# Réduction des dimensions des données en apprentissage artificiel

Y. Bennani, S. Guérif, E. Viennet

Université Paris 13, LIPN - CNRS UMR 7030
99, avenue J.B. Clément, F-93430 Villetaneuse

**Résumé.** Depuis plusieurs décénies, le volume des données disponibles ne cesse de croître ; alors qu'au début des années 80 le volume des bases de données se mesurait en mega-octets, il s'exprime aujourd'hui en tera-octets et parfois même en peta-octets. Le nombre de variables et le nombre d'exemples peuvent prendre des valeurs très élevés, et cela peut poser un problème lors de l'exploration et l'analyse des données. Ainsi, le développement d'outils de traitement adaptés aux données volumineuses est un enjeu majeur de la fouille de données. La réduction des dimensions permet notament de faciliter la visualisation et la compréhension des données, de réduire l'espace de stockage nécessaire et le temps d'exploitation, et enfin d'identifier les facteurs pertinents. Dans cet article, nous présentons un panarama des techniques de réduction des dimensions essentiellement basées sur la sélection de variables supervisée et non supervisée, et sur les méthodes géométriques de réduction de dimensions.

## 1 Introduction

La taille des données peut être mesurée selon deux dimensions, le nombre de variables et le nombre d'exemples. Ces deux dimensions peuvent prendre des valeurs très élevées, ce qui peut poser un problème lors de l'exploration et l'analyse de ces données. Pour cela, il est fondamental de mettre en place des outils de traitement de données permettant une meilleure compréhension de la valeur des connaissances disponibles dans ces données. La réduction des dimensions est l'une des plus vieilles approches permettant d'apporter des éléments de réponse à ce problème. Son objectif est de sélectionner ou d'extraire un sous-ensemble optimal de caractéristiques pertinentes pour un critère fixé auparavant. La sélection de ce sous-ensemble de caractéristiques permet d'éliminer les informations non-pertinentes et redondantes selon le critère utilisé. Cette sélection/extraction permet donc de réduire la dimension de l'espace des exemples et rendre l'ensemble des données plus représentatif du problème. En effet, les principaux objectifs de la réduction de la dimension sont :

- faciliter la visualisation et la compréhension des données,
- réduire l'espace de stockage nécessaire,
- réduire le temps d'apprentissage et d'utilisation,
- identifier les facteurs pertinents.

Les algorithmes d'apprentissage artificiel requièrent typiquement peu de traits (*features*) ou de variables (attributs) très significatives caractérisant le processus étudié. Dans le domaine