DATA

hackAtech

Shake science. Shape innovation.

#datascience

#coclustering

#visualisation

#bigdata

05-07 Mars 2020

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).

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)

QUELS AVANTAGES ?

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

FICHE IDENTITÉ

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

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.

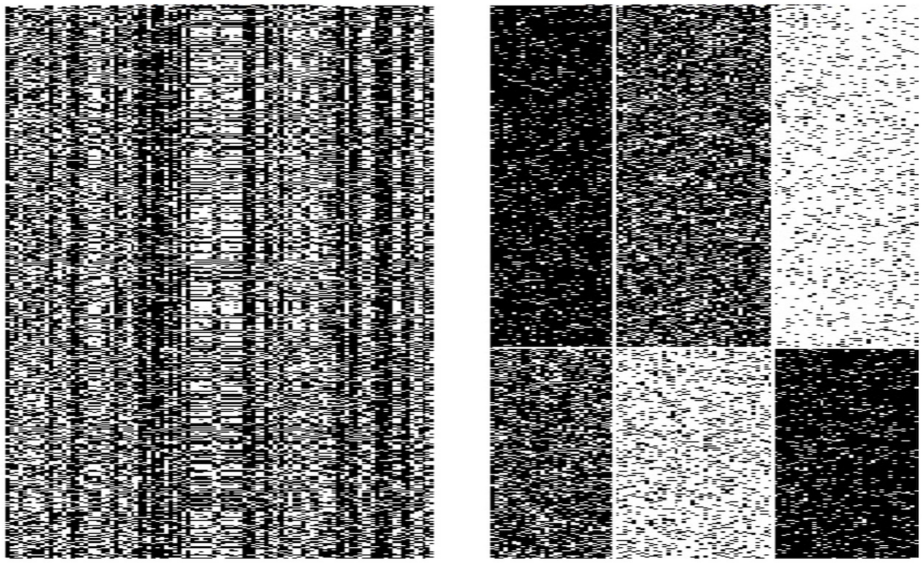
L'interprétation des classes se voit aussi bien par visualisation que par lecture des paramètres du modèle. L'utilisation d'un critère de choix du nombre de classes lignes/colonnes (critère ICL) assure la pertinence de l'analyse.

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)

INSTITUT CARNOT Inria

UNION EUROPÉENNE

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

Inria

DATA

hackAtech

Shake science. Shape innovation.

05-07 Mars 2020

#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

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.

QUELS AVANTAGES ?
- Interprétabilité des résultats
- Gestion des données hétérogènes et manquantes

SmartData, démonstrateur à Interface utilisant le logiciel MixtComp

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...

FICHE IDENTITÉ

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

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

hackAtech

Shake science. Shape innovation.

05 - 07 Mars 2020

#LinkedData

#Integration

#Decouvrir

#Structure

Shape Designer : Décrire et découvrir la structure de données

Inria

CARACTÉRISTIQUES

Le Web des données est un immense réseau de données et bases de connaissances libres et liées entre elles, gérées par des organisations publiques ou privées dans divers domaines d'activité.

Par exemple, il y a UniProt sur les protéines et leurs propriétés, GeoNames pour les données géographiques, Europeana une plateforme de données culturelles, ou encore WikiData, la plus grande base de données libre et collaborative créée sur le modèle de Wikipédia.

Les données utilisées sont des données semi-structurées, qui peuvent être par exemple une adresse mail, la civilité d'une personne, son métier, etc. Elles peuvent être libres ou propriétaires. Leur point commun est qu'on n'en connait pas la structure.

USE CASES

Il est possible d'utiliser la technologie pour n'importe quel domaine, par exemple :

Santé : utiliser les données existantes de protéines ou de maladies pour comprendre leur structure et leur contenu dans le but de les intégrer dans sa propre analyse.

Marketing : proposer des commerces à un utilisateur en fonction de sa position géographique. On peut donc utiliser ses propres données de recommandation (base de données de commerces dans une ville par exemple) enrichies par des bases de données géographiques.

PROCESSUS

Tout le monde peut contribuer à ces données, et tout le monde peut les utiliser. La difficulté consiste à intégrer de telles données hétérogènes dans son application.

Shape Designer permet de prendre un ensemble de données existant, provenant du web des données, et de découvrir sa structure. L'objectif, si l'on veut intégrer ces données à son application, est d'appréhender leur contenu afin de pouvoir les utiliser.

Par intégration, nous entendons lier des données extérieures à ses propres données, sans modifier l'une ou l'autre. Cela permet d'enrichir ses données, à partir de données extérieures. Les requêtes se font alors dans un ensemble dynamique et évolutif, constitué des différentes bases de données.

QUELS AVANTAGES ?

- Être capable d'intégrer ce genre de données dans sa propre application, ou inversement être capable d'enrichir des données existantes avec ses propres données.
- S'appuie sur les standards du web sémantique. Il n'est pas nécessaire d'avoir des connaissances préalables sur les standards pour utiliser l'outil.

FICHE IDENTITÉ

- Licence : Open source
- Langage de programmation : Multiplateforme (Java).
- Format des données : l'outil travaille sur du RDF
- Connaissances : Connaitre le Resource Description Framework (RDF), format de données pour le Web des données

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

FONCTIONNALITÉS GÉNÉRIQUES

- Définir simplement un schéma qui décrit la structure d'un dataset RDF.
ShapeDesigner permet de définir le schéma à partir d'exemples, sans connaissance préalable d'un langage de schémas. Il peut également s'utiliser pour apprendre un langage de schémas.

- Valider un dataset par rapport à un schéma.
L'interface graphique permet d'explorer directement les erreurs de validation et éditer les données pour les corriger.

- Extraire automatiquement le schéma d'un dataset existant. Cette fonctionnalité permet entre autres d'obtenir une description succincte du type de données contenues dans un dataset, et leur structure.

L'outil fonctionne avec des données disponibles localement ou à travers d'un SPARQL endpoint.
Il supporte les langages de schémas Shape Expressions et SHACL.

READ ME

https://gitlab.inria.fr/jdusart/shexjapp

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

DATA

hackAtech

Shake science. Shape innovation.

#heuristique
#optimisation
#multiobjectif
#parallélisation

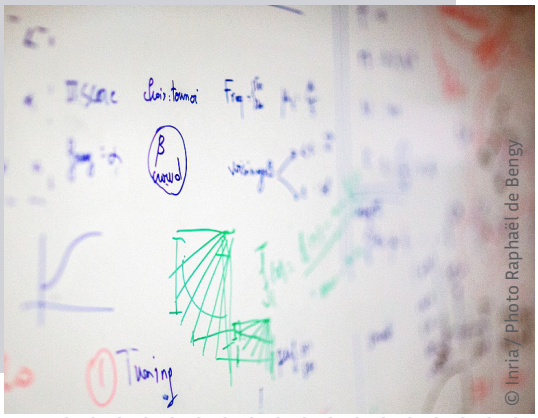
05-07 Mars 2020

Big Optimization et Calcul Haute Performance

Inria

CARACTÉRISTIQUES

Avec la quantité de données actuelle, un enjeu important est l'optimisation, c'est à dire résoudre un problème complexe en tenant compte des alternatives possibles afin d'obtenir le résultat le plus satisfaisant. On parle de «big optimization» quand il y a un grand nombre d'alternatives, de contraintes à respecter. Le focus est mis sur les problèmes de grande dimension en nombre de variables de décision et/ou de critères à optimiser, et/ou coûteux en temps d'évaluation des critères.



PROCESSUS

Le principe est donc d'identifier le problème d'optimisation, le modéliser, concevoir et implémenter une méthode de résolution parallèle approchée (méta-heuristique, c'est à dire en utilisant des algorithmes qui s'inspirent de la nature) ou exacte, et enfin la valider par expérimentation.

Optimisation multiobjectif

Il y a ici la prise en compte simultanée de plusieurs critères d'optimisation, souvent contradictoires, pour générer un ensemble de solutions.

Optimisation parallèle

Le calcul parallèle est indispensable pour adresser des problèmes de grande dimension, avec un grand nombre d'alternatives.
Le défi est de concevoir et implémenter des méthodes d'optimisation efficaces pour les environnements massivement parallèles et hétérogènes (multi-coeurs, GPUs, ARM...).
L'optimsation parallèle permet de décomposer le problème et/ou les objectifs afin de réduire le temps de calcul.

USE CASES

- Domotique : réalisation d'un algorithme permettant de gérer le fonctionnement de l'ensemble des appareils électriques de la maison en optimisant le rapport confort utilisateur/consommation énergie.
- Bio-médical : conception d'un logiciel permettant le dimensionnement de grands laboratoires (placement d'équipements) et la gestion temps réel optimisée des flux d'échantillons à analyser.

QUELS AVANTAGES ?

- Produire à moindre coût des solutions de bonne qualité,
- Résoudre des problèmes de grande dimension,
- Aider à la prise de décision multi-critère.

FICHE IDENTITÉ

- Logiciel : C/C++, python, CUDA, MPI, R, Multi-GPU, OpenMP, Chapel, ParadisEO

- Équipe projet : Bonus* - https://team.inria.fr/bonus/

FONCTIONNALITÉS GÉNÉRIQUES

Identifier la problématique d'optimisation.

Étape clé, elle permet d'identifier et comprendre la nature du problème et de définir les différents niveaux de complexité. Un problème industriel est souvent composé de plusieurs objectifs. Selon les problématiques, ils peuvent être traités successivement, chacune influençant les autres, ou simultanément : on parle d'optimisation multiobjectif.

Modéliser et implémenter.

La première phase est celle de modélisation de la solution et de la fonction objective adéquate au problème. La modélisation comprend la formulation mathématique des objectifs et des contraintes. Elle permet de donner une orientation à l'algorithme pour répondre aux besoins. Une fois la problématique identifiée et modélisée, on passe à la phase d'implémentation des différentes heuristiques et métaheuristiques ainsi que les approches de parallélisation possibles pouvant le résoudre. Cette phase se base sur l'expertise de l'équipe dans ce domaine.

Analyser les résultats.

Une fois les premiers résultats obtenus, une phase importante d'analyse et de paramétrage est nécessaire pour affiner l'approche. Cette phase se base sur l'expérimentation et l'analyse de ces résultats. Elle permet de trouver le bon équilibre entre l'exploration et l'exploitation de de l'espace de recherche.

Généraliser la solution.

La dernière étape est la proposition d'une solution avec une liste de paramètres à manipuler qui permettra à l'entreprise d'adapter l'approche pour différents sénarios possibles.

READ ME

T. CARNEIRO, J. GMYS, N. MELAB, D. TUYTTENS. Towards ultra-scale Branch-and-Bound using a high-productivity language, in «Future Generation Computer Systems», November 2019 https://hal.archives-ouvertes.fr/hal-02371238

S. CAHON, N. MELAB, E. TALBI. ParadisEO: A Framework for the Reusable Design of Parallel and Distributed Metaheuristics, in «J. Heuristics», 2004, vol. 10, no 3, pp. 357–380

O. ABDELKAFI, L. IDOUMGHAR, J. LEPAGNOT, J-L PAILLAUD, I. DEROCHE, L-A BAUMES, P. COLLET. Using a novel parallel genetic hybrid algorithm to generate and determine new zeolite frameworks. Computers and Chemical Engineering, Elsevier, 2017, 98, pp.50 - 60. https://hal.archives-ouvertes.fr/hal-01665051

Référent : Omar Abdelkafi

*Bonus est une équipe-projet Inria 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

Inria

DATA

hackAtech

Shake science. Shape innovation.

05 - 07 Mars 2020

#privacy

#decentralizedML

#moyenne

#GOPA

MyLocalInfo : calculer des statistiques d'enquête sans collecter les réponses individuelles

Inria

PROCESSUS

La quantité de données personnelles collectées lors de nos interactions quotidiennes avec des objets connectés offre de grandes opportunités pour le développement de services innovants, nourris par le machine learning, mais suscite aussi des inquiétudes sérieuses pour le respect de la vie privée des individus.

Le protocole GOPA, sur lequel repose l'application MyLocalInfo, est un protocole distribué qui permet à un grand nombre d'utilisateurs de calculer les valeurs moyennes de leurs données respectives sans les révéler, et sans que la moyenne ne soit calculée par un serveur central. Cela peut alors servir pour apprendre des modèles prédictifs.

Gopa permet de réaliser un calcul d'agrégat (comme la somme ou la moyenne) d'un ensemble de valeurs, chacune étant détenue par un utilisateur ne voulant pas la révéler. À l'issue de l'exécution de Gopa, tous les utilisateurs participant au calcul obtiennent le résultat, mais aucun ne peut déduire la valeur détenue par les autres participants. Les utilisations potentielles de cet algorithme sont de pouvoir réaliser des sondages, des votes, respectueux de la vie privée, c'est-à-dire en évitant le transfert de données sensibles vers un unique centre de stockage des données.

QUELS AVANTAGES ?

- Ne repose pas sur une tierce partie
- Assure la confidentialité face à un adversaire curieux ou malveillant

Une procédure de vérification offre une protection contre les utilisateurs malveillants qui utiliseraient le service dans le but de manipuler le résultat de l'algorithme.

USE CASES

Par exemple, dans le domaine médical, alors que les enquêtes auprès des patients jouent un rôle crucial, l'exactitude des données recueillies est meilleure si le sujet sait que ses réponses ne seront pas connues par d'autres personnes.

En particulier pour certaines questions sensibles : « Avez-vous pris correctement votre traitement ? » , « suivez-vous un régime ? », « étiez-vous en colère ? » ...

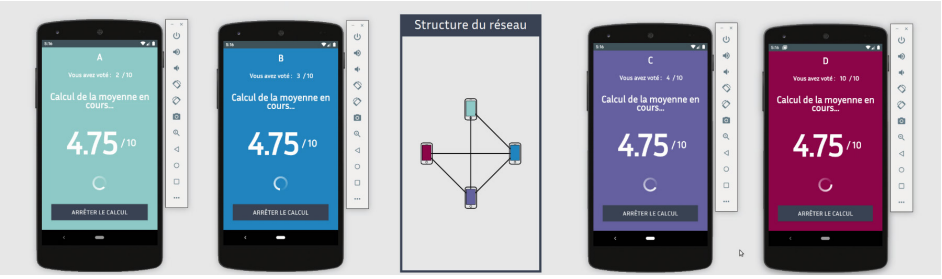
Notre technologie peut, dans ce cas, contribuer aux efforts des chercheurs pour renforcer la confiance des patients.

D'autres exemples existent dans les domaines du transport, marketing...

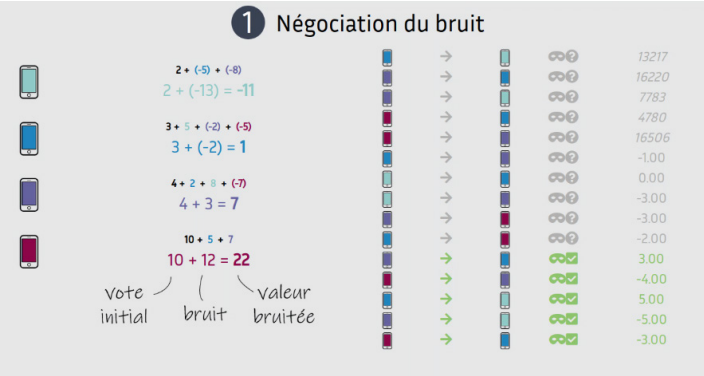
FICHE IDENTITÉ

- Langage de programmation : Java
- Équipe projet : Magnet* - https://team.inria.fr/magnet/

FONCTIONNALITÉS GÉNÉRIQUES

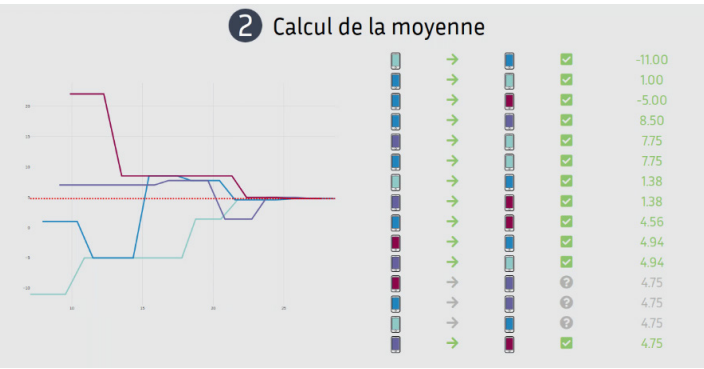


La version actuelle est une librairie Java, utilisée dans une application minimale Android qui calcule la moyenne entre les différentes valeurs stockées sur les devices des utilisateurs, sans révéler ses valeurs. Dans la simulation, des utilisateurs participent à un vote sur smartphone.



Lorsque les quatre utilisateurs ont voté, la première phase, la « négociation du bruit » se lance :
- B choisit une valeur (ex : 5), la crypte et envoie la valeur cryptée à A : il s'engage sur la valeur sans la dévoiler.
- A choisit une valeur (ex : 0) et l'envoie à B en clair (non cryptée).
- B envoie sa valeur (5) en clair à A. A crypte cette valeur avec la clé publique de B. A vérifie que cette valeur cryptée correspond bien à la valeur envoyée par B à l'étape 1.
- Le bruit final négocié entre A et B est la somme des 2 valeurs échangées (5 + 0 = 5). A va utiliser la valeur positive de ce bruit (5) et B, la valeur négative (-5).

Ce protocole de négociation s'effectue entre chaque paire de voisins. Chaque utilisateur va ajouter le bruit négocié avec chacun de ses voisins à la valeur initiale votée par l'utilisateur.



La deuxième phase, le « calcul de la moyenne », peut alors débuter :
- A envoie à l'un de ses voisins B sa valeur bruitée (ex : -11).
- B fait la moyenne entre la valeur reçue par A et sa propre valeur bruitée (ex : 1). La nouvelle valeur détenue par B est le résultat de cette moyenne ((-11)+1)/2 = (-5). Sur l'écran du téléphone B, la valeur s'actualise (elle passe à (-5)).
- B transmet cette nouvelle valeur à A sous forme d'acknowledge. A met également à jour sa valeur avec la nouvelle valeur reçue. Sur l'interface, le message original passe en vert. Sur l'écran de A, la valeur s'actualise (elle passe également à (-5)).

Ce protocole s'effectue en « pair-à-pair » entre chaque paire de voisins. Petit à petit, les valeurs détenues par les 4 utilisateurs vont converger vers la moyenne finale attendue.

READ ME

https://arxiv.org/abs/1803.09984

Référent : Jan Ramon

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


© Inria / Photo C. Morel

INSTITUT CARNOT Inria

UNION EUROPÉENNE

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

Inria