Université de Lorraine - IDMC
Masters SC & TAL – M1
2021-2022

Fiche de projet tutoré

How to evaluate the faithfulness of visual data projection?

Comment évaluer la pertinence d'une projection visuelle de données?

Encadrement / Supervisors

1. équipe, laboratoire / team, lab :
équipe ORPAILLEUR - LORIA
1. encadrant·e principal·e (nom, email) / main supervisor (name, email) :
Lydia Boudjeloud-Assala (lydia.boudjeloud@loria.fr) (Site de Metz)

Description / Description

1. projet global/global project

Certains algorithmes de fouille de données considèrent des concepts de voisinage en se basant sur les relations entre les données. Dans les ensembles à grandes dimensions, les données sont rares et éparses; et les notions de distance ou de voisinage perdent leur sens. L'éparpillement, la rareté dans les espaces de grandes dimensions implique que toute relation devient complexe et très difficile à interpréter.

Ceci illustre un des aspects de "la malédiction de la dimensionnalité" ou "curse of dimensionality" telle que définie par Belmann [1]. Pour surmonter la malédiction de la dimensionnalité, une des approches est d'utiliser des méthodes de réduction de dimensions telles que l'ACP (Analyse en Composantes Principales) [2] ou des méthodes de sélection de dimensions.

Les approches de réduction de dimensions telle que l'ACP [2], Multi Dimensional Scalling (MDS [3]) ou Stochastic Neighbor Embedding (T-SNE [4]) proposent des solutions en projetant les données dans des espaces qui préservent, le plus possible, les distances obtenues dans l'espace d'origine.

Cependant, en termes d'interprétation, ces techniques présentent quelques difficultés. Il est, par exemple, difficile d'interpréter le voisinage entre deux points obtenu dans la projection, par rapport à leur voisinage réel dans l'espace d'origine.

Les techniques de sélection de dimensions présentent également quelques difficultés, selon le critère utilisé, pour rechercher les sous espaces (sous-ensembles de dimensions) pertinents.

Un individu (ou un groupe d'individus) peut être décrit par plusieurs sous-ensembles de dimensions.

Cet individu peut se comporter comme un individu atypique (outlier) dans un sous espace, et

dans un autre sous espace il se mêle complètement dans la masse, faisant partie d'un cluster (groupe) homogène.

Projeter les données dans un espace réduit peut induire une perte d'information. De plus, il est généralement difficile d'interpréter la visualisation et d'évaluer l'information restituée visuellement par la projection. Evaluer la correspondance entre ce qui est projeté et visualisé dans le sous espace avec la structure des données réelles est aussi importante que l'efficacité et la précision de l'algorithme de fouille et/ou de sélection utilisé, et ces deux tâches peuvent être évaluées soit dans l'espace total des données et/ou dans différents sous espaces. Il est donc nécessaire de proposer des critères ou indicateurs pouvant guider l'utilisateur. C'est la seule façon d'avoir une confiance dans l'outil proposé et dans les résultats obtenus.
Dans cette perspective, le sujet porte sur l'évaluation de l'information restituée dans les projections de données pour l'exploration visuelle des données, particulièrement pour les ensembles de données de grandes dimensions.
L'approche que nous nous proposons de suivre, dans un premier temps, est d'étudier les critères utilisés par les méthodes de sélection de dimensions et de réduction de dimensions ainsi que les critères utilisés en visualisation d'information pour sélectionner les projections optimales de données.
Une procédure d'évaluation et de comparaison des différents critères d'évaluation de l'information restituées dans les différentes projections est à faire. Nous partons de l'hypothèse qu'il peut y avoir deux types d'évaluation.
Soit l'information restituée dans la projection correspond à ce qui existe réellement dans les données (Whole data Information), ou bien la projection nous permet de découvrir de nouvelles informations (Discovering Meanigful Information).

2. biblio. UE 705 (semestre 7)
Ce sujet représente un interet dans différentes communautés : Information Visualization [5], Visual Quality Metrics [6], High Dimensional Visualization [7] et Data Mining [8].
L'objectif principal est de fournir des outils et/ou crières permettant d'évaluer des visualisation de données.
Ces critères d'évaluation pourront ainsi fédérer les différentes communautés.
ces critères d'évaluation serviront d'étapes intermédiaires dans les outils interactifs et incrémentales de machine learning, afin d'obtenir des résultats pertinents aux diférentes étapes de l'algorithme.
La première étape de ce stage consiste à étudier l'état de l'art des communautés concernées pour tenter de référencer tous les critères d'évaluation de l'information existants à l'heure actuelle.

3. réalisation. UE 805 (semestre 8)
Dans un second temps, l'idée est de mettre en place un protocole de comparaison des différents critères selon le type d'évaluation et les techniques de projections utilisées.

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization
Encadrant à Metz

Livrables et échéancier / Deliverable and schedule
Réunions toutes les semaines par visio

Bibliographie /References (max. 4-5)
[*il ne s'agit pas de la bibliographie complète qui sera fournie aux étudiants au début du projet mais d'une bibliographie indicative pour aider à cerner le sujet*]

[1] R.E. Bellman. Adaptive Control Processes. Princeton University Press, Princeton, NJ, 1961.
[2] I.T. Jolli_e. Principal Component Analysis. Springer, NY, 2nd edition, 2002.
[3] J. B. Kruskal and M. Wish. Multidimensional Scaling. Beverly Hills and London : Sage Publications, 1978.
[4] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research, 9 :2579_2605, Nov 2008.
[5] C. Johnson, R. Moorhead, T. Munzner, H. P_ster, P. Rheingans, and T.S. Yoo. NIH/NSF Visualization Research Challenges Report. IEEE Press, 2006.
[6] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization : An overview and systematization. In Proceedings of the IEEE
Transaction on Visualization and Computer Graphics, volume 17, pages 2203_2212, 2011.
[7] Sara Johansson Fernstad, Jane Shaw, and Jimmy Johansso. Quality based guidance for exploratory dimensionality reduction. Information Visualization Journal, page 24, 2013.
[8] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection : A data perspective. CoRR, abs/1601.07996, 2016.