Méthodes topologiques pour la classification de données géométriques

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 les données 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 d'un nuage dépendent fondamentalement de la métrique qui lui est appliquée, et le choix d'une métrique particulière dépend donc de l'application visée. Toutefois, le même cadre général consistant à considérer les nuages de points comme des espaces métriques finis s'applique. La distance canonique entre de tels espaces 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.

Le travail de thèse s'appuiera en premier lieu sur un article récent co-écrit par plusieurs membres de l'équipe Geometrica [1], qui introduit les notions de topologie persistante et de diagramme de persistance dans le monde de la comparaison de formes. L'idée maîtresse de ce travail est de se servir des diagrammes de persistance pour construire des signatures pour espaces métriques finis, et de prouver leur stabilité vis-à-vis de la distance de Gromov-Hausdorff. Ainsi, deux espaces métriques proches pour cette distance ont des signatures proches. En d'autres termes, les distances entre signatures fournissent des bornes inférieures sur les distances de Gromov-Hausdorff entre les espaces correspondants, tout en étant plus faciles à calculer. Des expérimentations préliminaires en classification de formes 3d ont donné des résultats encourageants, mais plusieurs questions importantes restent ouvertes, dont les suivantes qui serviront de fil conducteur pour la thèse :

1. Les bornes sur la distance de Gromov-Hausdorff fournies par les distances entre signatures topologiques sont uniquement des bornes inférieures. Quelle est la pertinance de ces bornes ? Sont-elles toujours proches de la vraie distance de Gromov-Hausdorff, ou bien peuvent-elles au contraire être arbitrairement mauvaises ? Une autre manière de poser la question est de demander si l'on peut obtenir des bornes supérieures sur la distance de Gromov-Hausdorff par l'approche topologique. Ce sera l'une des grandes questions abordées dans la thèse.

2. Les signatures fournies par l'approche topologique sont globales, c'est-à-dire qu'elles caractérisent le nuage de point dans son ensemble et permettent donc uniquement la classification de nuages entiers. Peut-on définir des variantes plus locales de ces signatures, qui caractériseraient de manière pertinente certaines portions du nuage ? À la limite, peut-on définir des signatures topologiques ponctuelles, qui caractériseraient un point donné et son voisinage dans le nuage ? Répondre positivement à ces questions permettrait ensuite de faire de la comparaison et du matching partiel de formes. Ce sera l'un des objectifs principaux de la thèse. Ses implications sont nombreuses, par exemple en informatique graphique avec le problème du recalage de nuages de points, ou encore en bio-informatique avec le problème du docking de protéines.

3. Les questions de complexité inhérentes au calcul de signatures topologiques sont nombreuses et intimement liées à celles rencontrées plus généralement en topologie appliquée. En effet, les signatures topologiques sont élaborées à partir de familles de complexes simpliciaux à un paramètre, appelées des *filtrations*, dont la taille explose très rapidement avec le paramètre de la famille. L'approche typique en pratique consiste à ne construire que le début de la famille, et ainsi à calculer des signatures partielles. S'il est vrai que ces dernières sont garanties de fournir encore des bornes inférieures sur la distance de Gromov-Hausdorff, il n'est pas dit que ces bornes soient de qualité comparable à celles que l'on pourrait obtenir à partir des filtrations complètes. Deux questions se posent donc :

- Peut-on donner des garanties aux signatures partielles sous certaines hypothèses sur le seuil appliqué au paramètre de la filtration ? Si oui, quel compromis peut-on réaliser entre complexité du calcul et qualité du résultat ?
- Peut-on remplacer les filtrations actuellement utilisées par d'autres moins coûteuses dont les propriétés topologiques seraient comparables ? Cette dernière question est déjà étudiée dans un autre contexte par certains membres de l'équipe Geometrica [3], ainsi le travail effectué dans la thèse contribuera aux avancées sur le sujet.

L'essentiel du travail de thèse s'articulera autour des trois grands axes ci-dessus. Au passage, d'autres questions pourront être abordées, comme par exemple celle de la gestion du bruit et des anomalies générés lors de l'acquisition des données, dont l'importance et les implications dépassent largement le cadre de la classification de formes, et qui sont déjà sujets d'intenses efforts de la part de l'équipe Geometrica [2] et de la communauté de topologie appliquée dans son ensemble.

Références

[1] F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli, and S. Y. Oudot. Gromov-hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum (proc. SGP 2009)*, pages 1393–1403, 2009.

[2] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for measures based on distance functions. Preprint, submitted to the 26th Annual Symposium on Computational Geometry, 2010.

[3] B. Hudson, G. Miller, S. Y. Oudot, and D. Sheehy. Topological Inference via Meshing. Research Report RR-7125, INRIA, 2009.