Invitation à la soutenance publique de thèse de
Cyril DE BODT

Pour l’obtention du grade de Docteur en sciences de l’ingénieur et technologie
« Robust and Fast Neighbor Embedding Algorithms »

qui se déroulera
le vendredi 10 avril 2020 à 15h
par visioconférence
1348 Louvain-la-Neuve

Jury members :
Prof. Michel Verleysen (UCLouvain), supervisor
Prof. John Lee (UCLouvain), supervisor
Prof. Jean-Pierre Raskin (UCLouvain), chairperson
Prof. Jean-Charles Delvenne (UCLouvain), secretary
Prof. Barbara Hammer (Bielefeld University, Germany)
Prof. Benoît Frenay (UNamur, Belgium)
Prof. Laurent Jacques (UCLouvain)

In numerous machine learning (ML) settings, complex data mining tasks necessitate user interaction and cannot be completely automated. This interaction can take place in the context of a data exploratory phase, during which data visualization helps determining and refining the application needs. Since most ML databases are nowadays high-dimensional (HD), their visualization entails considering approaches from nonlinear dimensionality reduction (DR). This field aims at creating meaningful low-dimensional (LD) versions of HD data. The advent of neighbor embedding (NE) techniques impressively improved state-of-the-art DR performances.

Nevertheless, the data exploratory requirements in current ML applications involve generalizing modern NE algorithms in several aspects. Namely, they should be robustly able to handle unconventional data types, such as incomplete databases which are omnipresent in data analysis. Also, they must be sufficiently fast to process very large data sets, being ubiquitous presently. This thesis contributes to both of these aspects.

Regarding the ability to deal with incomplete data sets, common missing data imputation techniques are not suited to nonlinear DR, as they at best enable applying a DR scheme on the expected database. Since NE approaches are nonlinear, this differs from minimizing their expected cost function. The thesis addresses this limitation by proposing a general methodology to compute the LD embedding minimizing the cost function expectation, thanks to the multiple imputation framework.

As to the development of fast NE schemes, multi-scale techniques are of great interest among NE methods as they account for the global HD organization to define the LD space, delivering outstanding DR quality. Their time complexity in the number of data samples however prevents tackling large-scale databases. The thesis addresses this difficulty by presenting fast multi-scale NE algorithms which account for the dense nature of the multi-scale similarities, providing high quality embeddings of very big data sets.

The robust and fast NE algorithms designed in this thesis hence open the path to enhanced HD data exploration in ML through visualization.