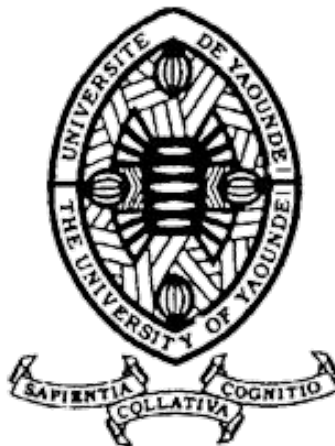


Parallel approaches to machine learning - A comprehensive survey

Fiche de lecture

Supervisé par :
Dr. MESSI



Faculté des Sciences

Nom :	DJIEMBOU TIENTCHEU VICTOR NICO
Article :	2
Publié le :	2012.11.16
Auteur :	Sujatha R. Upadhyaya
Mots clés :	Distributed and parallel ML, GPU, Map reduce
UE :	INF5099
Matricule :	17T2051
Département :	Informatique
Niveau - Option :	M2 - Sciences des Données

1 Contexte

La dernière décennie a été marquée par la proposition de nombreuses solutions scientifiques en particulier informatique avec l'avènement de l'apprentissage automatique pour la résolution de problèmes réels. Ses solutions au fil du temps font face à une augmentation de la calculabilité et suscitent l'intérêt de chercheurs pour pouvoir proposer des moyens de prendre en main l'explosion de la complexité de ses solutions temps sur l'aspect temporel que spatial.

Ainsi depuis la publication initiatrice de 1995 de K. Thearling [The96], la littérature a connu trois grandes tendances dans le but de solutionner ce problème. Une première qui visait à approcher de façon général le problème par **une exploration parallèle des données couplée à aux algorithmes d'apprentissage automatique** depuis 1995 ; En suite une deuxième qui démarre en 2000, **une exploration parallèle des données et algorithmes d'apprentissage automatique sur des GPUs** ; Enfin une dernière depuis 2005, **exploration parallèle des données et apprentissage automatique avec les techniques de map reduce**.

L'auteur va donc au moyen d'une revision systemic de la littérature proposer dans son article un grand plan sur ses différentes tendances.

2 Problème

En considérant un ensemble de catégories de problèmes ou encore un ensemble de modèles, le problème est de savoir quel est l'approche d'optimisation de la complexité sur les axes temporels et spatiales le plus recommandé selon la littérature.

3 motivations

Le domaine de l'optimisation de la complexité d'algorithmes d'apprentissage automatique est en vogue et il serait vraiment important de répondre à certaines questions comme :

- comment établir la relation entre un problème données et les approches de parallélisation ?
- comment faire le choix d'une approche ?
- quels sont les contraintes d'implémentation de ce choix ?

4 Approches parallèles générales de l'apprentissage automatique

Avant l'avènement des architectures GPUs, nombreuses approches assez généraliste de parallélisation ont existé et étaient principalement associées au contexte d'exploration de données en particulier l'extraction d'items fréquents et l'extraction de règles d'association. D'autres domaines co

Les avancées se sont principalement portées sur :

- la parallélisation des algorithmes de la grande famille **Apriori** en partitionnant les tâches sur les différents processeurs et en agrégeant le résultat progressivement (partitions basées sur le hachage).
- la parallélisation des algorithmes basés sur **Direct Hashing and Pruning (DHP)** qui vise à réussir à générer rapidement les ensembles d'items et réduire l'espace de stockage requis. Elle vient étendre les algorithmes parallèles d'Apriori en utilisant des techniques de hachage pour élaguer les arbres de recherche de hachage.
- **Exploration parallèle de données pour les règles d'association**, qui consiste en la génération en parallèle des ensembles de données candidats et en la détermination en parallèle des ensembles de données les plus fréquents au moyen des tables de hachages et des aggregations pour obtenir le comptage global.

- Algorithmes basés sur le DIC (Dynamic Itemset Counting), vient surmonter les limites des algorithmes DHP et Apriori qui nécessite plusieurs passage pour base de données stocké dans sur le processeur en réduisant le nombre de passages nécessaires pour générer l'ensemble d'éléments fréquents et le nombre d'éléments analysés par passage.
- L'extraction parallèle rapide (FPM), qui adopte des techniques d'élagage distribuées et globales.
- Algorithme de croissance parallèle des motifs fréquents, basé sur l'algorithme FPG qui extrait les motifs fréquents d'un arbre FP, sans génération d'ensembles d'éléments candidats.
- tentatives de parallélisation des tâches de l'apprentissage automatique tels que la validation croisée et aussi même de certains algorithmes tel que **SVM**, **Arbre de décision** et les **réseaux de neurones**.

5 Exploration de données, apprentissage automatique et efforts connexes sur les GPU

Les GPUs (Graphical Processing Unit) sont des unités de traitements conçus à des fin dans le domaine du traitement graphique. Mais sa capacité à facilement traiter les opération matricielles à suscité l'intérêt des chercheurs dans son usage dans la parallélisation des processus d'apprentissage automatique depuis 2005. A de nombreux expérimentations sur des algorithmes tels que K-means, SVM pour ne citer que ceux ci ont permit d'attester la haute performance de ses algorithmes sous des GPUs.

CUDA a été le langage qui est spécialement conçu pour la programmation sur les GPUs.

6 Exploration de données et apprentissage automatique à l'aide de la technique map reduce

La technique de map reduce a apporté un changement de paradigme dans la manière dont les données sont traitées dans les environnements distribués/parallèles. Ce changement à apporté certes de très grandes amélioration de performance mais n'a été exploré que dans très peu de domaines.

Ce cadre repose sur le principe que tous les algorithmes peuvent être exprimés sous forme de sommation. Les calculs sont effectués par différents mappers et la sommation finale est réalisée par le reducer.

Nombreuses autres applications ont été fait sur des algorithmes d'apprentissage automatique, tels que SVM; des fonctions telles les algorithmes de matching, multiplication matricielle, bien encore sur des GPU et des architectures multicoeurs.

7 Conclusion

L'amélioration des temps et d'espace de calcul depuis 1995 fait preuve d'extension d'approche et de nombreuse d'entre elles sont très intéressante et se doivent d'être exploré dans le contexte particulier de l'apprentissage automatique.

L'apprentissage automatique dans le cloud et le big data sont les perspectives intéressantes pour des approches tels que utilisation de la technique de map reduce et la parallélisation sur les GPUs.

Références

- [The96] Kurt Thearling, *Massively parallel architectures and algorithms for time series analysis*, Lectures in Complex Systems, Addison-Wesley (1996).
- [Upa13] Sujatha R Upadhyaya, *Parallel approaches to machine learning—a comprehensive survey*, Journal of Parallel and Distributed Computing **73** (2013), no. 3, 284–292.