Acronyme du projet : ADOC

Titre long du projet : Agrégation des DOnnées par Comparaisons Stochastiques

Nom du porteur et email : Johanne Cohen,

Mots clefs du projet : Agrégation, Comparaison stochastique, Monotonie stochastique, Algorithmique, Complexité.

Problématique adressée en 2017 et résultats saillants obtenus

Les ordres stochastiques : théorie et algorithmes

De nombreux traitements ne peuvent être réalisés de façon efficace sur des masses de données à cause de la complexité algorithmique combinée à la taille des données. Plutôt que sélectionner aléatoirement une partie de ces données, nous nous proposons de les agréger par des algorithmes qui fournissent des bornes stochastiques. Un des points forts et originaux de cette approche est que cette compression de données peut être effectuée en garantissant des bornes sur un traitement en prouvant sa monotonie par rapport à l'ordre considéré. Nous utilisons les ordres stochastiques pour agréger les données dans une distribution discrète avec un plus petit support, mais qui constitue une borne au sens de l'ordre stochastique considéré. C'est pourquoi nous avons développé une approche algorithmique des ordres stochastiques permettant d'agréger les données tout en garantissant l'ordre. Ces algorithmes sont polynomiaux et une de nos perspectives est d'avoir une complexité faible pour traiter des données de grande taille.

Nous avons, au cours de cette première année, conçu et prouvé des algorithmes pour les ordres de variabilité : convexe ou stop loss « ***cx*** », et concave « ***cv*** ». L'ordre convexe permet de considérer une distribution qui aura la même moyenne, mais dont la variance sera réduite (pour une borne inférieure) ou augmentée (pour une borne supérieure). Ces ordres permettent de borner l'aversion au risque (à moyenne constante) car la variance des échantillons traduit souvent un risque (au sens large) pour le phénomène étudié.

Plus précisément nous avons déterminé plusieurs opérations applicables aux atomes d'une distribution discrète et qui permettent de construire une distribution avec un atome de moins. Ces opérations sont assez intuitives comme le montrent les deux exemples suivants :

- Prendre trois atomes quelconques, enlever l'atome central et répartir sa masse sur les deux atomes extrêmes pour conserver la moyenne permet de construire une borne supérieure au sens convexe.
- Prendre deux atomes et les remplacer par un atome barycentre avec la même moyenne permet de construire une borne inférieure au sens convexe.

De plus nous avons prouvé un algorithme de type programmation dynamique qui permet de combiner ces actions élémentaires pour calculer une borne supérieure optimale (la plus précise pour une fonction de récompense donnée). Nous avons donc le choix entre des heuristiques utilisant les opérations élémentaires dans une séquence quelconque (mais obtenue rapidement) ou construire la séquence optimale des opérations (au prix d'une complexité algorithmique supérieure). Ce travail fait l'objet de la publication dans une revue internationale [1]. Deux stagiaires ont été associés à cette publication.

Les applications aux workflows

Nous comptons employer ces ordres de variabilité ainsi que l'ordre stochastique fort pour analyser des phénomènes d'attente (pour les patients dans les hôpitaux) ou de durée de calcul (pour un workflow). Dans ces modèles, il est classique de faire apparaître des séquences récursives stochastiques sur des opérateurs croissants et convexes (par exemple l'équation de Loynes pour une file d'attente). Cela suffit pour démontrer la monotonie au sens de ces ordres stochastiques et donc de garantir que si nous bornons la distribution des entrées, nous obtienons une distribution bornante de la sortie de l'algorithme comme nous l'avons déjà montré dans [1] pour des temps d'exécution de graphes de tâches. .

Un résultat obtenu mais non encore publié montre qu'un schéma d'agrégation classique consistant à remplacer une distribution empirique par des moyennes partielles introduit un biais systématique pour l'ordre convexe.

Nous avons appliqué le calcul de la borne supérieure d'une distribution au temps d'exécution de tâches à durée aléatoires dans les workflows scientifiques. Pour cela, nous avons travaillé avec des enseignants-chercheurs de l'université de Sao Paulo (Daniel Cordeiro, (laboratoire muti-disciplinaire) EACH), Kelly Braghetto (département de mathématique et statistique), qui travaillent sur les exécutions de ces workflows scientifiques (par exemple [X1]). L'objectif est de concevoir un algorithme calculant une distribution de temps de calcul dans les workflows. Pour cela, nous nous sommes concentrés sur les traces d'exécutions sur deux workflows particuliers (un lié à de l'analyse d'images [X2], l'autre pour les opérations de séquençage [X3] du génome). Actuellement, un nouvel algorithme (non encore publié) a été implémenté prenant en compte des propriétés structurelles de ces workflows (ce sont, ou ils contiennent, des graphes orientés série-parallèles). Ce travail est en cours de réalisation.

Les trajectoires de mobilité : l'enquête étudiée par le laboratoire PRINTEMPS sur la thématique Mobilité utilise des trajectoires de mobilité (liste de déplacements, de leur durée et de leur objet) ainsi que des données sociologiques sur les acteurs. L'idée est de sélectionner quelques individus les plus représentatifs. L'algorithme employé jusqu'à maintenant calcule une similarité (distance de Levenshtein) entre trajectoires (grâce à un algorithme issu de la comparaison de séquences génétiques) avant d'utiliser une méthode de classification. Cette distance prend en compte les insertions, destructions ou mutations d'une activité au sein d'une trajectoire et les poids élémentaires de ces opérations sont arbitraires. Les trajectoires retenues pour le moment se focalisent sur la séquence d'activité et non pas sur les durées. Cette approche a une complexité trop élevée et nous avons donc étudié plusieurs algorithmes de clustering et nous avons retenu un qui semblait le plus performant. Cet algorithme a été implémenté et peut être utilisé sur les données de mobilité. Il reste à étudier la qualité du résultat en terme de représentativité.

Les données de consommation électrique : Nous avons étudié les données de consommation électrique (PECAN Street) pour développer un contrôleur adapté à une grille. Dans une première étape, il est nécessaire de désagréger les consommations de chaque appareil (on suppose qu'il n'y a pas de capteur associé à chaque appareil) à partir de la consommation globale. Nous avons utilisé divers échantillonneurs de Gibbs en choisissant des variantes dont la complexité de calcul va être faible. On construit alors une trace d'utilisation pour les appareils les plus consommateurs. L'idée du contrôleur est de pouvoir décaler les utilisations de ces appareils (charge voiture, production eau chaude, air conditionné) afin de modifier la courbe de demande.

Liste des publications communes

[1] « Convex Stochastic Bounds and Stochastic Optimisation on Graphs », J. Cohen, A. Fauquette, J.M. Fourneau, G.C. Noukela, N. Pekergin, Electronic Notes on Theoretical Computer Science, Special Issue for the selected papers of Practical Aspects of Stochastic Modeling, Berlin, 2017.

Liste des rapports de stage

[R1] G.C. Noukela, « Algorithmique pour les ordres stochastiques sur les distributions discrètes » rapport de stage de fin de Master 1 Informatique, Univ. Creteil, 2017. Co-auteur du papier [1].

[R2] A. Fauquette, rapport de stage Initiation à la recherche, Centrale/Supelec, 2017. Co-auteur du papier [1].

[R3] Stephan Kunne, rapport de stage de fin de Master 1 Informatique, Univ. Paris XI, 2017. Sur une nouvelle approche pour la sélection de trajectoires représentatives pour des études de mobilité.

[R4] Arnaud Cadas, «The power disaggregation algorithms and their applications to demand dispatch », Rapport de stage M2 Statistique, UPMC. 2017.

[X1] « The Case for Resource Sharing in Scientific Workflow Executions », XVI Simpósio de Sistemas Computacionais de Alto Desempenho (WSCAD 2015). Florianópolis/SC, Brazil, 2015, R. Oda, D. Cordeiro, R. Ferreira da Silva, E. Deelman,and K. R. Braghetto.

[X2] « Montage: a grid portal and software toolkit for science-grade astronomical image mosaicking » Int. J. Comput. Sci. Eng., 4(2):73–8, J. C. Jacob, D. S Katz, G. B Berriman, et al. (2009).

[X3] USC Epigenome Center. http://epigenome.usc.edu/.

[X4] « A Numerical Analysis of Dynamical Fault Trees based on Stochastic Bounds », IEEE QEST, 2015, J.M. Fourneau, et N. Pekergin

[X5] « Computing stochastic bounds of network distributions of time before failure », IEEE RNDM 2015, J.M. Fourneau

[X6] « An Algorithmic Approach to Stochastic Bounds », Performance Evaluation of Complex Systems: Techniques and Tools, Performance 2002, Tutorial Lectures, Springer LNCS 2459, pp 64-88, J.M. Fourneau, et N. Pekergin