Logiciel « DtmVic »
Data and Text Mining: Visualisation, Inférence, Classification

Ludovic Lebart
Telecom-ParisTech, 46 rue Barrault, 75013, Paris
ludovic@lebart.org

1 Brève description

Ce logiciel est consacré à la visualisation des données multidimensionnelles, que ces données soient numériques, nominales ou textuelles. Les limitations de la version actuelle sont : 22 500 lignes (individus, observations), 1000 colonnes (variables numériques, variables nominales – une variable nominale = une colonne), 100 000 caractères pour les réponses textuelles d'un individu. Pour ce faire, dix-huit enchaînements de base sont proposés à l'utilisateur. On en décrira deux.

Exemple de l'enchaînement « PCA » : Analyse en composantes principales, classification des individus en k classes, description automatique des classes par les variables actives et illustratives. Le volet « VIC » (Visualisation, Inférence, Classification) permet alors d'obtenir des graphiques, des zones de confiance bootstrap, des cartes auto-organisées , etc.

Exemple de l'enchaînement « VISUTEX » : Analyse des correspondances de la table de contingence croisant les textes (données de base) et les mots les plus fréquents, mots caractéristiques des textes, lignes ou phrases caractéristiques des textes. Sériation de la table (re-ordonnancement des lignes et des colonnes). Mêmes compléments à partir du volet VIC.

2 Spécificité, accès

Le domaine d'application « coeur de cible » est *« le traitement statistique des enquêtes comportant des questions fermées et ouvertes ».*

- ✓ Complémentarité systématique des techniques de visualisation (Analyse en composantes principales, Analyse des correspondances simples et multiples) et de la classification automatique (méthode mixte combinant classification hiérarchique [critère de Ward] et centres mobiles [k-means]; cartes auto-organisées de Kohonen).
- ✓ Validation des techniques de visualisation : Ré-échantillonnage (bootstrap, bootstrap partiel, bootstrap total, bootstrap sur variables).
- ✓ Mise en oeuvre des méthodes d'Analyse de contiguïté et méthodes connexes.
- ✓ Prétraitement de texte (indépendant de la langue) : fusions et suppressions de mots.

La présente version de ce logiciel académique est accompagnée d'une batterie de 27 jeux de données. *[Les treize premiers exemples d'application sont commentés dans un tutoriel intégré au logiciel].* Pour la partie numérique, l'ouvrage de référence est : « *Statistique Exploratoire multidimensionnelle – Visualisation et Inférence en Fouilles de données* » par L. Lebart, M. Piron, A. Morineau, Dunod, 2006. Pour la partie textuelle : « *Statistique textuelle* » par L. Lebart et A. Salem, Dunod, 1994, téléchargeable en pdf à partir du site. Téléchargement libre de DtmVic: (version 4.1 de DTM) Site www.lebart.org.