Rapport de Julien Mairal et Jean-Philippe Vert sur la thèse de Mathieu Carrière

Sur les propriétés métriques et statistiques des descripteurs topologiques pour les données géométriques

La thèse de Mathieu Carrière étudie une approche originale pluri-disciplinaire consistant à représenter les données par leurs attributs topologiques. A la frontière de la géométrie computationelle, des statistiques, et de l'analyse de données, il s'agit d'un sujet conceptuellement élégant, mathématiquement très intéressant, mais qui est traditionnellement peu abordé par la communauté d'apprentissage statistique, dont les interactions avec la communauté de géométrie computationelle sont très rares. Dans sa thèse, Mathieu Carrière réussit le tour de force de contribuer de façon significative dans les deux disciplines et d'établir des ponts importants, qui pourront potentiellement inspirer une génération nouvelle de travaux.

En géométrie computationelle, ses contributions mathématiques sur la stabilité des mappers, ainsi que son travail sur les métriques permettant de comparer des graphes de Reeb ont ainsi donné lieu à trois publications dont l'une au Journal of Foundations of Computational Mathematics et deux à la conférence SoCG. A l'intersection des deux disciplines, une méthode de représentation des diagrammes de persistance pour la classification de formes 3D a été publiée dans la conférence SGP en 2015. Récemment, les travaux de Mathieu Carrière ont aussi été reconnus par la communauté d'apprentissage statistique, avec une publication à la conférence hautement sélective ICML (il s'agit d'une des deux conférences majeures d'apprentissage automatique, qui joue un rôle aussi important que les journaux internationaux). Cette dernière contribution est par ailleurs extrêment originale. Bien que moins technique que les précédentes, elle fait un lien entre l'analyse de données topologique, le transport optimal, et les méthodes à noyaux en apprentissage statistique, ce qui témoigne d'un spectre de compétences très large et d'une grande maturité scientifique.

Les chapitres introductifs 1 et 2 présentent respectivement une description des enjeux, des outils, et des contributions de façon non-formelle et intuitive, et une introduction des outils mathématiques utilisés dans la thèse (diagrammes de persistance, graphes de Reeb, Mapper). Le chapitre 1 fait preuve d'une grande pédagogie est constitue un point d'entrée clair et intuitif pour des membres de la communauté d'apprentissage statistique. Le chapitre 2 introduit les mêmes outils de façon abstraite et mathématique. Le chapitre est alors beaucoup plus aride, mais introduit en environ 25 pages, ce qui est usuellement introduit en plusieurs chapitres d'ouvrages.

CENTRE DE RECHERCHE GRENOBLE - RHÔNE-ALPES
Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex France
Tél. : +33 (0)4 76 61 52 00
Fax : +33 (0)4 76 61 52 52

ANTENNE Inria LYON LA DOUA
Bâtiment CEI-1
66 Boulevard Niels Bohr
69603 Villeurbanne France
Tél. : +33 (0)4 72 43 74 90
Fax : +33 (0)4 72 43 74 99

www.inria.fr

Le chapitre 3 apporte des justifications théoriques à l'utilisation de la distance "bottleneck" d_b pour comparer des graphes de Reeb. Celle-ci représente ces graphes par des diagrammes de persistance étendus et n'est donc qu'une pseudo-métrique (deux graphes de Reeb différents pouvant avoir le même diagramme de persistance étendu), ce qui rend discutable son utilisation pour des tâches discriminatives. Cependant, la métrique d_b a l'avantage d'être calculable en pratique, contrairement à d'autres métriques pour les graphes de Reeb, telles que la distortion fonctionnelle d_{FD} ou la distance de Gromov-Haussdorff d_{GH}. Le chapitre présente alors deux résultats positifs concernant la distance d_b. Tout d'abord, *localement*, d_b est équivalente aux métriques mentionnées précédemment. Par ailleurs, la métrique induite intrinsèque est *globalement* équivalente aux métriques intrinsèques induites par d_{FD} et d_{GH}. Bien que n'étant pas très bien placés pour juger de l'importance de cette contribution dans la littérature de géométrie computationnelle, nous avons trouvé les résultats très intéressants, ainsi que les arguments avancés compréhensibles et cohérents. Les deux résultats précédents nous semblent être des étapes significatives permettant de mieux comprendre la structure locale de l'espace des graphes de Reeb. Le chapitre se conclut par une liste de questions ouvertes intéressantes comme l'existence ou non de plus courts chemins dans l'espace des graphes de Reeb, qui permettrait d'utiliser en pratique la métrique intrinsèque induite par d_b.

Le chapitre 4 se focalise ensuite sur l'outil Mapper, qui est dérivé des graphes de Reeb, et qui a connu un certain nombre de succès applicatifs. Le chapitre étend la distance de bottleneck aux diagrammes de persistance d'une variante du Mapper (multinerve Mapper). L'intérêt est ensuite de montrer que cette pseudo-métrique est stable (non-expansive par rapport à la norme ℓ_∞). Etant donné que Mapper peut être vu comme une version pratique "pixelisée" du graphe de Reeb (le chapitre introduit par ailleurs un résultat de convergence du Mapper vers le graphe de Reeb), cette contribution permet de mieux comprendre le succès du Mapper, qui a démontré des qualités importantes dans des applications.

Le chapitre 5 s'intéresse à des aspects statistiques moins abstraits que les deux chapitres précédents, permettant de caractériser le taux de convergence de Mapper vers le graphe de Reeb, lorsque le Mapper est calculé sur un nuage de points échantillonné à partir d'une variété de dimension d. Ce chapitre permet de faire le lien entre les outils de géométrie computationnelle décrits dans les chapitres précédents, et des applications réelles où les données sont échantillonnées. Le résultat principal montre que le mapper converge en espérance vers le graphe de Reeb, en distance de bottleneck, avec un taux borné par $O((\log(n)/n)^{1/d})$ pour des fonctions régulières. Une borne inférieure est aussi calculée, montrant que le taux est optimal d'un point de vue minimax. Ce résultat a deux implications majeures. Tout d'abord, il montre que la dimension de la variété joue un rôle critique (ce qui n'est intuitivement pas surprenant, mais une justification théorique est tout à fait bienvenue). En second lieu, il donne

CENTRE DE RECHERCHE GRENOBLE - RHÔNE-ALPES
Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex France
Tél. : +33 (0)4 76 61 52 00
Fax : +33 (0)4 76 61 52 52

ANTENNE Inria LYON LA DOUA
Bâtiment CEI-1
66 Boulevard Niels Bohr
69603 Villeurbanne France
Tél. : +33 (0)4 72 43 74 90
Fax : +33 (0)4 72 43 74 99

www.inria.fr

aussi lieu à une heuristique pratique permettant de choisir une couverture spécifique pour le Mapper. Une application intéressante de l'analyse statistique permet aussi de calculer des diagrammes de persistance incluant des intervalles de confiance.

Enfin, le chapitre 6 présente des distances pratiques permettant de comparer des diagrammes de persistance, et de les utiliser dans des applications d'apprentissage automatique. La contribution de ce chapitre est particulièrement originale, et permet d'établir un point entre géométrie computationelle, transport optimal, et méthodes à noyaux. La distance obtenue utilise le noyau "sliced Wasserstein" et il est démontré, que sous certaines hypothèses légères, la distance induite par le noyau est équivalente à la distance de Wasserstein pour les diagrammes de persistance (bien que la distance de Wasserstein ne soit pas une métrique Hilbertienne). Le chapitre aborde alors des questions pratiques computationelles, permettant d'approcher cette distance, et enfin des applications convaincantes de la distance pour des tâches de classification. Il s'agit ainsi d'un chapitre très complet, abordant des questions de modélisation (design d'une métrique adaptée aux données, d'un point de vue topologique), des questions théoriques (stabilité de la métrique), et aussi pratiques avec plusieurs applications, qui ne manqueront pas d'éveiller l'intérêt de la communauté d'apprentissage statistique.

Pour conclure, Matthieu Carrière a effectué une thèse excellente, et a contribué de façon significative à plusieurs domaines scientifiques, ce qui est remarquable. Il a su résoudre certains problèmes ouverts théoriques demandant une analyse mathématique fine, proposer de nouveaux modèles et outils, et a su appliquer (de façon parcimonieuse) ses recherches à des problèmes pratiques. En conséquence, nous émettons un avis extrêmement favorable à l'autorisation de soutenance de ses travaux, en vue de l'obtention de son diplôme de doctorat de l'université Paris Saclay.

Julien Mairal

Jean-Philippe Vert

CENTRE DE RECHERCHE GRENOBLE - RHÔNE-ALPES
Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex France
Tél. : +33 (0)4 76 61 52 00
Fax : +33 (0)4 76 61 52 52

ANTENNE Inria LYON LA DOUA
Bâtiment CEI-1
66 Boulevard Niels Bohr
69603 Villeurbanne France
Tél. : +33 (0)4 72 43 74 90
Fax : +33 (0)4 72 43 74 99

www.inria.fr