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

in questo lavoro si affronteranno diverse tematiche relative alla Big Data Analytics e si propongono soluzioni inerenti ad ognuna di esse con esempi sviluppati ed applicati a dati reali. Partendo dalla curse of dimensionality e la feature extraction (dnet), passando per la visualizzazione dei dati con le NN fino alla eterogeneità dei dati (chimera)

definire feature come variable e dire che nel resto del testo verranno usati in maniera indistinta i due termini

Chapter 1

Feature Selection - DNetPRO algorithm

After the end of the Human Genome Project (HGP, 2003) [11] there has been growing interest on biological data and their analysis. At the same time, the availability of this type of data increased exponentially with the technological improvement of data extractors (High-Throughput technologies) [13] and with lower production costs. Lower costs and efficiency in time extraction are the main factors that allow us to go into the new scientific era of Big Data. Biological Big Data works with very large and complex datasets which are typically impossible to store, handle and analyze using standard computer and techniques [9]. Just think that we need around 140 Gb for the storage of the DNA of a single person and an Array Express, a compendium of public gene expression data, contains more than 1.3 million of genomes which have been collected in more than 45000 experiments [4]. Since the number of available data is getting greater, we need to design several storage databases to organize, classify and moreover to extract informations from them. The Bioinformatics European Institute (EBI) at Hinxton (UK), which is part of the European Laboratory of Biological Molecular and one of the biggest repositories of biological data, stores 20 petabytes of data and genomics and proteomics back-ups. The amount of the genomics data is only 2 petabytes, and it doubles every year: it is not worth to remark that these quantities represent about a tenth of data stored by CERN of Ginevra [10]. On the other hand, the ability of processing data and the computational techniques of analysis do not grow the same way. Therefore the gap between the great growth of the number of available data and our ability to work with them is getting bigger.

From a computational point of view, the Bioinformatics new-science is looking for new methods to analyze these large amount of data. The common Machine Learning methods, i.e computational algorithms able to identify significant patterns into large quantities of data, needs to be optimized and modified to increase their computational and statistical performances. To optimize the computational times we need to extend existing methods and algorithms and to develop new dimensionality reduction techniques. In Machine Learning, in fact, as the dimensionality of the data increases, the amount of data required to perform a reliable analysis grows exponentially[1]. The dimensionality reduction techniques are methods able to identify the more significant variables of a given problem or a combination of them, where “significant” means that this smaller number of variables (or features) preserves the information about the problem as much as possible. So this huge amount of high-dimensional omics data (e.g. transcriptomics through microarray or NGS, epigenomics, SNP profiling, proteomics and metabolomics, but also metagenomics of gut microbiota) poses enormous challenges as how to extract useful information from them. One of the prominent problems is to extract low-dimensional sets of variables – sig-

[1] Often this phenomenon is called “curse of dimensionality”.

natures – for classification and diagnostic purposes, for example to better stratify patients for personalized intervention strategies based on their molecular profile [14, 2, 8, 1].

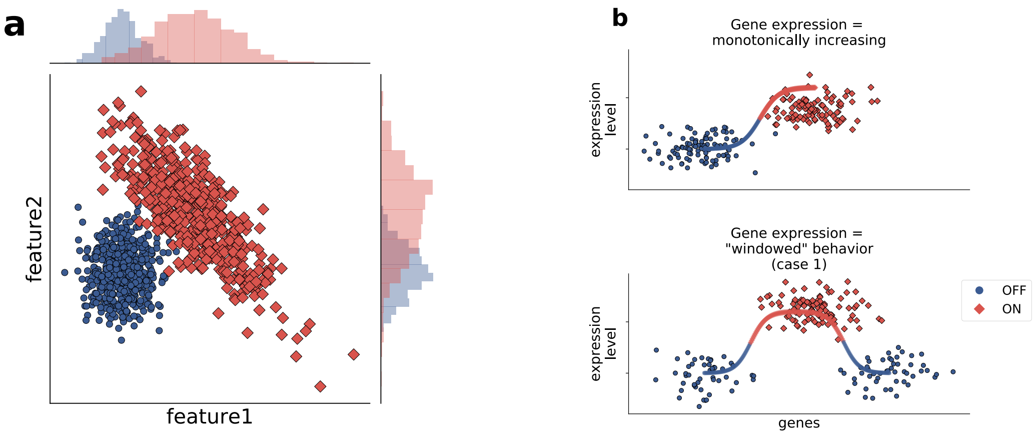

Figure 1.1: (**a**) An example in which single-parameter classification fails in predicting higher-dimension classification performance. Both parameters (*feature1* and *feature2*) badly classify in 1-D, but have a very good performance in 2D. Moreover, classification can be easily interpreted in terms of relative higher/lower expression of both probes. (**b**) Activity of a biological feature (e.g. a gene) as a function of its expression level: top) monotonically increasing, often also discretized to an on/off state; center, bottom) “windowed” behavior, in which there are two or more activity states that do not depend monotonically on expression level. X axis: expression level, Y axis, biological state (arbitrary scales).

Many approaches are used for these classification purposes [5], such as Elastic Net [7], Support Vector Machine, K-nearest Neighbor, Neural networks and Random Forest [12]. Some methods select signature variables by means of single-variable scoring methods [3, 6] (e.g. Student’s t test for a two-class comparison), while others search for projections in variable space, and then perform a dimensionality reduction by thresholding the projection weights, but these approaches could fail even in simple two-dimensional situations (Fig. 1.1).

Methods that select variables for multi-dimensional signatures based on single-variable performance can have limits in predicting higher-dimensional signature performance. As shown in Fig. 1.1(a), in which both variables taken singularly perform poorly, but their performance becomes optimal in a 2-dimensional combination, in terms of linear separation of the two classes.

It is known that complex separation surfaces characterize classification tasks associated to image and speech recognition, for which Deep Networks are used successfully in recent times, but in many cases biological data, such as gene or protein expression, are more likely characterized by a up/down-regulation behavior (as shown in Fig. 1.1(b) top), while more complex behaviors (e.g. a “windowed” optimal range of activity, Fig. 1.1(b) bottom) are much less likely. Thus, discriminant-based methods (and logistic regression methods alike) can very likely provide good classification performances in these cases (as demonstrated by our results with DNetPRO) if applied in at least two-dimensional spaces. Moreover, the “linearity” of these methods (that generate very simple class separation surfaces, i.e. linear or quadratic) guarantee that a “buildup” of a signature based on lower-dimensional signatures is feasible.

This consideration are relevant in particular for microarray data where we face on a small number of samples compared to a huge amount of variables (gene probes). This kind of problem, often called “large p, small n” problem (where p is the number of features, i.e variables, and n is the number of samples), tend to be prone to overfitting[2] and they are classified to ill-posed. The difficulty on the feature extraction can also increase due to noisy variables that can drastically affect the machine learning algorithms. Often is difficult to discriminate between noise and significant variables and even more as the number of variables rises.

In this thesis I propose a new method of features selection - DNetPRO, *Discriminant Analysis with Network PROcessing* - developed to outperform the mentioned above problems. The method is particularly designed to gene-expression data analysis and it was tested against the most common feature selection techniques. The method was already applied on gene-expression datasets but my work focused on the benchmark of it and on its optimization for Big Data applications. The pipeline algorithm is made by many different steps and only a part of it was designed to biological application: this allow me to apply (part of) the same techniques also in different kind of problems with good results (see next sections).

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.

[2] A solution to a problem is classified as "overfitted" if small fluctuations on the data variance produce classification errors.

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

[1] J. S. Beckmann and D. A. Lew. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. In *Genome Medicine*, 2016.

[2] I. S. Chan and G. S. Ginsburg. Personalized medicine: Progress and promise. *Annual Review of Genomics and Human Genetics*, 12(1):217–244, 2011. PMID: 21721939.

[3] L. Eckhard. A universal selection method in linear regression models. *Open Journal of Statistics*, 2, 2012.

[4] C. Greene, J. Tan, M. Ung, J. Moore, and C. Cheng. Big data bioinformatics. *Journal of cellular physiology*, 229(12), 2014.

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 2002.

[6] R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[7] J. J. Hughey and A. J. Butte. Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Research*, 2015.

[8] T. M. Johnson. Perspective on precision medicine in oncology. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 37(9):988–989, 2017.

[9] D. Kumari and R. Kumar. Impact of biological big data in bioinformatics. *International Journal of Computer Applications*, 101(11):22–24, 2014.

[10] V. Marx. The big challenges of big data. *Nature Reviews*, 498(255), 2013.

[11] M. McKinney. Human genome project information. *Reference Reviews*, 26(3):38–39, 2012.

[12] H. Pang, S. L. George, K. Hui, and T. Tong. Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE*, 2012.

[13] J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Molecular Cell*, 2015.

[14] K. Scotlandi, D. Remondini, G. Castellani, M. C. Manara, F. Nardi, L. Cantiani, M. Francesconi, M. Mercuri, A. M. Caccuri, M. Serra, S. Knuutila, and P. Picci. Overcoming resistance to conventional drugs in ewing sarcoma and identification of molecular predictors of outcome. *Journal of Clinical Oncology*, 27(13):2209–2216, 2009. PMID: 19307502.

Acknowledgment