Avis de Soutenance

Monsieur Nguyen Viet Dung NGHIEM

Informatique

Soutiendra publiquement ses travaux de thèse intitulés

Clustering et intégration de connaissances

dirigés par Madame Christel VRAIN et Madame Thi Bich Hanh DAO

Ecole doctorale : Mathématiques, Informatique, Physique Théorique et Ingénierie des Systèmes - MIPTIS
Unité de recherche : LIFO - Laboratoire d'Informatique Fondamentale d'Orléans

Soutenance prévue le ***mercredi 15 décembre 2021*** à 14h00
Lieu : LIFO, bâtiment 3IA, 6 Rue Léonard de Vinci, 45067 Orléans
Salle : Amphi H

Composition du jury proposé

Mme Christel VRAIN	Université d'Orléans	Directrice de thèse
Mme Thi-Bich-Hanh DAO	Université d'Orléans	Co-directrice de thèse
M. Pierre MARQUIS	Université d'Artois	Examinateur
M. Antoine CORNUEJOLS	AgroParisTech	Examinateur
M. Tias GUNS	KU Leuven, Belgium	Rapporteur
M. Dino IENCO	UMR TETIS	Rapporteur

Mots-clés : clustering sous contraintes,apprentissage profond,approche déclarative,,

Résumé :

Le clustering sous contraintes (une généralisation du clustering semi-supervisé) vise à exploiter les connaissances des experts lors de la tâche de clustering. Ces connaissances peuvent prendre des formes diverses : des relations entre instances, des conditions sur les clusters, telles que leur cardinalité, … mais aussi des connaissances plus sémantiques comme par exemple, obtenir des clusters équitables. Les contraintes peuvent être intégrées à différentes étapes du processus de clustering : en pré-traitement, par exemple en apprenant une nouvelle métrique entre points, pendant le processus de clustering ou dans une étape de post-traitement. La plupart des travaux intègrent des contraintes pendant le clustering, et ils peuvent être divisés en deux approches : modifier des algorithmes de clustering existants pour gérer des contraintes spécifiques / modéliser le problème dans des cadres déclaratifs, tels que la Programmation Linéaire en Nombres Entiers (PLNE), SAT ou la Programmation par Contraintes. Dans cette thèse, nous proposons trois contributions : (1) une méthode déclarative modifiant une partition existante pour satisfaire des contraintes ; (2) un cadre générique pour intégrer plusieurs types de contraintes dans un modèle de clustering par apprentissage profond ; (3) la définition et la formulation de nouveaux types de contraintes. Notre première contribution porte sur une méthode de post-traitement, en sortie d'un algorithme de clustering, pour assurer la satisfaction de contraintes. L'originalité est de considérer une matrice d'allocation qui donne les scores d'attribution des points à chaque cluster et de trouver la meilleure partition satisfaisant toutes les contraintes. Nous formulons ce problème comme un problème d'optimisation en PLNE. Des expérimentations montrent que cette méthode est efficace tout en étant compétitive en termes de qualité du clustering par rapport à l'état de l'art. Utiliser une matrice d'allocation permet de post-traiter le résultat

de divers algorithmes de clustering, qu'ils soient probabilistes ou issus d'un modèle d'apprentissage profond. Alors que dans la première contribution, les contraintes sont traitées après un algorithme d'apprentissage, notre deuxième contribution vise à exploiter ces contraintes directement dans un modèle d'apprentissage profond. Les avancées en apprentissage profond permettent de trouver une représentation des données dans des espaces de dimension plus faible grâce à des plongements non linéaires. Elles ont conduit au développement du Deep Clustering, des méthodes de clustering fondées sur l'apprentissage profond. Pour introduire des contraintes lors de l'étape d'apprentissage, il est courant de disposer d'une fonction de perte pénalisant la non-satisfaction des contraintes. Les travaux actuels introduisent une fonction différente pour chaque type de contrainte (contraintes de taille, contraintes par paires, triplet, …). Dans notre travail, nous proposons un cadre unifié pour intégrer les contraintes générales en les formalisant en logique et en considérant leurs modèles. A notre connaissance, nous sommes les premiers à proposer pour le clustering profond un cadre générique pour intégrer des contraintes expertes. Nous proposons deux formulations de la fonction de perte, fondées sur la notion de modèles et nous montrons qu'elles peuvent être calculées de manière efficace grâce à des techniques de Weighted Model Counting. Les résultats expérimentaux sur des jeux de données connus montrent que notre approche est compétitive avec d'autres méthodes spécifiques aux contraintes, tout en étant générale. Outre ces deux méthodes génériques, nous avons défini et formulé de nouveaux types de contraintes en clustering. Premièrement, la contrainte de couverture de cluster limite le nombre de clusters auxquels un groupe de points peut appartenir. Deuxièmement, l'équité combinée prend en compte à la fois l'équité de groupe et l'équité individuelle.