Complex Days

5 & 6 février 2024, Crowne Plaza à Nice

Titre

« Un nouvel estimateur de densité utilisant le phénomène de percolation et les complexes simpliciaux. »

Title

“A new density estimator based on percolation phenomenon and simplicial complexes.”

Par/By

Louis Hauseux, Centre Inria d’Université Côte d’Azur, Sophia-Antipolis. Encadré par/Directed by Konstantin Avrachenkov (équipe/team NEO) et Josiane Zerubia (équipe/team AYANA).

Mots-clefs

percolation ; estimateur de densité ; algorithme de clustering ; graphe géométrique ; complexe simplicial ; Structures à Large Échelle de l’univers.

Keywords

percolation; density estimator; clustering algorithm; geometric graph; simplicial complexe; Large Scale Structures of universe.

Résumé

Les galaxies ne se répartissent pas uniformément au sein de l’univers mais se regroupent au sein de « structures à grandes échelles » : 1° des super-amas de galaxies (petits volumes hyper-denses de $\mathbb{R}^3$) ; 2° des feuillets ou « murs » de galaxies (surfaces) ; 3° des « filaments » de galaxies (courbes). Ces différents clusters délimitent de grandes régions quasiment vides de galaxies.

Identifier correctement ces différents clusters à partir du nuage de galaxies (vues comme des points de $\mathbb{R}^3$) nécessite des outils qui sachent tirer profit de cette structure hiérarchique. On pense par exemple à l’algorithme de clustering H-DBSCAN [1] (‘H’ pour « hiérarchique »). H-DBSCAN part de l’estimateur de densité des k-Plus-Proches-Voisins. Pour notre problème d’identification des clusters de galaxies, d’autres estimateurs de densité ont également été proposés, construits à partir de la triangulation de Delaunay [2] ou du diagramme de Voronoï.

Nous nous intéresserons au phénomène mathématique qui est derrière ces différents algorithmes, à savoir : la **percolation**.

Ce mécanisme est fondamental : bien le comprendre permet d’expliquer les bons résultats – dans un premier temps – et – plus important encore – de les améliorer. Car on peut faire « percoler » à peu près ce que l’on veut : des boules centrées en les points du nuage (comme dans le modèle Booléen de la percolation continue ; cela revient à faire de l’analyse variationnelle sur un graphe géométrique) ; des clusters à haut niveau

de densité comme dans le cas de H-DBSCAN ; des triangles (complexes simpliciaux de dimension 2) comme dans le cas de l'estimateur de Delaunay ; des chaînes de complexes simpliciaux (comme dans le cas de l'homologie persistante).

Pour mesurer la capacité de tels algorithmes à bien distinguer des zones voisines de densités proches, nous définissons un critère : la « vitesse de percolation ». Le bon recouvrement des niveaux de densité se fait d'autant mieux que cette vitesse est grande.

Muni de ces outils, nous pouvons alors comparer différents types de percolation. Il ressort empiriquement que plus la notion de connexité employée est restrictive, plus la vitesse de percolation est grande, nous incitant à travailler avec des complexes simpliciaux et la q-*connectivity* telle que définie par Atkin [3].

Dans le cas d'une discrétisation $\mathbb{R}^d \rightarrow \mathbb{Z}^d$ du phénomène de percolation, nous pouvons même calculer asymptotiquement cette vitesse de percolation sur les clusters de haut niveau de densité des *k*-Plus-Proches-Voisins (ce qui revient peu ou prou à travailler avec des complexes de Čech), expliquant ainsi le choix du *k* dans la pratique.

Abstract

Galaxies are not distributed uniformly in the universe but are grouped together in 'Large Scale Structures': 1° superclusters of galaxies (small, hyper-dense volumes of $\mathbb{R}^3$); 2° sheets or "walls" of galaxies (surfaces); 3° "filaments" of galaxies (curves). These different clusters delimit large regions that are quasi-empty of galaxies.

Correctly identifying these different clusters from the cloud of galaxies (seen as points in $\mathbb{R}^3$) requires tools that take advantage of this hierarchical structure. One example is the H-DBSCAN [1] clustering algorithm ('H' for "hierarchical"). H-DBSCAN is based on the *k*-Nearest Neighbors density estimator. Other density estimators have also been proposed, based on the Delaunay triangulation [2] or the Voronoi diagram.

We will look at the mathematical phenomenon behind these different algorithms: **percolation**.

This mechanism is fundamental: it allows us to explain the good results and to improve them. We can 'percolate' anything we want: balls centered on the points of the cloud (as in the Boolean model of continuum percolation; this is equivalent to variational analysis on a geometric graph); high-density clusters as in the case of H-DBSCAN; triangles (simplicial complexes of dimension 2) as in the case of the Delaunay estimator; chains of simplicial complexes (as in the case of persistent homology).

To measure the ability of such algorithms to distinguish between neighboring areas of close (but distinct) densities, we define a criterion: the "percolation rate".

The higher the rate, the better the density levels recovering. With these tools, we can then compare different types of percolation. Empirically, it appears that the most restrictive notions of connectivity improve the percolation rate, encouraging us to work with simplicial complexes and q-*connectivity* as defined by Atkin [3].

In the case of a $\mathbb{R}^d \rightarrow \mathbb{Z}^d$ normalization of the percolation phenomenon, we can even asymptotically calculate this percolation rate on clusters of high *k*-Nearest Neighbors density (which is more or less the same as working with Čech complexes), thus explaining the choice of *k* in practice.

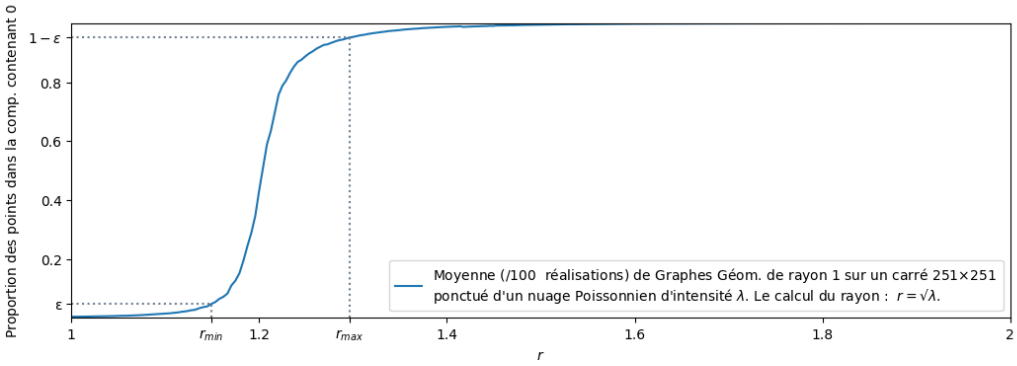

Figure 1. Vitesse de percolation du modèle Booléen : r_{min}/r_{max}.
Boolean model percolation rate: r_{min}/r_{max}.

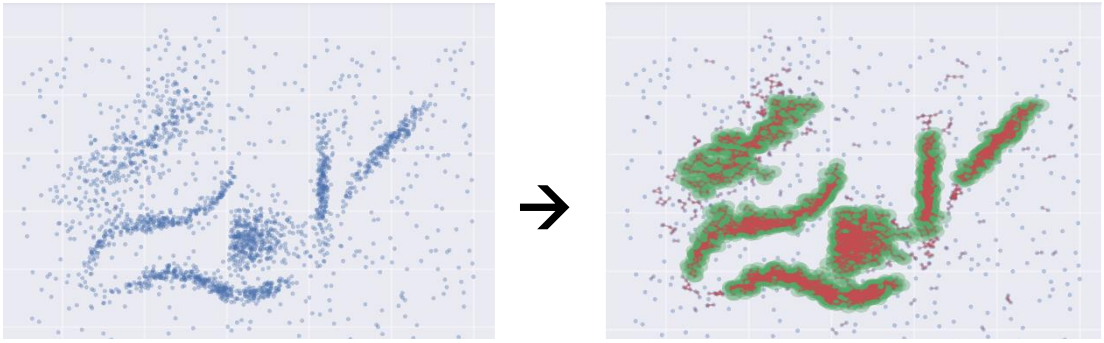

Figure 2. Application au clustering sur un nuage de points.
Application to clustering on a point cloud.

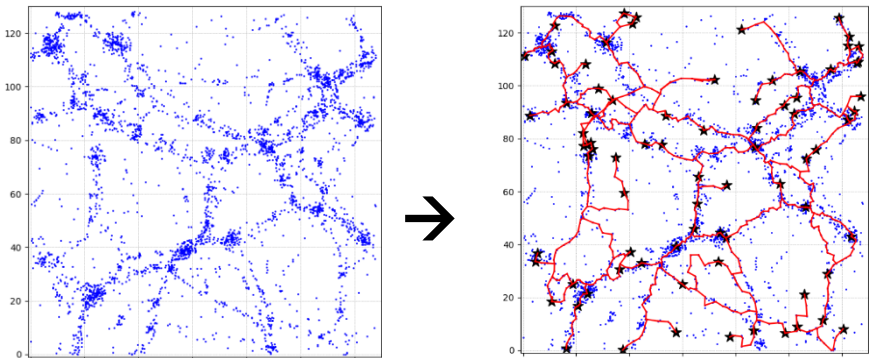

Figure 3. Extraction des filaments de galaxies.
Extraction of galaxy filaments.

Bibliographie/Bibliography

[1] Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: *Advances in Knowledge Discovery and Data Mining*, pp. 160–172 (2013). DOI:10.1007/978-3-642-37456-214

[2] Schaap, W.E.: Dtfe: the Delaunay tessellation field estimator. Ph.D. thesis, Proefschrift Rijksuniversiteit Groningen (2007).

[3] Atkin, R.: From cohomology in physics to *q*-connectivity in social science. In: *International Journal of Man-Machine Studies*, 4-2, pp. 139-167 (1972). DOI:10.1016/S0020-7373(72)80029-4

P.-S. : les diapositives pour l'éventuelle présentation ou le poster seront en anglais. La présentation pourra elle-même se faire en anglais si nécessaire.