




# LES CARTES DE KOHONEN

## PLAN DU COURS

- ▶ Introduction
  - ▶ Notations et définitions
  - ▶ Carte topologique
  - ▶ Algorithme de Kohonen
  - ▶ Carte de Kohonen et classification
  - ▶ Exemple
- 

## Introduction

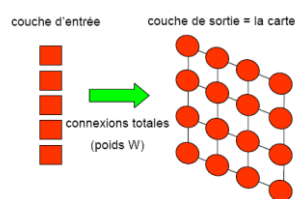
- ▶ Dans les méthodes non supervisées, nous avons vu la CAH, les k-means
- ▶ Ce sont les méthodes les plus connues
- ▶ Il existe d'autres méthodes :
  - ▶ Modèles de mélange
  - ▶ Cartes de Kohonen
- ▶ Les cartes de Kohonen
  - ▶ cartes topologiques auto-organisatrices
  - ▶ Self Organised Map (SOM T.Kohonen 1981)
- ▶ Quand les utiliser ?
  - ▶ Utilisées dans un but descriptif : comprendre et analyser la structure des données multidimensionnelles

## Introduction

- ▶ Carte auto-organisatrice ou carte de Kohonen
  - ▶ Algorithme de classification développé par Teuvo Kohonen dès 1982
- ▶ Propriétés
  - ▶ Apprentissage non supervisé
  - ▶ Méthode de quantification vectorielle
  - ▶ Regroupement des informations en classes tout en respectant la topologie de l'espace des observations
    - ▶ définition a priori d'une notion de voisinage entre classes
    - ▶ des observations voisines dans l'espace des données appartiennent après classement à la même classe ou à des classes voisines
    - ▶ compression de données multidimensionnelles tout en préservant leurs caractéristiques

## Carte de Kohonen

- ▶ Deux espaces indépendants
  - ▶ l'espace des données généralement de grande dimension
  - ▶ l'espace des représentations (la carte) de dimension réduite
    - ▶ les noeuds de la carte sont disposés géométriquement selon une topologie fixée a priori
  - ▶ Trouver la projection entre les deux espaces
    - ▶ la projection doit conserver la topologie des données
- ▶ Un noeud dans la carte possède des
  - ▶ coordonnées *fixes* sur la carte
  - ▶ coordonnées *adaptables*  $W$  dans l'espace d'entrée original

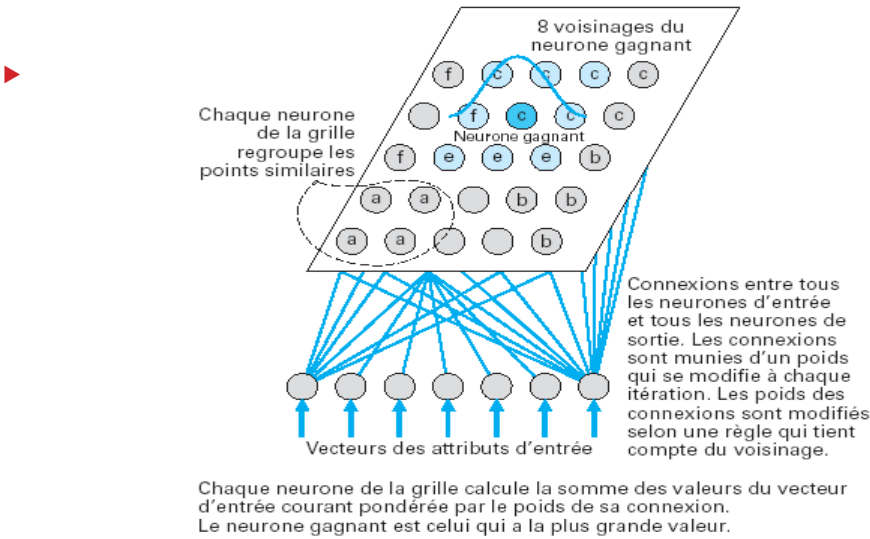


## Autre définition

- ▶ Transposer l'idée à un système **d'apprentissage non supervisé** compétitif où l'espace d'entrée est « mappée » dans un **espace réduit** (souvent rectangulaire) avec un principe fort :  
des individus similaires dans l'espace initial seront projetés dans le même **neurone** ou tout du moins dans **deux neurones proches** dans l'espace de sortie (préservation des proximités).

Sert à la fois pour la **réduction de la dimensionnalité**, la **visualisation** et la **classification automatique** (clustering, apprentissage non supervisé).

## Carte de Kohonen

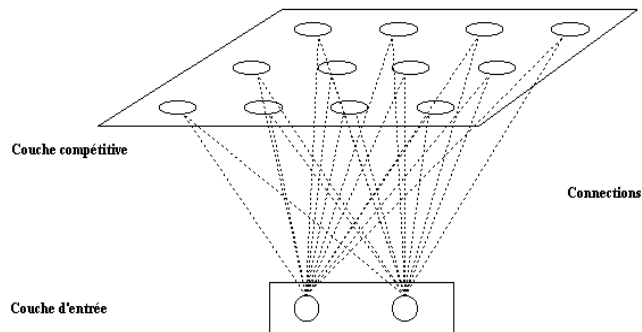


## Cartes de Kohonen

- ▶ Méthode basée sur les réseaux de neurones
- ▶ Processus incrémental d'auto organisation qui cherche à projeter des données dans un espace de faible dimension = carte topologique
- ▶ Notion de distance et de voisinage

## Architecture neuronale

### ► Réseau à 2 couches entièrement connectés



## Architecture neuronale (2)

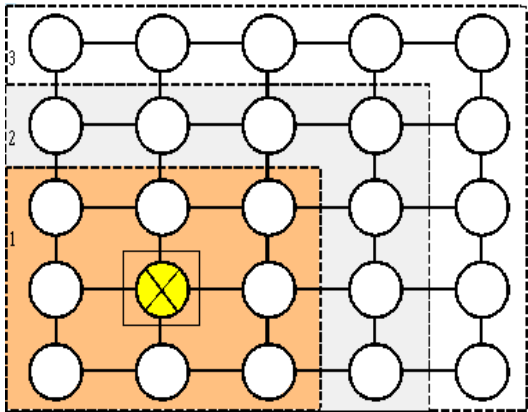
- couche d'entrée à  $p$  neurones : présentation des observations de dim  $p$  – les données
- couche d'adaptation = treillis des neurones = carte
- vecteur référent  $w_c$  associé à un neurone  $c$  de la carte  
= vecteur des connexions (ou vecteur de poids) qui arrive au neurone  $c$ .

## Structure topologique

- ▶ Les voisinages entre classes peuvent être choisis de manière variée, mais en général on suppose que les classes sont disposées sur une grille rectangulaire qui définit naturellement les voisins de chaque classe.
- ▶ Mais on peut choisir une autre topologie.

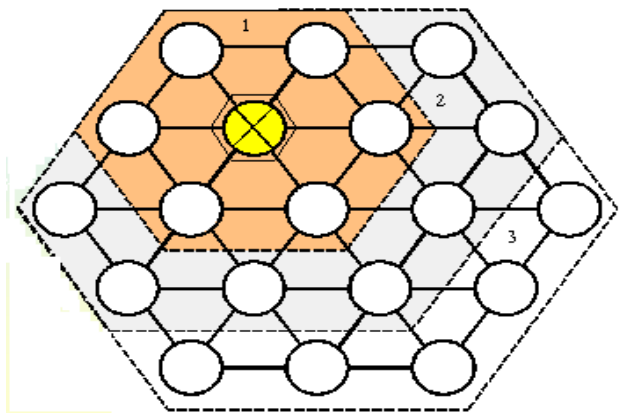
## Structure topologique

- ▶ Grille Rectangulaire



## Structure topologique

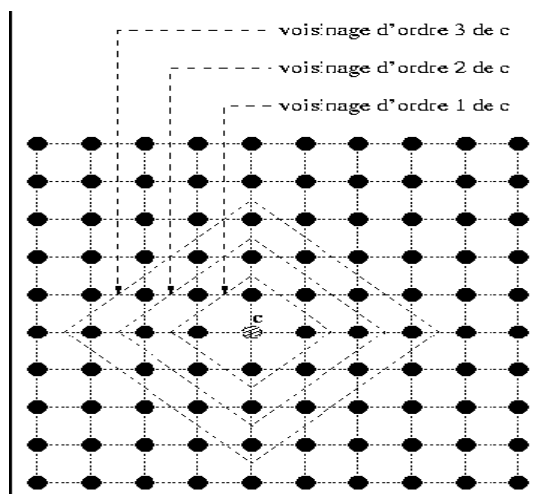
► Grille Hexagonale



## Distance sur la carte

- Carte = ensemble de neurone liés par une structure de graphe non orienté
- $\delta(c, r)$  distance discrète sur le graphe : le plus court chemin entre les neurones  $c$  et  $r$
- Pour chaque neurone  $c$ , cette distance discrète permet de définir la notion de voisinage d'ordre  $d$  de  $c$

## Voisinage topologique



## Principe général de l'algorithme de Kohonen

- ▶ algorithme original de classification qui a été défini par T. Kohonen, dans les années 80.
- ▶ Apprentissage non supervisé
- ▶ regroupement des observations en classes
- ▶ auto-organisation, respect de la topologie de l'espace des observations
- ▶ les réponses associées à des entrées voisines sont voisines



## Principe général de l'algorithme de Kohonen

- ▶ Nombreuses applications
  - ▶ En représentation de données de grande dimension sur des réseaux de dimension 1 ou 2
  - ▶ En classification où la notion de classes voisines a un sens

## Apprentissage de la carte

- ▶ L'apprentissage met en correspondance l'espace des entrées et la carte
  - ▶ Adaptation des poids  $W$  de telle manière que des exemples proches dans l'espace d'entrée sont associés au même neurone ou à des neurones proches dans la carte

## Algorithme

- Initialisation aléatoire des poids  $W$
- A chaque itération  $t$ 
  1. Présentation d'un exemple d'apprentissage  $X(t)$ , choisi au hasard, à l'entrée de la carte
  2. Comparaison de l'exemple à tous les vecteurs poids, le neurone gagnant  $j^*$  est celui dont le vecteur poids est le plus proche de l'entrée  $X(t)$  (*phase de compétition*)
  3. Évaluation du voisinage du neurone gagnant dans la carte
  4. Mise à jour des poids pour tous les neurones de la carte, l'adaptation est d'autant plus forte que les neurones sont voisins de  $j^*$  (*phase de coopération*)

## Algorithme - détails

- Phase de complétion :
  - On utilise comme distance (2) :  $d_N(X(t), W_{j^*}(t)) = \min_j d_N(X(t), W_j(t))$
  - L'évaluation du voisinage se fait avec (3) :  $h_{j^*}(j, t) = h(d(j, j^*), t)$

- Phase de coopération :
  - La mise à jour des poids se fait avec (4) :

$$W_j(t+1) = W_j(t) + \Delta W_j(t)$$
$$\Delta W_j(t) = \varepsilon(t) \cdot h_{j^*}(j, t) \cdot (X(t) - W_j(t))$$

$d_N$  distance dans l'espace d'entrée

$d$  distance dans la carte

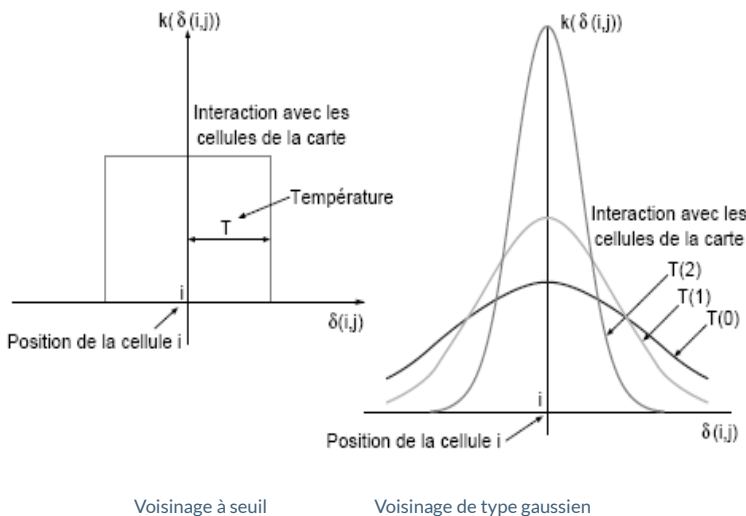
$\varepsilon$  pas d'apprentissage

$h$  fonction voisinage

## Paramètres de l'algorithme

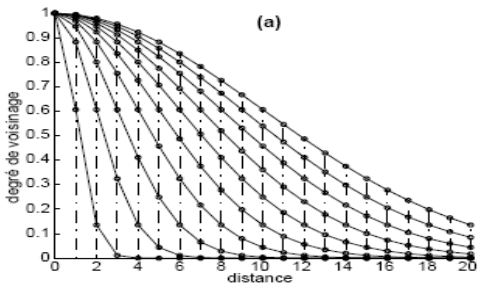
- ▶ La fonction de voisinage dans la carte est une fonction continue de forme gaussienne
  - ▶ La décroissance de la taille du voisinage s'obtient par diminution de l'écart type  $s$
  - ▶  $s$  grand, beaucoup de neurones se rapprochent de  $X(t)$ ,  $s$  petit, l'adaptation reste très localisée
- ▶ Le pas de l'apprentissage contrôle la vitesse d'apprentissage
  - ▶  $e$  trop petit, le modèle ne s'adapte pas assez aux données,  $e$  trop grand, risque d'instabilité
- ▶ Condition de convergence
  - ▶ Choisir les valeurs  $s$  et  $e$  grandes au départ et les réduire graduellement (on choisira des paramètres à décroissance exponentielle)

## Fonctions de voisinage usuelles



## Taille du voisinage

- ▶ plus  $T$  est petit et plus le nombre de neurones inclus dans le voisinage est faible



- ▶ La procédure de SOM réalise plusieurs minimisations de la fonction de coût locale en faisant décroître  $T$

## Exemple d'application

- ▶ Description des données
  - ▶ 53 pays décrits par 6 indicateurs sociaux et économiques (1991) : croissance annuelle (%), mortalité infantile (‰), taux d'analphabétisme (%), taux de scolarisation (%), PNB par habitant, augmentation annuelle du PNB (%)
  - ▶ Représentation des données sous la forme d'une table de 53 lignes et 6 colonnes
- ▶ Construction de la carte auto-organisatrice à partir des données
  - ▶ un vecteur d'entrée de la carte = une ligne de la table (6 dimensions d'entrée)
  - ▶ paramètres de la carte : carte carrée 8 lignes, 8 colonnes (64 neurones), voisinage hexagonal
- ▶ Après apprentissage, chaque individu (un vecteur d'entrée) est associé à un neurone de la carte (le neurone le plus proche de l'individu)

## Représentations

- ▶ Nous allons étudier quelques représentations graphiques de la carte avec R
- ▶ Ces représentations peuvent être proches de celles de l'ACP sans contrainte linéaire

## Classification automatique et Kohonen

- ▶ Les cartes de Kohonen ne sont pas exactement des méthodes de classification automatique
- ▶ Pas vraiment de notion de classe
- ▶ Si on met autant de neurones que de classes, on obtiendra des résultats peu intéressants et proche des k-means
- ▶ regroupement des neurones de la carte d'une manière statistique : on utilise la CAH
- ▶ et recours à l'expertise après la CAH
- ▶ Exemple : couleur de l'océan

## Exemple d'application en télédétection (Niang et al LODYC, Université Pierre et Marie Curie )

- ▶ La problématique :
  - ▶ analyse de la couleur de l'océan permet de mesurer l'intensité de l'activité biologique
  - ▶ Mesure de la réflectance spectrale marine résultat de l'interaction du rayonnement solaire avec les composants de l'atmosphère et de l'océan
  - ▶ correction atmosphérique pour retirer du signal la contribution de l'atmosphère(nuages, aérosols)

## Carte topologique et recherche documentaire

- ▶ SOM sur données textuelles non numériques
- ▶ Websom, créé par Kohonen et son équipe
- ▶ organisation selon le contenu de 7 millions de textes en une seule base de données documentaire
- ▶ visualisation de la base une fois organisée
- ▶ recherche documentaire originale

## Carte topologique et recherche documentaire (2)

- ▶ Modification des algorithmes de base pour :
  - ▶ introduire une connaissance linguistique qui permette la manipulation de textes
  - ▶ entraîner des cartes de grande dimension afin de rendre aussi vaste que possible l'ensemble des documents à prendre en compte pendant la recherche documentaire
  - ▶ réaliser un système de visualisation performant véritable guide de la recherche
  - ▶ de réduire le temps de la recherche documentaire

## Websom

- ▶ six semaines d'apprentissage sur une machine à six processeurs (SGI O2000).
- ▶ Les performances obtenues sur la base des 7 millions de texte atteignent 64 % de bonne classification.
- ▶ Comme pour toutes les applications en fouille de données, l'aspect de visualisation a été très soigné
- ▶ la carte sous la forme d'une suite de pages HTML
- ▶ Exploration par clic de souris sur la carte : atteindre les documents, les visualiser et les lire.

## D'autres applications

- ▶ Plus de 4000 papiers de recherche recensés, de nombreuses applications industrielles
- ▶ Applications
  - ▶ Analyse exploratoire de données, clustering, classification
  - ▶ Quantification, détection de données erronées, sélection de variable
  - ▶ Données manquantes, prédiction, diagnostic
- ▶ Domaines
  - ▶ *Télécommunication* : analyse et détection de la fraude sur les cartes FT (Lemaire, FTR&D), diagnostic de l'état du réseau à partir de mesures de sondes (Fessant, Clérot, FTRD)
  - ▶ *Socio Economie* : analyse du marché immobilier de Paris (Ibbou, U Paris I), segmentation du marché du travail (Gaubert, U Paris I), dépenses de formation en entreprise (Perraudin, U Paris I)
  - ▶ *TextMining* : organisation d'une grande base de documents, le système WEBSOM (Kohonen, UTH), recherche de mots clés dans de grands textes (Kohonen, UTH)
  - ▶ *Processus industriels* : optimisation du dosage d'un coagulant pour un problème de traitement de l'eau (Valentin, Suez)
  - ▶ *Energie* : prédiction de la consommation électrique (Rousset, U Paris I)
  - ▶ *Téledétection* : analyse de la couleur de l'océan à partir d'images satellites (Thiria, U de Versailles)

## Application – classification non supervisée avec R et Kohonen



## Quelques références

- ▶ ANOUAR F., BADRAN F., THIRIA S. [1997], Self Organized Map, A Probabilistic Approach, *Proceedings of the Workshop on Self-Organized Maps*, Helsinki University of Technology, Espoo, Finlande, 4-6 juin 1997.
- ▶ THIRIA S., LECHEVALLIER Y., GASCUEL O., CANU S. [1997], *Statistique et méthodes neuronales*, Dunod.
- ▶ KOHONEN T. [2001], *Self Organizing Maps*, Springer, 3e édition.