Alma Mater Studiorum · Università di Bologna

Physics and Astronomy Department
PhD Thesis in Applied Physics

Implementation and optimization of algorithms in Biological Big Data Analytics

Supervisor:

Prof. Daniel Remondini

Correlator:

Prof. Gastone Castellani
Prof. Armando Bazzani

Presented by:

Nico Curti

Session 2019/2020

""*No one know nothing,*
everyone know something,
but something is nothing to someone,
while
something is important to everybody"

Daudi, Manyara

Abstract

Contents

Introduction

Chapter 1

Feature Selection - DNetPRO algorithm

Introduction to feature selection problem and theoretical background.
Focus on biological Big Data and problems related.

1.1 DNetPRO algorithm

Method description. Efficiency on a biological toy model.

1.2 Algorithm implementation

Description of the algorithm implementation in C++. Parallelization of the algorihtm. Use of BGL for network processing (filter node using view) Wrap in Python for Sklearn use Time Performances on different machines.

1.3 Synapse dataset

Description of the synapse datasets. Application of the DNetPRO on the Synapse dataset (mRNA, miRNA, RPPA) of Yuan et al. with two different pipelines. Discussion on obtained performances compared to the most common machine learning methods. Discussion on the ranking. Discussion on the extracted signature.

1.4 Cytokinoma dataset

Description of the cytokinoma dataset with statistics. Application of the DNetPRO on the Cytokine dataset. Discussion on the obtained signature and biological interpretation of the Alzheimer disease.

1.5 Bovine Paratuberculosis

Description of the bovine dataset with biological background. Application of the DNetPRO on the Bovine dataset with the description of the two singatures extracted. Discussion on biological interpretation of the genes.

Chapter 2

Deep Learning - Neural Network algorithms

Description of the modern deep neural networks. Computational problems and potential applications

2.1 Neural Network laboratory - NumPyNet

Description of the Neural Network laboratory developed in pure numpy. Study of the neural network functionality. Testing of the code against tensorflow.

2.2 Replicated Focusing Belief Propagation

Description of the rFBP library as optimization of the Julia code. Pure c++ implementation with Python wrap (sklearn compatibility). Scorer library as performance evaluation tool with parallel evaluation of scorers.

2.3 Build YouR Own Neural network - Byron library

Limits of the most common neural network frameworks. Neural Network library for parallel computing developed in C++. Pyron as python wrap of the library. Description of the algorithms used to optimize the computation (ex. im2col vs winograd).

2.4 Object Detection - Yolo architecture

Introduction on the image classification and detection with Yolo architecture. Implementation in Byron with description of performances against darknet (original implementation). Focus on performances (time, memory, cpu).

2.5 Super Resolution - WDSR architecture

Introduction on Super Resolution problem with focus on state-of-art neural network architecture. Description of the Byron implementation and application on NMR data with the most common measurements. Super-resolution allows better detection!

2.6 Image Segmentation - UNet architecture

Introduction on Image Segmentation problem. Creation of the datasets with common image-processing methods Application of Unet (Byron implementation) on femur images.

Chapter 3

Biological Big Data - CHIMeRA project

Many public datasets available. Description of the database used in chimera. Problems about the intersections and partial informations (single db).

3.1 Data extraction - Web scraping

Description of the web scraping techinques used to obtain the "no-public" datasets. Reference to the github project.

3.2 The CHIMeRA project

What is CHIMeRA project and which is its potentiality. Description of the database created and of the query implemented to obtain the results

3.3 CHIMeRA query

Some query examples like leukemia subnetwork and PRNP subnetwork. Description of the information extracted by these subnetworks.

Conclusions

Appendix A - Bioinformatic Pipeline Profiling

Bibliography

Acknowledgment