Computer Sciences

Semantic data exploration with Self-Organizing Maps

Exploration de données sémantiques avec des cartes auto-organisées

Mots-clés : réseau de neurones, Kohonen, apprentissage non supervisés, cartes topologiques

Project description: Une carte auto-organisée de Kohonen – en anglais Self-Organizing Maps (SOM) – du nom de son inventeur le finlandais Teuvo Kohonen [1], est un modèle d'apprentissage non supervisé, destiné aux applications dans lesquelles le maintien d'une topologie entre les espaces d'entrée et de sortie est important. La caractéristique notable de cet algorithme est que les vecteurs d'entrée qui sont proches – similaires – dans un espace dimensionnel élevé sont également mappés aux noeuds voisins dans l'espace 2D. Il s'agit essentiellement d'une méthode de réduction de dimensionnalité, car le réseau "mappe" des entrées de grandes dimensions à une représentation discrétisée de faible dimension (généralement bidimensionnelle, mais parfois 34-dimensionnelle) et conserve la structure sous-jacente de son espace d'entrée.

Un détail : tout l'apprentissage se déroule sans supervision, c'est-à-dire que les noeuds sont *auto-organisés*. Ces réseaux de neurones sont également appelées cartes de caractéristiques car elles recyclent essentiellement les caractéristiques des données d'entrée et se regroupent simplement en fonction de la similitude entre elles. Cela a une valeur pragmatique pour visualiser des quantités complexes ou importantes de données de grande dimension et pour représenter la relation entre elles dans un champ faible, généralement bidimensionnel, pour voir si les données non étiquetées données ont une structure.

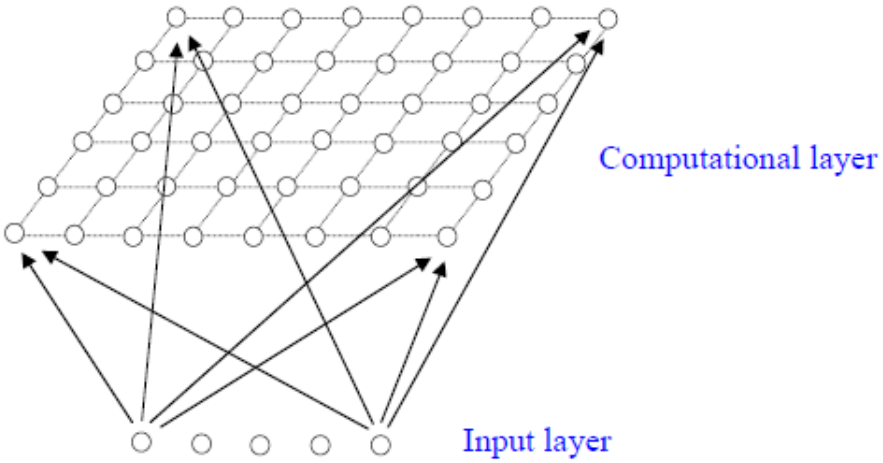

Figure 1: Dans les SOM, les neurones ou cellules sont généralement placés dans une grille régulière. Le réseau neuronal est formé de manière compétitive de telle sorte que chaque neurone cartographie les éléments de données qui lui sont les plus similaires. Après l'entraînement, les cellules qui se ressemblent sont situées les unes à côté des autres, formant ainsi des quartiers. Cette propriété préserve la topologie de l'espace d'entrée d'origine dans la grille ou le treillis.

L'entrée d'une SOM est un vecteur $\mathbf{x}$ dimensionnel m et la couche de sortie est formée par un réseau de y_j pour $j = 1, \ldots, n$ neurones. Chaque neurone de sortie y_j a un vecteur de poids dimensionnel m w_j. Une SOM avec 15 neurones de sortie et 4 neurones d'entrée est illustré à la figure 1 [2].

Une SOM diffère des réseaux de neurones typiques à la fois par son architecture et ses propriétés algorithmiques. Premièrement, sa structure comprend une grille 2D linéaire monocouche de neurones, au lieu d'une série de couches. Tous les noeuds de cette grille sont connectés directement au vecteur d'entrée, mais pas les uns aux autres, ce qui signifie que les noeuds ne connaissent pas les valeurs de leurs voisins, et ne mettent à jour le poids de leurs connexions qu'en fonction des entrées données. La grille elle-même est la carte qui s'organise à chaque itération en fonction de l'entrée des données d'entrée. En tant que tel, après le regroupement, chaque noeud a sa propre coordonnée (i, j), ce qui permet de calculer la distance euclidienne entre 2 noeuds au moyen du théorème de Pythagore.

Variables;

- t est l'itération courante
- n est le nombre maximal d'itération,
- λ est la constante de temps, utilisée pour décomposer le rayon et le taux d'apprentissage
- i, j sont les coordonnées de ligne/colonne de la grille de noeuds
- d est la distance entre un noeud et la *best matching Unit* (BMU)
- w est le vecteur de poids
- $w_{ij}(t)$ est le poids de la connexion entre les noeuds i, j de la grille et l'instance du vecteur d'entrée à l'itération t
- x est le vecteur d'entrée
- $x(t)$ est l'instance du vecteur d'entrée à l'itération t
- $\alpha(t)$ est le taux d'apprentissage, décroissant avec le temps dans l'intervalle [0,1], pour assurer la convergence du réseau.
- $\beta_{ij}(t)$ est la fonction de voisinage, décroissante de manière monotone et représentant un noeud i, la distance de j par rapport à la BMU et l'influence qu'elle a sur l'apprentissage à l'étape t.
- $\sigma(t)$ est le rayon de la fonction de voisinage, qui détermine dans quelle mesure les noeuds voisins sont examinés dans la grille 2D lors de la mise à jour des vecteurs. Il est progressivement réduit au fil du temps.

Les mises à jour et modifications des variables se font selon les formules suivantes:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha_i(t)[x(t) - w_{ij}(t)] \tag{1}$$

ou

$$w_{ij}(t+1) = w_{ij}(t) + \alpha_i(t)\beta_{ij}(t)[x(t) - w_{ij}(t)] \tag{2}$$

Les applications de ces programmes sont nombreuses en économie, en informatique ou en biologie. Citons par exemple l'analyse des trajectoires projssionnelles, les graphes de coocurences des personnages au cinema,

La cartographie de données non numériques ou des tableaux de contingences est très active également en recherche.

Résultats attendus: programmer un modèle de Kohonen pour des données continues et catégorielles. Proposer des améliorations pour la convergence de ces données. Mettre en oeuvre un prototype de ce modèle pour l'analyse de données de patients diabétiques.

Profil et compétences requises Capacité à comprendre et développer des algorithmes d'apprentissage adaptatif et à traiter les données médicales, les indexer et les utiliser dans un système opérationnel pour réaliser la mission décrite ci-dessus.Compétences en programmation: Python ou en C / C ++. Une pratique de Tensorflow et Pytorch serait un plus. La pratique du français nn'est pas obligatoire.

Qualités professionnelles recherchées: autonomie, sens de la relation pour interagir avec les équipes de recherche et de société, motivation pour les nouvelles technologies, créativité pour mettre en place une solution innovante.

Contact: vincent Vigneron, Hichem Maaref, leonardo Duarte
({`vincent.vigneron,hichem.maaref`}`@ibisc.univ-evry.fr`, `leonardo.duarte@fca.unicamp.br`)
Phone: +33 6 635 687 60

References

[1] Teuvo Kohonen. Essentials of the self-organizing map. *Neural Networks*, 37:52–65, 2013.

[2] Aleksey A. Pastukhov and Alexander A. Prokofiev. Kohonen self-organizing map application to representative sample formation in the training of the multilayer perceptron. *St. Petersburg Polytechnical University Journal: Physics and Mathematics*, 2(2):134–143, 2016.