DATA

hack A tech
Shake science. Shape innovation.

#datascience
#coclustering
#visualisation
#bigdata

BLOCKCLUSTER : grandes découvertes en grande dimension

Inria

CARACTÉRISTIQUES

BlockCluster est un logiciel qui permet de classifier des données de grande dimension (plusieurs milliers ou dizaines milliers de variables ou de descripteurs), là où la plupart des autres méthodes travaillent sur des données de dimensions plus réduites (plusieurs centaines de descripteurs en général).

Par ailleurs, BlockCluster autorise nativement le traitement sur des tableaux de données de divers types :

- Continues (valeurs de capteurs sur de grands équipements)

- Comptage/entier (analyse du contenu de grands textes)

- Catégorielles (analyse de très grands questionnaires, données de navigation web)

TRAITEMENT DES DONNÉES

Le logiciel détermine à la fois des groupes d'individus (les lignes du tableau de données) et des groupes de variables (les colonnes du tableau de données).

Cette classification non supervisée dite « croisée » permet d'obtenir d'une part des résultats valides en très grande dimension et d'autre part d'offrir à tous types d'utilisateurs une compréhension simplifiée de très grands tableaux
(les classes individus/variables produisent des visualisations « naturelles » par réorganisation du tableau de données).

QUELS AVANTAGES?

- Gestion de très grands tableaux de données, en particulier avec beaucoup de variables
- Interprétabilité des résultats

USE CASES

Tous les domaines d'activité, dont :

Marketing, (cyber)sécurité : découverte d'activités sur le web par historique de navigation,

Santé : analyse de données génomiques ,

Industrie, transport : traitement de données textuelles non structurées (annotations « terrain » par exemple).

DATA

FICHE IDENTITÉ

- **Licence :** GPL
- **Langage de programmation:** C++
- **Propriété intellectuelle :** Inria - Université de Lille – Université de Technologie de Compiègne – CNRS
- **Équipe projet** : Modal* - https://www.inria.fr/fr/modal

FONCTIONNALITÉS GÉNÉRIQUES

Le logiciel implémente le principe dit du latent block model (LBM) pour produire la classification croisée non supervisée. Les modèles disponibles pour décrire les données sont les suivants : lois multinomiale, Gaussienne, et de Poisson.

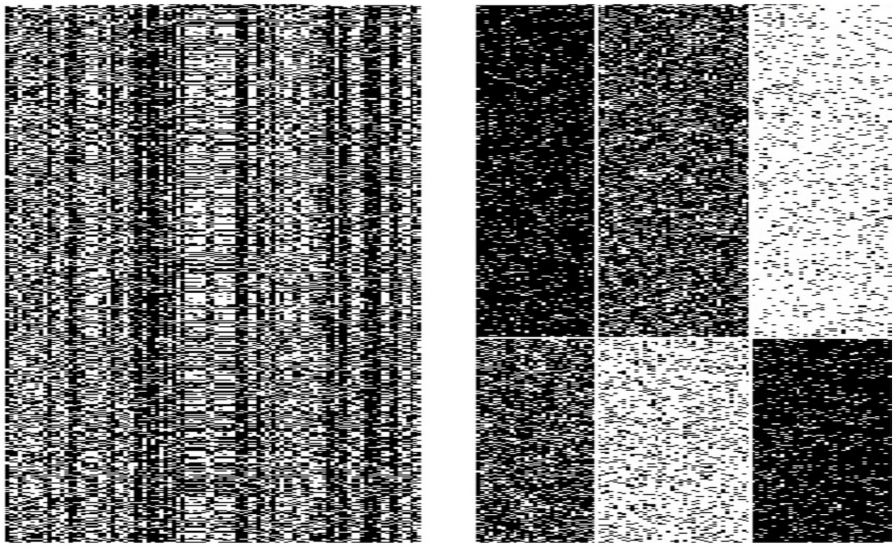
L'interprétation des classes se voit aussi bien par visualisation que par lecture des paramètres du modèle. L'utilisation d'un critère de choix du nombre de classes lignes/colonnes (critère ICL) assure la pertinence de l'analyse.

Inputs :

Fichier Data (données fortement multivariées à classifier, types continus / comptage / catégorie s autorisés) : .csv

Outputs :

Paramètres constituant les classes lignes / colonnes, valeurs d'ICL (Integrated Classification Likelihood) pour connaître le nombre de classes. Sortie graphique customisée sous R



READ ME

Package R sur le CRAN : http://cran.r-project.org/web/packages/blockcluster/index.html

Essai rapide en mode SaaS (plateforme MASSICCC) : https://massiccc.lille.inria.fr

Référent : Christophe Biernacki

* Modal est une équipe-projet commune à Inria et au laboratoire Paul Painlevé (CNRS, Université de Lille)

© Inria / Photo Kaksonen

DATA

hackAtech
Shake science. Shape innovation.

#bigdata
#clustering
#prediction
#datascience

MIXTCOMP
Faites parler vos données !

Inria

CARACTÉRISTIQUES

Données hétérogènes

MixtComp permet de traiter des données hétérogènes (nombre, intervalle, courbe...), là où la plupart des autres méthodes travaillent sur des données homogènes, donc d'un seul type.

Traitement de données multivariées fortement hétérogènes :
- Continues (indicateur de température)
- Comptage/entier (nombre d'enfants)
- Catégorielles (statut marital : marié, pacsé, célibataire...)
- Rangs (classement de lieux de vacances par ordre de préférence)
- Fonctionnelles (température en fonction de l'heure)

Données manquantes et imprécises

SmartData, démonstrateur à Interface utilisant le logiciel MixtComp

TRAITEMENT DES DONNÉES

Classification non supervisée
Le logiciel détermine des groupes partageant des caractéristiques communes.

Prédiction
Les résultats de la classification peuvent être utilisés pour faire de la prédiction, sur des données manquantes ou partielles.

Interprétabilité
Contrairement à beaucoup d'autres approches, MixtComp permet d'interpréter nativement les groupes issus de la classification.

USE CASES

Tous les domaines d'activité, dont :

Santé : analyse des effets de traitement
Retail : prédiction des stocks
Marketing : segmentation client
Industrie : identification de modes de fonctionnement...

QUELS AVANTAGES?

- Interprétabilité des résultats
- Gestion des données hétérogènes et manquantes

DATA

FICHE IDENTITÉ

- **Licence :** AGPL3
- **Langage de programmation:** C++
- **Propriété intellectuelle :** Inria - Université de Lille - CNRS
- **Équipe projet :** Modal* - https://www.inria.fr/fr/modal

FONCTIONNALITÉS GÉNÉRIQUES

Use case	Objectifs
Je sais à quelle classe appartient chaque objet.	Définition des paramètres (caractéristiques) constituant les classes. Aide à l'interprétation des classes existantes.
Je connais le nombre de classes mais je ne sais pas à quelle classe appartient chaque objet.	Affectation de chaque objet à une des classes. Évaluation du risque de ce classement (probabilité d'erreur).
Je ne connais pas le nombre de classes.	Recherche du nombre de classes optimal. Définition des paramètres (caractéristiques) constituant les classes correspondantes : aide à l'interprétation de ces nouvelles classes.

Les modèles disponibles pour décrire les données sont les suivants : Lois multinomiale, Gaussienne, Poisson, Weibull, Negative binomiale ainsi que les modèles Insertion Sort Algorithm et ClustSeg.

· Détection, extraction de structures dans les données. Imputation des données manquantes ou des intervalles.

· Description des classes à partir:

- des paramètres (caractéristiques) estimés constituant les classes : interprétation des classes,
- des valeurs de critères de choix du nombre de classes (ICL, BIC...) : pertinence d'un nombre de classes.

Apprentissage / Classification

Inputs :

- Fichier *Data* (données multivariées hétérogènes à classifier) : .csv, .json
- Fichier *Descripteur des modèles* à utiliser sur les données : .csv.

Outputs :

- Paramètres constituant les classes, valeurs d'ICL (Integrated Classification Likelihood) et BIC (Bayesian Information Criteria) : .json, sortie graphique customisée sous R.

Prédiction

Inputs :

- Fichier Data (nouvelles série de données) : .csv.
- Fichier Descripteur des modèles à utiliser sur les données : .csv.

Outputs :

- Probabilité pour la nouvelle série de données d'appartenir aux différentes classes apprises sur une base historique : .json et sortie graphique customisée sous R.

READ ME

https://github.com/modal-inria/MixtComp

Essai rapide en mode SaaS (plateforme MASSICCC) : https://massiccc.lille.inria.fr

Référent : Christophe Biernacki

* Modal est une équipe-projet commune à Inria et au laboratoire Paul Painlevé (CNRS, Université de Lille)

DATA

#basededonnéesgraphes

#straming

#NoSQL

#PostgreSQL

hack A tech

Shake science. Shape innovation.

Data management

Inria

Extraire, ranger et nettoyer des gros jeux de données pour préparer leur analyse.

PROCESSUS

Prépare et nettoie des gros jeux de données bruitées, afin de les exploiter, visualiser, analyser en amont du traitement (statistique, ML, etc.)

Identifier une ou plusieurs sources de données à collecter (en fonction des objectifs de traitement), par exemple :

- Aligner des données commerciales entre plusieurs filiales, plusieurs logiciels et des sources extérieures (web)
- Construire des schémas de BDD efficaces pour la mise à jour centralisée de plusieurs sources de données
- Construire des schémas de BDD distribuées pour la scalabilité horizontale de l'infrastructure

© Inria / Photo Raphaël de Bengy

USE CASES

- fusion-acquisition de données
- dans le retail : recherche d'informations peu structurée pour les moteurs de recherche (indexation, web sémantique)
- extraction et ingestion d'informations pour la visualisation et le clustering de données
- construction de dashboard data visualisation pour l'ADULM pour l'analyse de données géospatiales
- débogage de schémas postgreSQL pour l'INSERM durant la crise Covid

QUELS AVANTAGES ?

Construction de pipelines de données efficaces pour permettre leur mise à jour, leur réutilisation dans plusieurs contextes, ainsi que leur évolution.

DATA

FICHE IDENTITÉ

- Utilisation avancée de PostgreSQL (distribution / intégration ML in DB / intégration ETL)
- Extraction d'informations en streaming haute performance (optimisation SIMD bas niveau) Intégrations PostgreSQL et Python, C
- Utilisations de bases de données graphes (RDF, web sémantique) Intégration de NetworkX (Python)

- Équipe projet : Links* - https://www.inria.fr/fr/links

FONCTIONNALITÉS GÉNÉRIQUES

À partir de plusieurs sources de données hétérogènes, on peut construire un jeu de données cohérent et prêt à être traité efficacement.

Deux exemples de logiciels :

NetworkDisk

NetworkDisk est une librairie Python qui permet de manipuler des graphes directement sur le disque. L'objectif est d'être parfaitement compatible avec la librairie NetworkX, tout en allégeant les contraintes de mémoire et en apportant de la persistance pour les graphes. NetworkDisk est conçu pour les utilisateurs de NetworkX souhaitant manipuler des graphes sans avoir à se soucier des technologies de bases de données associées ni avoir à apprendre de nouveaux langages de bases de données.

Shex

Le langage ShEx (Shape Expressions) fournit un schéma structurel pour les données RDF (Resource Description Framework, modèle de graphe destiné à décrire formellement les ressources Web et leurs métadonnées, afin de permettre le traitement automatique de telles descriptions. - Wikipédia). Ce langage peut être utilisé pour documenter des APIs ou des jeux de données, ou explorer des jeux de données hétérogènes.

READ ME

https://networkdisk.inria.fr/

https://github.com/iovka/shex-java

Référent : Charles Paperman

* Links est une équipe-projet commune à Inria et au laboratoire CRIStAL (Centrale Lille, CNRS, Université de Lille)

INSTITUT CARNOT Inria

UNION EUROPÉENNE

l'Europe s'engage en Hauts-de-France avec le FEDER