournit un exemple concret d'application de cette méthode d'attribution (calcul du FAR) pour déterminer si l'influence de l'homme sur le climat est à l'origine de la vague de chaleur qui a frappé l'Europe en 2003. A partir de l'analyse des températures d'été au-dessus de l'Europe dans des simulations avec forçage historique HIST (hypothèse factuelle avec forçages x_1 et x_2) et forçage uniquement naturel NAT (hypothèse contre-factuelle avec forçage x_2 seul) Stott et co-auteurs (Nature, 2003) ont pu calculer p_0 et p_1 et donc le FAR. Ils ont conclu qu'il était très probable que l'influence de l'homme ait au moins doublé le risque d'une vague de chaleur de cette ampleur. Cependant il convient de prendre gare à l'interprétation de ces résultats car ils pointent vers une cause nécessaire mais en rien ne précisent si cette cause est également suffisante. Par expérience, il est d'ailleurs admis que pour les évènements extrêmes, plusieurs causes nécessaires sont souvent identifiées mais rarement de causes suffisantes. Nous retiendrons

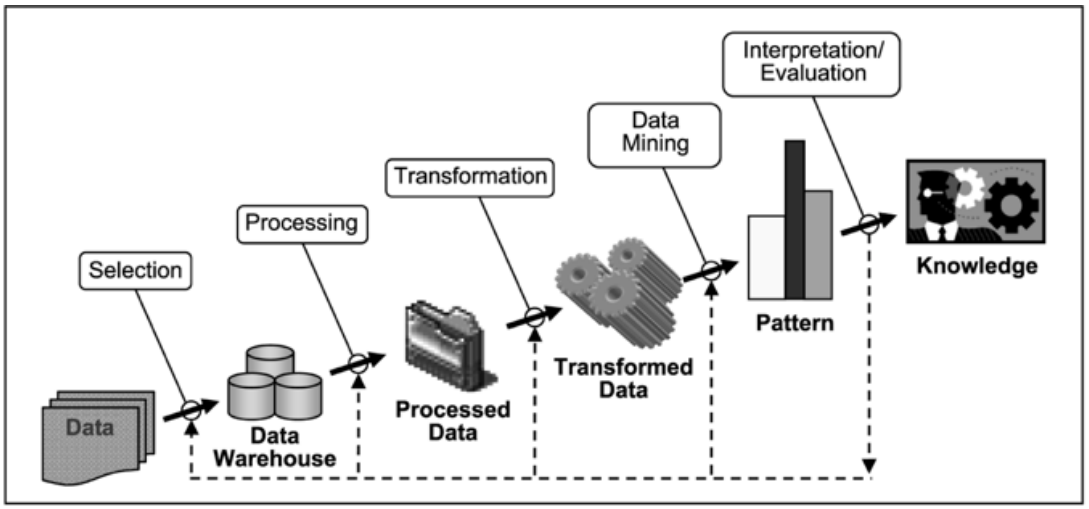

FIGURE 5: Représentation standard du processus KDD, découverte de connaissances dans une base de données

donc que des réponses précises sur les causes requièrent une définition tout aussi précises des questions posées. "Est-ce que les émissions de CO2 ont causé la vague de chaleur ?" n'amenant pas la même réponse que "est-ce que les émissions de CO2 ont augmenté le risque de vague de chaleur ?". P. Naveau conclura en discutant deux perspectives des caractères nécessaires et suffisants des causes. D'une part la perspective "ex post" (après l'évènement) cherche les causes à blâmer pour un évènement. C'est la position du juge (ou de l'assureur) pour qui les causes nécessaires sont importantes. D'autre part la perspective "ex ante" (avant l'évènement) identifie les causes à supprimer pour empêcher l'évènement d'avoir lieu. C'est la position du preneur de décision pour qui les causes suffisantes, resp. nécessaires, sont importantes car elles donnent le coût de l'inaction, resp. le bénéfice de l'inaction.

Nous retiendrons d'une part que les difficultés de détection sont en grande partie liées à notre manque de connaissance de la variabilité du système climatique terrestre à toutes ses échelles de temps et d'espace. Détecter les changements climatiques repose sur une meilleure observation et compréhension de la variabilité naturelle du système. Nous retiendrons également que pour attribuer des changements à des causes, il est nécessaire de parfaitement bien poser le vocabulaire et de distinguer les causalités nécessaires de celles suffisantes. Attribuer les changements climatiques repose sur notre capacité à décomposons et casser les chaînes de causalité et donc en grande partie à pouvoir générer des simulations numériques climatiques réalistes pour passer d'une approche probabiliste observationnelle à interventionniste.

3 Introduction à la fouille de données par l'exemple de l'analyse d'images en télédétection

Présentation donnée par Pierre Gançarski (iCube, Strasbourg)
Disponible en ligne : `http://goo.gl/4pUI1y`

La première partie du cours est une introduction à la discipline de "fouille" des données. Elle consiste à extraire des informations intéressantes (non-triviales, implicites, inconnues jusqu'alors, utiles) d'une collection de données. Pour cela, il y a deux types de méthodes : les descriptives et les prédictives, mais nous ne reviendrons pas ici sur les définitions déjà données par V. Kumar (voir ci-dessus). Nous retiendrons cependant que ces méthodes prennent place dans un processus de "découverte de connaissances dans les bases de données" (KDD/ Knowledge Discovery in Database) que l'on représente souvent sous la forme du schéma donné Figure 5 Représentation standard du processus KDD, découverte de connaissances dans une base de données.

Après cette introduction, P. Gançarski nous a décrit les caractéristiques des **images obtenues par télédétection** (aériennes et satellites) et pour lesquelles il nous montrera l'usage des méthodes de classification supervisées ou non. Nous retiendrons que le niveau d'analyse dépend de la résolution spatiale des images (la surface couverte par un pixel) qui va de 300 mètres à quelques dizaines de centimètres, et de la résolution spectrale (qui dépend du capteur) qui elle va du bleu à l'infra-rouge. L'interprétation d'une image doit par ailleurs être clairement définie. En effet il existe des différences entre l'interprétation visuelle des informations spectrales et l'interprétation sémantique des pixels. Comme la sémantique n'est pas toujours contenue explicitement dans l'image et dépend du contexte et de l'expertise de l'opérateur, on peut parler de **fossé sémantique**. Dans ce contexte, il s'agit du manque de concordance

entre l'information de bas-niveau (celle extraite automatiquement) et l'information de haut-niveau (celle de l'analyse de l'expert).

L'analyse d'une image peut se faire par pixel. Mais cette approche ne fonctionne bien qu'avec les images de basses résolutions. Avec la plupart des images, qui sont haute résolution, un pixel n'est très souvent qu'une partie d'un objet thématique. Il est donc nécessaire de faire appel à une analyse basée sur les objets (Object-Based Image Analysis, OBIA) plutôt que les pixels. Pour cela, il faut accomplir deux tâches : (i) segmenter l'image et (ii) caractériser les segments obtenus. Un objet, ou région, est la somme du segment et de ses caractéristiques. Mais des difficultés naissent du fait qu'il existe une multitude de possibilités pour segmenter une image et qu'il n'y a en général, ni de bonnes ni de mauvaises solutions, seulement des plus ou moins utiles suivant les objectifs. On retiendra de cette introduction le fait que l'analyse d'une image - par pixel ou par objet - fait appel au processus de classification automatique pour attribuer un label à chacun des pixels ou objets. C'est pourquoi la suite de la présentation porte en détail sur les méthodes de classification, dans une première partie supervisées puis non-supervisées.

3.1 Classification supervisée

La classification supervisée est une méthode de fouille prédictive qui consiste à construire un modèle qui à chaque donnée associe une catégorie - ou une classe, ou un label - à partir d'une base de donnée d'apprentissage qui contient déjà des associations entre données et classes. Une donnée (un élément de la base) peut avoir plusieurs attributs, chaque attributs a une liste possible de valeurs : ceux sont les catégories ou classe ou label de l'attribut. Prenons un exemple pour la suite. Les données sont des éléments à qui nous pouvons associer 3 attributs : le type de capteur, l'année de mise en service et un flag qualité. Les types de capteurs possibles sont rouge ou bleu, les années sont des entiers entre 2000 et 2014 et les flags qualités bon ou mauvais. Un modèle de classification supervisée permettra d'attribuer une catégorie (eg : bon ou mauvais) d'un attribut (eg : flag qualité) à une nouvelle donnée qui ne faisait pas partie de la base d'apprentissage. Les méthodes pour déterminer un tel modèle sont par exemple les machines à vecteurs de support (l'objet de la session de travaux dirigés menés par S. Canu, voir ci-après), les arbres de décisions, les réseaux de neurones, les k plus proches voisins etc...

L'**évaluation des performances** d'un tel modèle est un problème complexe que P. Gançarski nous présente ensuite. Pour mener cette évaluation, nous pouvons utiliser une **matrice de confusion**. Celle-ci consiste à recenser les différentes occurrences des vrai/fausse réponses du modèle et permet de calculer la justesse/fidélité/exactitude du modèle (accuracy en anglais) ainsi que la précision et le rappel du modèle pour une classe (precision and recall). Par exemple si la classe est une couleur, disons rouge ou bleu, la matrice de confusion est donnée par la Table 1.

		Classe prédite	
		rouge	bleu
Classe réelle	rouge	A	B
	bleu	C	D

TABLE 1: Exemple de matrice de confusion pour un problème à deux classes possibles rouge ou bleu. A est le nombre de données étant vraiment de classe rouge, B les faux rouges, C les faux bleus et donc D les vrais bleus.

Avec le nombre total de données classées N = A + B + C + D, la justesse du modèle est donnée par :

$$A_{cc} = \frac{A+D}{N} \quad (1)$$

c'est à dire simplement le ratio du nombre de données correctement classées (bonnes réponses) et du nombre total de données classées (toutes les réponses). Le lecteur remarquera la différence avec la précision du modèle sur une classe. Pour la classe rouge elle sera donnée par le ratio $\frac{A}{A+C}$, pour la classe bleue $\frac{D}{B+D}$; c'est à dire le nombre de données justement attribués à une classe par rapport au nombre total de données attribuées dans cette classe. Le rappel du modèle sur une classe sera en revanche calculée avec la somme sur les lignes de la matrice de confusion, c'est à dire que le rappel d'une classe est la fraction des résultats correctement prédits du nombre total réel d'éléments dans cette classe. Dans notre exemple pour la classe rouge, le rappel est $\frac{A}{A+B}$.

Le problème est que la justesse du modèle sur des données déjà classées n'indique en rien sa justesse sur la classification des données encore inconnues. Une alternative consiste donc à ne pas utiliser toutes les données dont on connaît la classe pour construire le modèle puis en estimer la justesse, c'est la **validation croisée**. On pourra donc diviser la base de données en : un jeu d'apprentissage (training set) qui servira à construire le modèle et un jeu de test (test set) sur lequel on évaluera le modèle. En poussant cette logique, on pourra également proposer de diviser la base

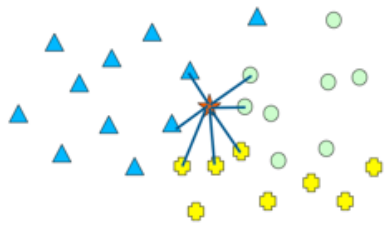

FIGURE 6: Les K plus proches voisins. A quelle classe appartient l'étoile dépend trop du K.

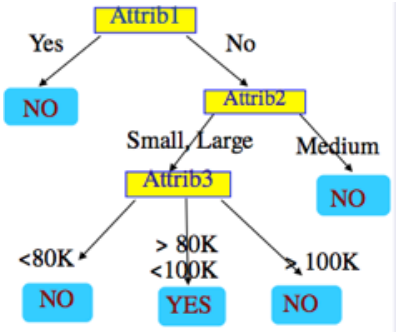

FIGURE 7: Arbre de décisions. Les nœuds (attributs) sont en jaune, les feuilles (classes) en bleu.

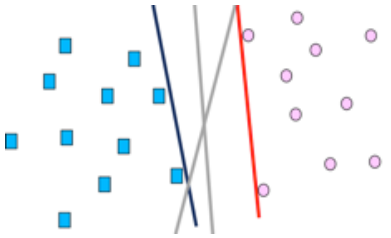

FIGURE 8: Hyperplans. Quelle est la meilleure séparation entre la classe bleue et la rose ?

Illustration de trois méthodes de classification supervisée

de données en K sous-ensembles, de construire le modèle sur K-1 sous-ensembles et de l'évaluer sur le sous-ensemble restant. On renouvelant cette opération plusieurs fois, on pourra sélectionner le modèle le plus juste.

Mais quelque soit la méthode, ce qui est déterminé par l'ancien n'indiquera rien de juste sur le nouveau si l'apprentissage s'est fait sur des données biaisées vers une classe (imaginer le cas extrême où les données sont quasiment toutes de la classe rouge et que les nouvelles données sont plutôt de classe bleue).

P. Gançarski nous présente ensuite plus en détails trois des méthodes permettant la construction d'un modèle dans une tâche de classification supervisée : celle des k plus proches voisins (K-nearest-neighbor), les arbres de décisions et les hyperplans.

3.1.1 Plus proches voisins La méthode des plus proches voisins, comme d'autres, repose sur l'hypothèse que plus deux données sont proches, plus la probabilité qu'elles appartiennent à la même classe est grande. Pour une donnée à classer, on calculera donc une distance entre elle et K voisins et déterminera sa classe en fonction de celle de ses voisins par une mesure de similarité. Les problèmes de cette méthode sont qu'elle ne construit pas de modèle explicite, son coût numérique est important, la trop grande influence du paramètre K et bien sûr qu'elle repose sur une mesure de similarité (souvent une simple distance euclidienne). On devine sur l'illustration Figure 6 ces difficultés.

3.1.2 Arbre de décision Les arbres de décisions consistent en une "hiérarchie de décisions" (voir illustration Figure 7) déterminées par une partitionnement récursif des données d'apprentissage. L'algorithme peut se décrire simplement de la manière suivante :

- initier la procédure en plaçant toutes les données dans un nœud d'origine, considérer ce nœud comme le nœud courant C
- choisir un attribut de test
- si C ne contient que des éléments d'une même classe, alors C devient une feuille labellisée avec cette classe,
- si C contient des éléments appartenant à plusieurs classes, utiliser l'attribut de test pour séparer les éléments en ensembles plus petits, qui seront chacun associés à un nœud fils
- appliquer cette procédure à tous les nœuds fils de manière récursive

Il y a bien sûr un certains nombres de problèmes avec cette méthode. En effet, comment choisir au mieux l'attribut de test ? cela peut dépendre du type d'attribut. Comment déterminer la meilleur séparation ? elles pourront dépendre de mesure de l'entropie ou de l'index de Gini.

3.1.3 Hyperplans Les méthodes basées sur les hyperplans cherchent à déterminer le meilleur plan qui séparera deux classes d'un attribut dans un jeux de données (voir illustration Figure 8). Comme ce sujet est au cœur du cours de S. Canu, nous ne retiendrons ici que les remarques générales non abordées par S. Canu. et qui concernent toutes les méthodes de classifications automatiques.

Plus particulièrement, le lecteur retiendra d'une part que si la séparation entre les classes est non linéaire (voir Figure 9), une projection dans un autre plan ou une transformation des données peut rendre le problème linéaire et plus facile à traiter. D'autre part le lecteur prendra garde aux notions de sous et sur-ajustement d'un modèle aux données (under et overfitting). En effet, dans le cas illustré en Figure 8, les deux classes sont parfaitement bien séparées dans le plan et il n'existera qu'une seule meilleure solution, linéaire. Mais dans les cas autrement plus complexes, la séparation entre les classes est moins claire. Prenons l'illustration donnée par P. Gançarski et représentée ici Figure 9. Sur la solution de gauche (panneau A), une simple droite est utilisée par séparer les deux classes. Cette solution linéaire est un cas de sous-ajustement où la frontière entre les deux classes n'est pas bien représentée et donnera des résultats flous

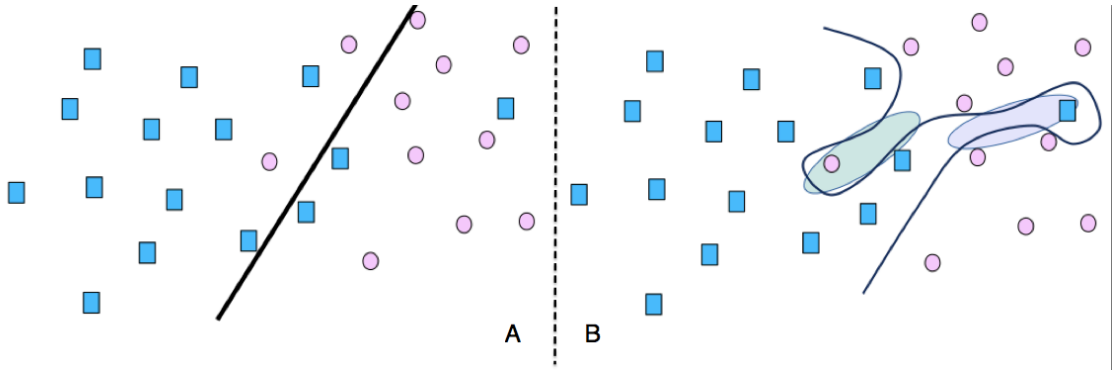

FIGURE 9: Illustration du cas de deux classes dont la séparation est non-linéaire. A/ Le plan (droite noire linéaire) est trop simple pour bien séparer les classes bleue et rose. C'est un cas de sous-ajustement aux données. B/ Le plan (courbe noire non linéaire) est très complexe. Même s'il sépare parfaitement les deux classes, il sera incapable de bien classer une nouvelle donnée apparaissant dans les régions verte ou bleue. C'est un cas de sur-ajustement aux données.

pour de nouvelles données. La situation inverse est illustrée sur le panneau B Figure 9. La séparation entre les deux classes est tellement non-linéaire qu'elle parvient à parfaitement séparer le plan. Mais c'est un cas typique de sur-ajustement où la solution passe bien par tous les points du problème mais n'a plus aucune capacité prédictive. Si une nouvelle donnée apparaît dans les régions grisées en vert et bleue, la classification sera très probablement mauvaise. James Faghmous, de l'université du Minesota et qui était présent à OBIDAM14, explique cette notion dans les termes suivants : **Pensez au sur-ajustement comme à une mémorisation, en opposition à un apprentissage** (Faghmous, 2013). C'est un peu comme pour un enfant qui pour apprendre à additionner apprendrait par cœur toutes les sommes d'entiers entre 1 et 5 mais ne pourrait pas calculer la somme de 6 avec 7 parce qu'il ne les a jamais vu.

Le sur-ajustement est de plus en plus probable au fur et à mesure que la complexité du modèle d'apprentissage augmente. Comment éviter cet écueil et déterminer la complexité du modèle à adopter ? On remarquera que l'erreur d'apprentissage, c'est à dire 100% soustrait de la justesse calculée avec l'ensemble d'apprentissage (voir Equation 1), va normalement diminuer et tendre vers 0 avec l'augmentation de la complexité du modèle. L'erreur sur la validation, c'est à dire 100% soustrait de la justesse calculée avec l'ensemble de test, va en revanche diminuer dans un premier temps mais finira par augmenter ou stagner quand la complexité du modèle lui fera perdre tout pouvoir prédictif. Le sur-ajustement intervient donc quand les courbes des deux erreurs commencent à diverger.

3.2 Classification non-supervisée

La classification supervisée présentée dans la section précédente repose sur une connaissance apriori des classes à attribuer aux éléments d'une base de données. En pratique, pour les sciences de l'environnement, ces classes sont difficiles à choisir et peuvent être très nombreuses. En effet, prenons l'image aérienne d'une ville, les classes possibles pour les éléments de l'image peuvent être les routes, habitations, plans d'eau, ruelles, véhicules, ombres, etc... Le nombre de classe possible dépend donc principalement de la quantité d'information contenu dans la base, ici la richesse de détail de l'image, donc sa résolution. Comment fournir assez d'exemple pour chaque classe quand il y en a énormément ? On touche ici à la limitation des méthodes d'apprentissage supervisées qui seront donc principalement utilisées pour les images à basse résolution ou avec peu de classes. Ces limitations mènent donc naturellement vers l'apprentissage non-supervisé qui va chercher à identifier la structure des données sans en avoir une connaissance apriori.

L'apprentissage non-supervisé est comme nous l'avons déjà indiqué, une méthode descriptive qui permet d'identifier des groupes d'objets de telle manière que les objets d'un groupe seront similaires (ou liés) tout en étant différents (ou non-liés) des objets des autres groupes. Les groupes identifiés, parfois appelés cluster, pourraient se voir attribuer une classe thématique par un expert, ce pourquoi l'apprentissage non-supervisé permet de faire une classification non-supervisée des données que l'on appelle parfois clustering. Les méthodes de clustering sont des méthodes de regroupement hiérarchique ou de partitionnement des données.

3.2.1 Méthode de partitionnement La plus connues des méthodes de partitionnement des données est le Kmoyens ou Kmeans en anglais. Etant donnée un jeu de données $(x_1, x_2, ..., x_n)$ où chaque donnée est un vecteur de dimension d, le partitionnement par Kmoyens cherche à séparer les n données en K ensembles S=$S_1, S_2, ..., S_K$ de manière à

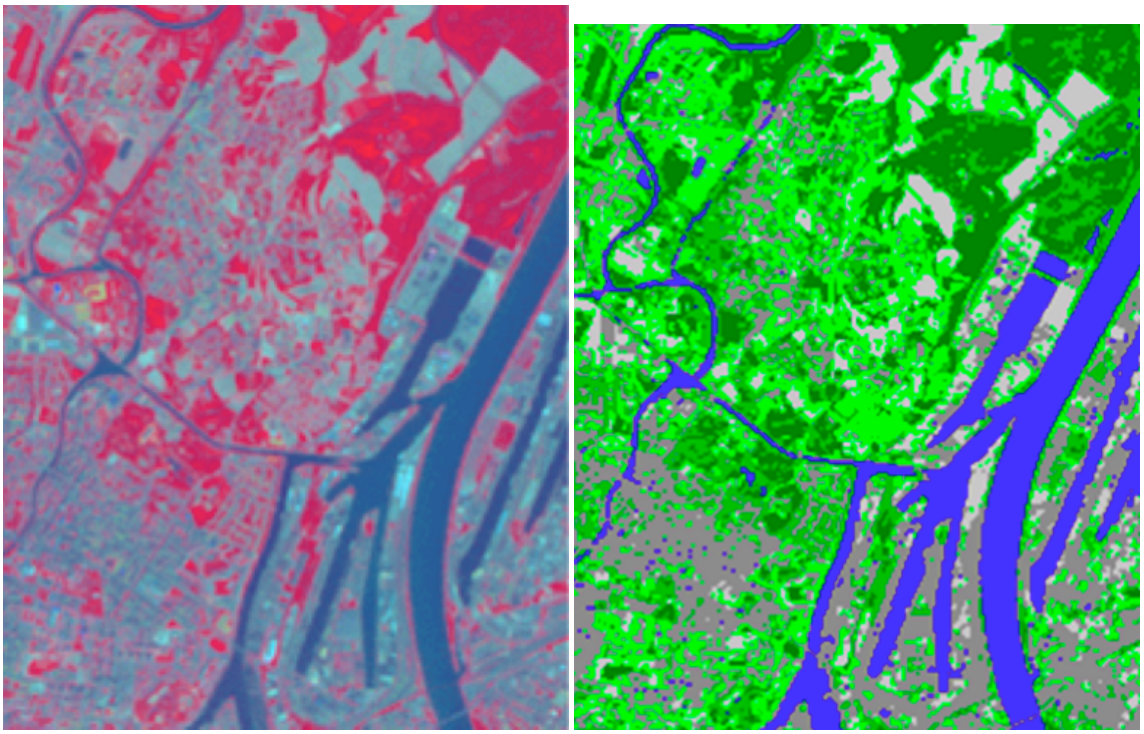

FIGURE 10: Exemple d'application du Kmeans sur une image SPOT. Gauche/A Image spot. Droite/B La même image avec 5 clusters colorés et identifiés par la méthode Kmeans.

minimiser la somme des carrés à l'intérieur des clusters :

$$\arg\min_{S} \quad = \quad \sum_{i=1}^{K} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2}$$

où μ_i est la moyenne des données de S_i. Il existe plusieurs algorithmes pour calculer cette minimisation et P. Gançarski nous montre simplement leur logique itérative en utilisant une image SPOT comme exemple, reproduite ici sur la Figure 10.

L'algorithme est le suivant :

1. choisir le nombre K des clusters
2. choisir au hasard K centroïdes dans l'espace des données
3. Itérer :
   - assigner chaque donnée au cluster dont le centroïde est le plus proche
   - recalculer la position des centroïdes à partir des données qui lui sont assignées
4. Répéter l'étape 3 jusqu'à ce que les centroïdes ne bougent plus.

Bien sûr derrière une telle simplicité se cache des difficultés certaines liées au choix du nombre de clusters. Kmeans souffre également d'une grande sensibilité aux outliers et aux conditions initiales (l'algorithme identifie un minimum local mais sans la garantie que ce soit le minimum global) tout en étant limité aux clusters ayant des géométries convexes. Malgré tout, Kmeans reste une méthode très employée à cause de sa simplicité de mise en œuvre et de son efficacité pour les jeux de données relativement bien structurés.

3.2.2 Méthode de regroupement hiérarchique Les méthodes de regroupement hiérarchique peuvent être bottom-up (logique agglomérative) ou top-down (logique divisive). Le cas bottom-up, ou classification ascendante hiérarchique (ou CAH), est illustré sur les Figures 11-12. Il s'agit de fusionner les pairs les plus similaires et de s'arrêter quand toutes les données ont été fusionnées dans un seul cluster. A l'état initial, chaque donnée représente donc un cluster et à l'état final on dispose d'une hiérarchie de clusters permettant a posteriori de choisir un nombre de cluster approprié. L'avantage de cette méthode est qu'elle ne fait pas intervenir de moyennes, en revanche elle est très limitée pour de grandes bases de données car il est nécessaire de calculer les distances entre toutes les pairs possibles, le coût de calcul est donc important.

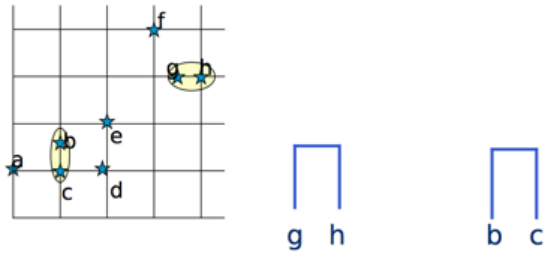

FIGURE 11: Deuxième étape de la CAH. Les données b,c et g,h ont été fusionnées parce qu'elles étaient les plus proches

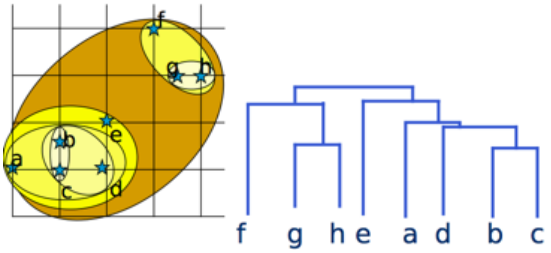

FIGURE 12: La CAH est achevée. La représentation hiérarchique permet a posteriori de choisir un nombre de clusters pertinents pour séparer les données. Ici, k=2 paraît intéressant. Le 1er cluster a les données f, g et h ; le 2nd les données e, a, d, b et c.

Illustration de la classification ascendante hiérarchique.

4 Machines à vecteurs supports et noyaux : classification linéaire et non-linéaire

Présentation donnée par Stéphane Canu (LITIS, INSA Rouen, Université de Normandie)
Disponible en ligne : http://goo.gl/91gfep

Parmi les méthodes de classification supervisée on trouve les machines à vecteurs de support (SVM). Dans cette session d'OBIDAM14, S. Canu nous a présenté mathématiquement les SVMs en détail et nous a permis de les tester pendant les séances de travaux dirigés. Le lecteur trouvera donc dans ce compte-rendu une présentation simplifiée et sans détails analytiques des SVM. Les machines à vecteurs de support sont aussi appelées les séparateurs à vaste marge. Elles sont de nature prédictive et s'utilisent donc dans les situations où l'on cherche à modéliser la séparation entre deux classes dans un jeu de données labellisées (en maximisant la marge avec les données les plus proches, nommées vecteurs supports). Nous l'avons déjà illustré en Figure 8 (cas linéaire) et Figure 9 (cas non-linéaire), mais nous pouvons reprendre l'exemple plus concret fourni par S. Canu et donné ici Figures 13-14-15.

Les SVMs reposent sur deux principes qui permettront de traiter les problèmes de classification non-linéaire et de reformuler le problème pour qu'il devienne un problème d'optimisation quadratique.

4.1 Marge maximale

Le premier principe est celui de marge maximale, illustré Figure 16. Comme il existe une multitude de droites pouvant séparer deux classes, il faut un critère pour en sélectionner une. Celui-ci est basé sur la performance prédictive de la séparatrice, c'est à dire sa capacité à généraliser la classification. On peut montrer qu'il existe une seule et unique solution à cette optimisation : **la ligne de décision optimale, ou frontière de séparation, est celle qui maximise la marge**. La marge est la distance minimum entre la séparatrice et les données, là encore on pourra montrer que pour déterminer la marge maximale, seules les données les plus proches de la séparatrice jouent un rôle ; on les appelle **vecteurs supports** (d'où le nom de la méthode). Les autres données n'interviennent pas dans la détermination, ce qui a son importance pour le coût numérique de calcul avec de grandes bases de données. La détermination de l'hyperplan/séparatrice est donc bien sûr un problème d'optimisation où il s'agira de chercher le minimum d'une fonction sous certaines contraintes. Ici la fonction est la distance des vecteurs supports à la séparatrice, c'est donc une fonction quadratique et pas linéaire. Ainsi, le problème sera résolu en le reformulant mathématiquement comme une optimisation quadratique (QP) que l'on pourra résoudre avec des méthodes classiques type multiplicateurs de Lagrange. Comme tout problème classique d'optimisation quadratique, il peut se présenter sous 2 perspectives, équivalentes : les formes primale et duale pour lesquels des solveurs numériques existent et seront utilisés pendant la séance de travaux dirigés. S. Canu présentera les conditions d'optimalité liées à la convexité du problème. Le lecteur est donc renvoyé à sa présentation pour la formulation mathématique de ces problèmes. Nous retiendrons ici que la formulation primale sera particulièrement adaptée quand le nombre de données sera beaucoup plus grand que le nombre de dimensions de chaque donnée ; la formulation duale sera adaptée à la situation inverse.

Nous aborderons également le cas où les données sont bruitées, qui conduit au cas dit non séparable du problème, illustré Figure 17. Il s'agit fondamentalement de pouvoir tenir compte d'erreurs potentielles dans les données qui, sans être prises en compte, empêcheraient strictement la détermination de la séparatrice optimale. Pour modéliser ces erreurs, on introduit des variables dites molles (slack variables en anglais) pour chacune des données du problème. Cette nouvelle dérivation des formulations primale et duale (qui ont maintenant 2 fois plus de contraintes) nous mènera

FIGURE 13: L'objectif ici est de déterminer si les objets identifiés sur l'image (technique OBIA, cf section précédente) sont des bateaux ou des maisons. Crédit : A Gentle Introduction to Support Vector Machines in Biomedicine A. Statnikov, D. Hardin, I. Guyon and C. F. Aliferis

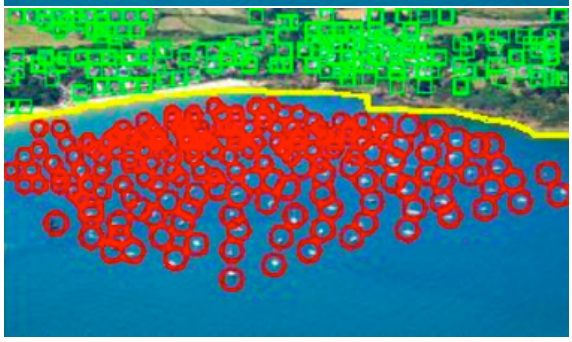

FIGURE 14: Une base de données est construite. Avec une liste d'objets ayant chacun des coordonnées et une classe rouge/bateau, vert/maison.

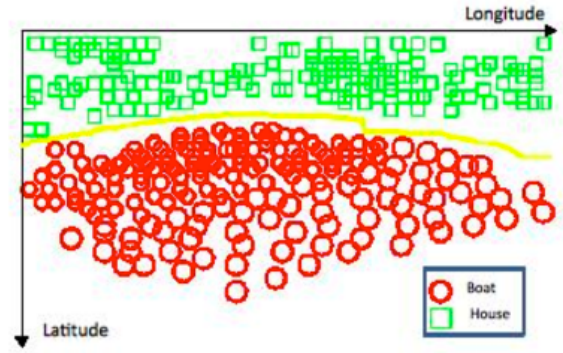

FIGURE 15: A partir des exemples déjà labellisés, on cherche à déterminer la règle de décision (courbe jaune) qui permettra de classer un nouvel objet sur l'image.

vers la mesure de perte de Hinge. En effet, prendre en compte le bruit sur les données et le minimiser dans le problème QP n'implique pas que la séparatrice soit parfaite. Il faut donc pouvoir mesurer la performance de la classification et pour cela nous pourrons utiliser la perte de Hinge qui tend vers 0 pour une bonne prédiction du classifieur et vers 1 sinon.

4.2 Fonction noyau

On voit bien la limite de la prise en compte d'erreurs sur les variables dans le cas non séparable du problème : il existe bien sûr des distributions de données pour lesquelles la séparatrice ne peut pas être une droite (on dit non-linéairement séparable). Le second principe sur lequel s'appuie les SVMs consiste à changer d'espace de représentation des données pour un espace, peut-être de dimension supérieure, dans lequel le problème sera linéairement séparable. Un exemple simple est donné sur la Figure 18 où l'on peut voir deux classes qui ne sont pas linéairement séparables dans l'espace cartésien, mais qui le seront dans l'espace de coordonnées polaires. Mais les fonctions pouvant exprimer cette transformation peuvent être beaucoup plus complexes et si on s'en tenait là, la solution de l'hyperplan ferait intervenir le produit scalaire du nouvel espace, qui possiblement est à très grandes dimensions, une solution coûteuse numériquement. Pour lever ces limitations, les SVMs s'appuient sur un principe connu sous le nom de kernel trick qui consiste à utiliser une fonction noyau pour exprimer la transformation non-linéaire de l'espace d'origine vers l'espace de redescription. L'intérêt de la fonction noyau est qu'elle fait disparaître de la solution de l'hyperplan séparateur la mention explicite de la fonction complexe de la transformation d'un espace à l'autre.

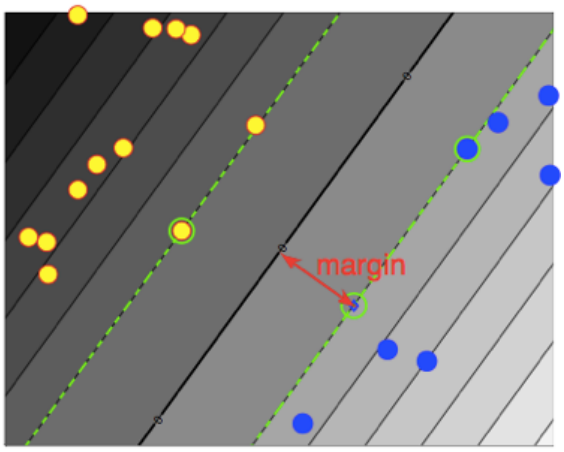

FIGURE 16: La marge. Illustration pour le cas séparable où toutes les données des deux classes peuvent être séparées par une droite (linéairement séparable).

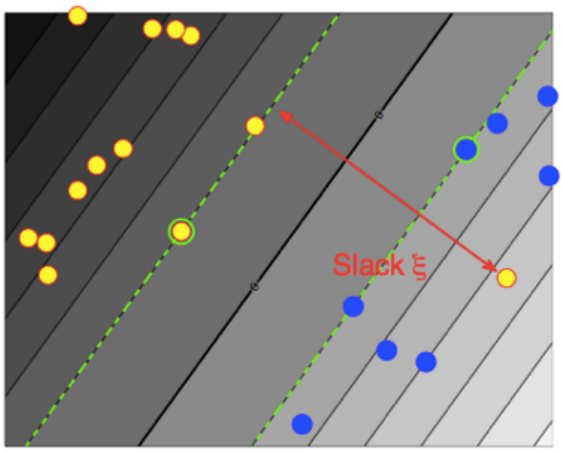

FIGURE 17: Variable molle. Distribution de données où certaines sont bruitées, donnant naissance au cas dit ”non séparable” de SVM.

Principe de séparateur à vaste marge, ou SVM. Les données sont les points, de classe jaune ou bleue ; les vecteurs de support sont entourés en vert. Il s'agit de déterminer la séparatrice (courbe noir) qui maximisera la marge avec les vecteurs supports.

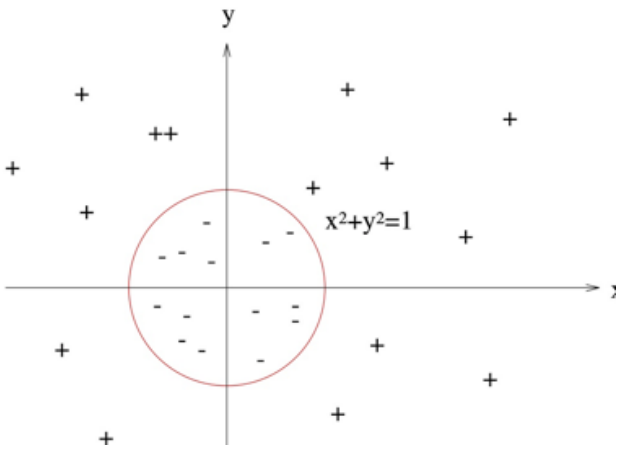

FIGURE 18: Exemple de transformation d'un cas non séparable (coordonnées cartésiennes) en cas séparable (coordonnées polaires). C'est un exemple trivial où les deux espaces ont la même dimension.

Références

Faghmous, J., M. Le, M. Uluyol, V. Kumar, and S. Chatterjee, 2013 : A parameter-free spatio-temporal pattern mining model to catalog global ocean dynamics. *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 151–160, doi :10.1109/ICDM.2013.162.

Faghmous, J. H., 2013 : New to machine learning ? avoid these three mistakes, URL https://t.co/Qm9iM7ipzw.

Hagmayer, Y., S. A. Sloman, D. A. Lagnado, and M. R. Waldmann, 2007 : Causal reasoning through intervention. *Causal learning : Psychology, philosophy, and computation*, 86–100, URL http://else.econ.ucl.ac.uk/papers/uploaded/199.pdf.

Hume, D., 1748 : *An enquiry concerning human understanding*. Broadview Press.

Kawale, J., M. Steinbach, and V. Kumar, 2011a : *Discovering Dynamic Dipoles in Climate Data*. Society for Industrial and Applied Mathematics, doi :doi:10.1137/1.9781611972818.10, URL http://dx.doi.org/10.1137/1.9781611972818.10.

Kawale, J., et al., 2011b : *Data Guided Discovery of Dynamic Climate Dipoles.* 30-44 pp.

Kawale, J., et al., 2013 : A graph-based approach to find teleconnections in climate data. *Statistical Analysis and Data Mining*, **6 (3)**, 158–179, doi :10.1002/sam.11181, URL http://dx.doi.org/10.1002/sam.11181.

Pearl, J., 2000 : *Causality : models, reasoning and inference*, Vol. 29. Cambridge Univ Press, URL http://bayes.cs.ucla.edu/BOOK-99/book-toc.html.