WinSitu, un nouveau paradigme pour l'analyse exploratoire de données basée sur des projections

Michaël Aupetit*

*CEA, LIST, Laboratoire Information Modèles et Apprentissage,
F-91191 Gif-sur-Yvette, France.
michael.aupetit@cea.fr

Résumé. Dans cet article, nous discutons des limites pratiques de l'analyse exploratoire de données basée sur les techniques de projection non linéaires continues. Nous montrons que ces méthodes de projection sont inutilisables en l'état pour permettre une inférence quelconque sur les données originelles. Nous présentons une méthode de visualisation *in situ* et montrons au travers de différentes expériences, qu'elle est indispensable à leur interprétation. Ce processus implémente le paradigme *WinSitu* d'analyse exploratoire visuelle basée sur des projections que nous introduisons pour la première fois dans ce travail. Ce changement de paradigme permet de rendre aux méthodes de projection toute leur utilité.

1 Introduction

L'objectif essentiel de l'analyse exploratoire visuelle de données est d'inférer de leur représentation graphique des propriétés sur leur structure originelle. Ces données sont fournies sous forme d'un tableau individus-variables $N \times D$ ou d'une matrice de dissimilarités inter-individus $N \times N$. Deux types d'informations sont recherchées :

- Des classes de variables similaires, ou des variables atypiques, mises en évidence par un ensemble d'individus.
- Des classes d'individus similaires, ou des individus atypiques, mis en évidence par un ensemble de variables.

La similarité entre variables peut être mesurée par exemple par la corrélation de Fisher ou de Pearson. Celle entre individus peut être mesurée par exemple par la distance euclidienne, ou par la distance d'édition ou fournie directement sous forme matricielle. Lorsque les individus sont représentés sous forme d'un nuage de points dans un repère cartésien orthonormé à deux dimensions, la mesure de corrélation entre les variables codées par les axes de ce repère peut être estimée visuellement, de même que la présence d'individus atypiques ou de classes d'individus similaires apparait aussi immédiatement. Les difficultés surviennent lorsque l'on cherche à analyser des données multi-dimensionnelles.

L'analyse exploratoire de données multi-dimensionnelles passe soit par l'estimation automatique de paramètres d'un modèle fournissant un résumé de telle ou telle caractéristique recherchée dans les données (adéquation à une loi de densité de probabilité, degré d'appartenance à des classes de structure prédéfinie, classification des variables ou des individus...)