Signatures pour les formes géométriques

Domaine : analyse géométrique et topologique de données

Encadrant : Steve Oudot

Institution d'accueil : Inria Saclay Ile-de-France, équipe Geometrica

Lieu : campus de l'Ecole polytechnique, bâtiment Turing
1 rue Honoré d'Estienne d'Orves, 91190 Palaiseau

Contact : steve.oudot@inria.fr

Problématique générale

Les domaines de l'acquisition et de la génération de données géométriques connnaissent un essor constant depuis plusieurs décennies, si bien qu'aujourd'hui de larges banques de données de formes géométriques sont disponibles. Dans le but d'organiser ces données, il est primordial de définir des notions pertinentes de similarité entre formes géométriques, qui soient invariantes à l'échantillonnage et aux différentes poses que peuvent prendre les objets ou les êtres que ces formes représentent. Des questions de cette nature apparaissent dans de nombreux domaines scientifiques, comme par exemple la biologie structurale, dont l'un des objectifs est de classifier les différentes configurations d'une protéine, ou l'informatique graphique, dont l'un des problèmes consiste à recaler les différents nuages de points acquis par un scanner sur un objet, ou encore l'intelligence artificielle, dont l'un des objectifs est la classification de formes 2d ou 3d et plus généralement de données en toutes dimensions.

Les jeux de données géométriques sont généralement constitués d'un nuage de points entre lesquels une distance (euclidienne, géodésique, de diffusion, etc) est appliquée. Les propriétés intrinsèques du nuage dépendent fondamentalement de la métrique qui lui est associée. Si le choix d'une métrique particulière dépend clairement de l'application visée, le cadre général consistant à considérer les nuages de points comme des espaces métriques reste le même. La distance canonique entre espaces métriques est la distance de Gromov-Hausdorff, qui est invariante aux isométries et fournit une quantification objective de la similarité ou de la dissimilarité entre espaces. Le problème est que la distance de Gromov-Hausdorff est trop coûteuse à calculer pour être utilisée en pratique. L'objectif devient alors de l'estimer au mieux afin de se doter d'outils capables de comparer les nuages de points (et leurs formes sous-jacentes) entre eux.

Développée initialement dans le cadre de l'analyse topologique de données, la *théorie de la persistance* [6] a été récemment proposée comme approche potentielle pour définir des signatures sur des formes 3d [1,2]. Les signatures obtenues, appelées *diagrammes de persistance*, ont la propriété fondamentale d'être stables vis-à-vis des perturbations potentielles de leur forme associée. Ceci est une conséquence du fameux *théorème de stabilité* [3], un résultat central de la théorie. Cette approche n'est toutefois pas exempte de défauts, dont les principaux sont les suivants :

- Les signatures obtenues vivent dans un espace où la métrique est coûteuse à calculer, où les notions de moyenne et de barycentre sont mal définies, enfin où il est difficile de définir des noyaux pour les méthodes d'apprentissage classiques [7].
- Ces signatures sont globales par nature, c'est-à-dire qu'elles caractérisent une forme 3d dans son ensemble. Elles sont donc inappropriées pour la reconnaissance ou la classification de parties de formes, comme par exemple les différents membres d'un animal.
- Enfin, le calcul de ces signatures en pratique engendre des problèmes majeurs de gestion de la mémoire, car ils reposent sur la construction de familles de complexes simpliciaux dont la taille explose rapidement avec la dimension et la taille des données. En pratique on doit donc se restreindre et ne calculer qu'une partie de chaque signature, perdant ainsi une part peut-être

importante de l'information qu'elle contient.

Ces raisons rendent les signatures basées sur la persistance difficiles à mettre en oeuvre dans des cas pratiques. Toutefois, une de leurs grandes qualités est d'être très générales, ce qui permet de les utiliser au-delà du cadre des formes 3d, sur des jeux de données en grandes dimensions par exemple.

Contributions et programme de recherche

Le but de la thèse sera de définir de nouvelles signatures basées sur la persistance qui ne souffrent pas des défauts mentionnés plus haut. On considérera en particulier les points suivants :

1. **Construire des signatures locales**, caractérisant un point et son voisinage sur une forme géométrique (3d ou non). C'est l'un des aspects clés de la thèse. L'approche qui semble la plus naturelle a priori consiste à regarder la fonction distance au point considéré le long de la forme (qui peut être facilement approchée à partir d'un échantillonnage de la forme) et à calculer son diagramme de persistance. D'autres fonctions sont également envisageables, et la grande question est celle de la stabilité de tels diagrammes, qui sort du cadre de la théorie actuelle. On s'attend à un travail de fond de nature très topologique sur ce point.
2. **Définir des noyaux** à partir de diagrammes de persistance. Alors que l'on sait que cette tâche est difficile dans l'espace des diagrammes, l'idée serait d'envoyer les signatures dans un autre espace, typiquement un espace vectoriel, quitte à perdre au passage une partie de l'information contenue dans ces signatures. Le compromis à realiser se situe entre d'une part la quantité d'information perdue, et d'autre part l'intérêt et la praticité des noyaux obtenus. Le travail ici est de nature très géométrique, en lien avec les applications potentielles en apprentissage. Une partie importante du travail sera d'ailleurs d'évaluer la qualité des noyaux obtenus pour la classification supervisée ou semi-supervisée de formes géométriques, en particulier de formes 3d. Pour cela on s'appuiera sur des banques de données de formes 3d qui font référence à l'heure actuelle, comme par exemple celle du concours SHREC (http://www.aimatshape.net/event/SHREC).
3. **Rendre les signatures plus faciles a calculer**. Actuellement on ne sait calculer qu'une partie de chaque signature, ce qui revient à laisser de côté une part potentiellement importante de l'information qu'elle contient. Pour éviter cet écueil, il faut définir des variantes des signatures qui soient proches (pas forcément identiques) et beaucoup plus faciles à calculer. Des travaux dans ce sens ont été réalisés dans un contexte différent [4,5], et il sera possible de s'appuyer dessus pour démarrer. L'impact potentiel est énorme, puisque cela permettra d'exploiter tout le potentiel des signatures basées sur la persistance en pratique, chose qui reste impossible à l'heure actuelle.
4. **S'attaquer à la délicate question des bornes inférieures**. En effet, aucun résultat théorique ne vient aujourd'hui appuyer l'observation empirique que les signatures basées sur la persistance sont informatives. En particulier, on ne sait pas borner inférieurement la distance entre ces signatures en fonction de la distance de Gromov-Hausdorff entre les formes qu'elles représentent : ainsi, en principe, deux formes très différentes géométriquement peuvent avoir des signatures très proches, voire même identiques. Prouver des bornes inférieures théoriques, quitte à faire des hypothèses (raisonnables) supplémentaires sur les données, est reconnu comme une question fondamentale mais difficile, qui devra donc être abordée sur le long terme, probablement une fois que le doctorant aura acquis une expérience suffisante du sujet.
5. Au-delà des formes 3d, le doctorant pourra **s'intéresser à la classification et à l'analyse de jeux de données en plus grandes dimensions**. En effet, les constructions utilisées pour définir les signatures sont indépendantes de la dimension des données, toutefois leur complexité augmente avec la dimension. Il conviendra donc de réfléchir à la meilleure stratégie pour faire passer les constructions à l'échelle. A la clé, il sera possible d'utiliser les signatures sur des jeux de données en grandes dimensions, ce afin de les pré-traiter en détectant par exemple leurs symétries.

Ce ne sont là que certains des aspects qui pourront être regardés durant la thèse. Le sujet est suffisamment vaste pour laisser l'imagination et l'esprit de conquête du doctorant s'exprimer. Par ailleurs, la concurrence sur ce sujet n'est pas encore trop forte, ce qui laisse un peu de marge de manoeuvre à l'heure actuelle. Il semble que ce soit le bon moment pour l'aborder.

Publications associées

Publications de l'équipe :

1. F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli, S. Y. Oudot. Gromov-Hausdorff Stable Signatures for Shapes using Persistence. *Computer Graphics Forum* (proc. SGP 2009), pages 1393-1403.
2. F. Chazal, V. de Silva. S. Oudot. Persistence Stability for Geometric Complexes. *Geometriae Dedicata*, pages 1-22, 2013.
3. F. Chazal, V. de Silva, M. Glisse, S. Oudot. The Structure and Stability of Persistence Modules. Monograph under review, see arXiv:1207.3674v1 [math.AT].
4. M. Buchet and F. Chazal and S. Y. Oudot and D. R. Sheehy. Efficient and Robust Topological Data Analysis on Metric Spaces. arXiv:1306.0039 [cs.CG].
5. S. Y. Oudot and D. R. Sheehy. Zigzag Zoology: Rips Zigzags for Homology Inference. *Proc. 29th Annual Symposium on Computational Geometry*, June 2013.

Autres publications :

6. H. Edelsbrunner and J. L. Harer. Computational Topology. An Introduction. American Mathematical Society, 2010.
7. Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Frn’echet means for distributions of persistence diagrams. arXiv preprint arXiv:1206.2790, 2012.