Présentation de la mise en place d'une analyse par cartes auto-organisatrices

Valentin Daniel

Chaire maritime

Chaire de recherche sur les dynamiques des activités humaines en mer et la planification de l'espace maritime

Fiche thématique n°3

Présentation de la mise en place d'une analyse par cartes auto-organisatrices

Valentin Daniel

Nantes Université

Fondation Nantes Université

Préambule

Tout comme la fiche précédente, cette fiche a pour but de présenter les éléments techniques de la méthode de classification de l'économie maritime. Elle sera complétée par la fiche 4 qui détaillera les outils d'interprétation avancés.

Ainsi la lecture de cette fiche n'est pas obligatoire pour la compréhension de l'outil. Cependant la mise en ligne d'un tel document est importante dans un souci d'éclairage, d'autant plus lorsque l'application directe des méthodes présentée ici fait appel à la big data et aux données socio-économiques.

Enfin la Chaire se positionnant sur l'expérimentation de nouvelles méthodologies, l'explication de la démarche et des aspects techniques revêt une importance particulière, en vue notamment de s'assurer de la reproductibilité des travaux.

Principes généraux

VOCABULAIRE SPECIFIQUE

Puisque les réseaux de Kohonen sont issus des théories du machine learning il convient avant tout de définir le vocabulaire spécifique qui sera utilisé dans cette partie.

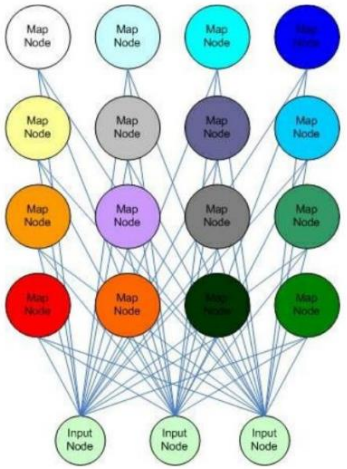
- Un nœud est un élément de structure des cartes
- Les poids associés aux nœuds correspondent aux valeurs des variables dans ces nœuds
- Un vecteur de poids est donc un ensemble de valeurs de variable
- Un vecteur d'input (ou nœud d'input) correspond simplement à une observation au sens commun du traitement des bases de données

Cette fiche se basera majoritairement sur l'article explicatif des cartes de Kohonen de Guthikonda (2005), afin d'illustrer les concepts théoriques décrits ici.

STRUCTURE DES CARTES

Les cartes auto-organisatrices présentent certaines restrictions, les variables retenues pour la formation du modèle doivent toutes être de nature quantitative et subir une transformation statistique (centrée réduite). Dans le cadre de l'application que nous verrons par la suite, aucune variable qualitative n'avait besoin d'être retenue, c'est l'une des raisons pour lesquelles la méthode des SOM a pu être envisagée. Des variables qualitatives peuvent tout de même être intégrées dans les cartes si elles sont transformées au préalable, par exemple par la méthode du one-hot-encoding (T. Montzka, 2018). Le one-hot-encoding consiste à créer pour chaque modalité une variable dichotomique, prenant la valeur de 0 ou de 1.

La structure des cartes auto-organisatrices est assez simple, elle peut se schématiser de la manière suivante :



La structure présentée ici est un réseau de 4 par 4, soit 16 nœuds, chacun d'entre eux est relié à chacun des 3 nœuds d'input. Pour cette carte on a donc 16x3 = 48 connections.

Principes généraux

Les nœuds de la carte ne sont quant à eux pas connectés entre eux, la manière dont ils sont organisés permet entre autres d'obtenir une grille en deux dimensions afin d'avoir une meilleure visualisation des résultats. Chaque nœud aura une coordonnée unique associée. Cela permet de les identifier et surtout de calculer la distance entre chacun d'eux et donc par la suite d'en regrouper certains si nécessaire. Cependant puisque les nœuds ne sont pas connectés, la valeur du nœud voisin n'aura pas d'impact direct. Une carte (aussi appelée réseau) est donc constituée des nœuds de carte, un nœud de carte contient quant à lui un poids et ses coordonnées dans le réseau, les nœuds d'inputs contiennent également leurs poids respectifs. Les vecteurs de poids doivent être les mêmes entres les nœuds de la carte et ceux d'input pour que l'algorithme fonctionne correctement.

PRINCIPE ALGORITHMIQUE DES CARTES AUTO-ORGANISATRICES

L'algorithme des cartes auto-organisatrices peut se décomposer en 6 étapes (S. Guthikonda 2005). Certaines d'entre elles seront détaillées par la suite dans un exemple d'application.

- Les poids associés à chaque nœud sont initialisés.
- Un vecteur d'input (une observation et les variables associées) est exposé au réseau.
- On cherche le nœud du réseau dont les poids correspondent le plus au vecteur d'input. Le nœud sélectionné est appelé BMU (Best Matching Unit)
- Un rayon de voisinage est calculé pour le BMU. Au début du processus le voisinage concerne tout le réseau puis diminue petit à petit.
- Les nœud se trouvant dans le rayon de voisinage du BMU sont réajustés pour correspondre plus fidèlement au vecteur d'input. Plus un nœud du voisinage est proche du BMU, plus son poids sera modifié pour correspondre à l'input.
- L'opération est répétée pour chaque observation. Cela mène à un ajustement permanent du réseau.

Cet algorithme nécessite l'utilisation d'un ensemble de fonction.

Ensemble de fonctions

1. CALCUL DE LA « BEST MATCHING UNIT »

$$InputDistance^2 = \sum_{i=0}^{i=n} (I_i - W_i)^2$$

Avec :

- I : le vecteur d'input utilisé à ce moment dans l'algorithme
- W : le vecteur de poids du nœud considéré
- n : le nombres de poids

Cette équation permet de calculer à partir des variables du modèle, une distance Euclidienne élevée au carré entre l'input et le nœud.

2. CALCUL DU RAYON DE VOISINAGE

$$\sigma(t) = \sigma_0 e^{(-t/\lambda)}$$

Avec :

t : le numéro de l'itération en cours

λ : la constante de temps (*c.f équation 3*)

σ_0 : le rayon total de la carte

Le rayon de voisinage décroit avec le nombre d'itérations. Plus celui-ci sera faible, plus le réseau aura appris et moins on aura besoin d'affecter de nœuds autour du BMU pour en améliorer la qualité. Au départ ce rayon doit être quasiment équivalent à celui de la carte dans son ensemble, alors qu'en fin de processus, le rayon du BMU se trouve être le BMU lui-même.

Ensemble de fonction

3. CALCUL DE LA CONSTANTE DE TEMPS

$$\lambda = T/\sigma_0$$

Avec :

T : le nombre total d'itérations

σ_0 : le rayon total de la carte

4. CALCUL DES NOUVELLES VALEURS DES POIDS D'UN NOEUD

$$W(t+1) = W(t) + \Theta(t)L(t)(I(t) - W(t))$$

Avec :

W : le vecteur de poids du nœud considéré

I : le vecteur d'input utilisé à ce moment dans l'algorithme

L : le ratio d'apprentissage (c.f équation 5)

Θ : appelé « influence », il s'agît d'un indicateur de distance entre le nœud considéré et le BMU (c.f équation 6)

Cette fonction est celle qui va définir l'apprentissage dans le réseau, elle permet de recalculer la valeur des poids pour le nœud traité. Au fil des itérations, de moins en moins de nœuds seront sélectionnés, car le rayon de voisinage va se réduire, mais lorsqu'un nœud sera traité, il apprendra d'autant plus que la valeur de ses poids sera différente de celle des poids du vecteur d'input. Autrement dit les valeurs des poids du nœud seront plus fortement modifiées.

Ensemble de fonction

5. CALCUL DU RATIO D'APPRENTISSAGE

$$L(t) = L_0 e^{(-\frac{t}{\lambda})}$$

Avec :

t : le numéro de l'itération en cours

λ : la constante de temps (c.f équation 3)

L_0 : la valeur choisie pour le ratio d'apprentissage initial. Le ratio d'apprentissage, décroit avec le nombre d'itérations de la même manière que le rayon de voisinage. La valeur optimale de L_0 est sujet à débat. Dans la pratique, elle est souvent laissée par défaut dans les logiciels.

6. CALCUL DE L'INFLUENCE

$$\Theta(t) = e^{-\frac{d^2}{2\sigma^2(t)}}$$

Avec :

d : le nombres de nœuds entre celui considéré et le BMU.

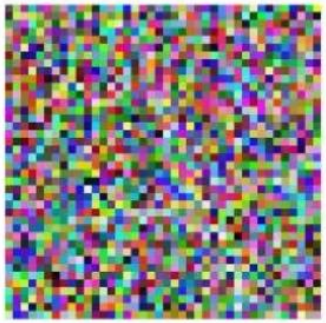
En prenant en compte la distance entre le nœud et le BMU, cet indicateur permet de pondérer l'apprentissage du nœud en question. Plus le nœud est proche, plus il apprendra. Bien entendu, si il se trouve en dehors du rayon de voisinage alors le nœud ne sera pas considéré du tout. Afin de mieux comprendre le fonctionnement réel de l'algorithme, il convient d'en présenter un exemple d'application simple avec un jeu de données de faible envergure.

Exemple d'application : classification des couleurs

INITIALISATION DU RESEAU

Cet exemple est tiré de l'article sur les cartes auto-organisatrice de Shyam M. Guthikonda (2005) et porte sur la classification des couleurs. La classification des couleurs par les cartes de Kohonen est l'exemple le plus utilisé pour introduire cette méthode. En effet, en colorant les nœuds aux couleurs associées la carte deviendra très visuelle et parlante même pour les non-initiés.

Pour effectuer les classifications des couleurs, trois poids seulement seront utilisés pour les nœuds : les valeurs rouge, vert et bleu. Par exemple l'observation correspondant au rouge aura comme vecteur de poids w = (1,0,0). En reprenant l'idée de base d'une carte nous devrions avoir le rouge, le bordeaux et le rose dans des zones proches. La première étape de l'algorithme est illustrée par le graphique ci-dessous, le réseau est initialisé, chaque carré représente en fait un nœud du réseau, un vecteur de poids leur est assigné de manière aléatoire (il en résulte donc une couleur).



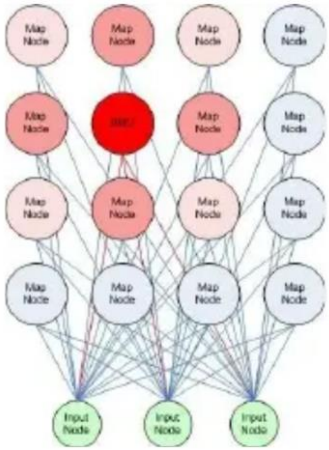
On remarque ici que la disposition des couleurs est complètement aléatoire. Les étapes suivantes vont contribuer à harmoniser l'ensemble. Lors de la deuxième étape, un vecteur d'inputs est sélectionné. Dans cet exemple il y a 8 vecteurs qui correspondent à 8 couleurs. Puis chaque nœud est parcouru pour trouver celui dont les poids correspondent le plus au vecteur.

Pour faire un parallèle avec la classification de l'économie, il faudrait voir ici chaque petit pixel comme une entreprise fictive. On obtient donc une économie fictive à laquelle on va ajouter nos vraies entreprises, de manière à ce qu'elles soient placées à côté des entreprises fictives qui leur ressemblent le plus. Puis cet ajout va modifier les valeurs des entreprises fictives pour qu'elles correspondent davantage aux entreprises réelles. A la fin de la modélisation les entreprises fictives ne sont plus présentes dans les données.

Exemple d'application : classification des couleurs

RAYON DE VOISINAGE ET INTENSITE D'APPRENTISSAGE

La figure suivante montre la sélection du BMU et l'identification du voisinage. Cette étape, répétée plusieurs fois, est très importante puisqu'elle va modeler la carte, et donc structurer de plus en plus la représentation des classes :

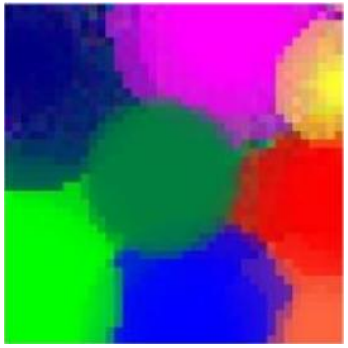


Le BMU est ici en rouge, les nœuds du voisinage sont eux colorés en rouge plus ou moins clair en fonction de leur distance. Tous ces nœuds voient leur vecteur de poids respectif impacté par la fonction d'apprentissage (c.f équation 4). Plus un nœud est proche du BMU plus il sera impacté, les autres nœuds n'étant pas modifiés. Enfin l'opération est répétée pour chaque vecteur d'input, jusqu'à obtenir la carte finale.

Exemple d'application : Résultats

CARTE FINALE

La carte finale obtenue regroupe les couleurs proches entres elles :



On obtient ici une structure de notre palette de couleur, tout comme on pourrait obtenir une structure de notre économie. On observe que le résultat est plutôt bon, le rouge, l'orange, le jaune et le rose forme la partie droite de la carte alors que la partie gauche est représentée par les teinte bleu-vert. La lecture de ce type de résultat est assez instinctive puisqu'on peut rapidement faire le parallèle avec une carte géographique, on y voit la « région » des teintes rouge, des teintes vertes etc…

Cependant la représentation ne parait pas optimale, en effet le bleu clair et le bleu foncé sont assez éloignés.

Cela peut être dû à plusieurs phénomènes. Cependant, à ce stade la principale piste de réponse est la suivante : visuellement les couleurs sont proches, mais en réalité les valeurs RGB qui les composent sont légèrement éloignées.

Ce genre de résultats peut également se retrouver lors de l'analyse de phénomène complexe. Il convient alors de croiser les résultats de la carte finale avec ceux de cartes intermédiaires. Celles-ci aideront alors à comprendre en profondeur la structure du phénomène étudié et les raisons de l'éloignement de certaines « régions ».

A l'aide de cet exemple simple nous avons donc vu qu'il était possible d'agencer un ensemble d'agent en fonction de caractéristiques multiples. La fiche suivante s'attachera donc à présenter les méthodes d'analyses basées sur ces mêmes cartes auto-organisatrices, permettant d'approfondir la connaissance du phénomène étudié.