Using game theory and similarity spaces in biological data; salvation or damnation?

Eleni Papakonstantinou[1,2], Thanasis Mitsis[2], Andrew Papadopoulos[3], Io Diakou[2,3], Panagiotis Panagopoulos[4,*]

[1] Division of Endocrinology and Metabolism, Center of Clinical, Experimental Surgery and Translational Research, Biomedical Research Foundation of the Academy of Athens, 11527 Athens, Greece

[2] Laboratory of Genetics, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, 11855 Athens, Greece

[3] Faculty of Natural & Mathematical Sciences, King's College London, Bush House, Strand, London, U.K.

[4] School of Informatics, National and Kapodistrian University of Athens, Athens, Greece

**Correspondence*: panagopoulos.png@gmail.com

Abstract

In bioinformatics, advances in full text indexing, time series indexing and graph indexing, combined with compressed storage, are now offering scalable and dynamic solutions. Mining and learning are constantly optimized by the use of parallelization and clustering methods such as dimensionality reduction approaches. However, there is no common perspective approach – and in this regard, recent research has suggested the use of similarity spaces. Such an approach allows for the efficient indexing, learning, and mining of different data types in common space, regardless of whether they are vectorial or not. It also benefits from parallelization approaches and GP-GPU computing or cloud infrastructure, supporting great scalability and allowing for efficient high-throughput analyses.

Introduction

Advances in full text indexing, time series indexing and graph indexing, combined with compressed storage, are now offering scalable and dynamic solutions. Mining and learning are constantly optimized by the use of parallelization and clustering methods such as dimensionality reduction approaches. However, there is no common perspective approach – and in this regard, recent research has suggested the use of similarity spaces. Such an approach allows for the efficient indexing, learning and mining of different data types in common space, regardless of whether they are vectorial or not. It also benefits

from parallelization approaches and GP-GPU computing or cloud infrastructure, supporting great scalability and allowing for efficient high-throughput analyses.

Learning and mining using game theory

The analysis of big data in terms of mining and learning is directly linked to parallelization. There already exists a variety of systems dealing with parallel classification and clustering. The literature includes ensemble-based parallelization, with parallel AdaBoost on multi-core processors (Chen et al., 2008) and tree ensemble learning with the PLANET system. There exists parallel learning in probabilistic graphical models (Panda et al., 2009), which are widely used in textual, image and event processing domains. Related to probabilistic models, recent research has managed to devise parallel Gibbs samplers, namely the Chromatic and Parallel Splash samplers. The Splash sampler allows linear scaling of performance in multicore settings. Related to classification, recent research brought decision trees to the world of parallelization, based on distributed histogram building. They show speedup that is significant (4.5x in a setting of 100,000 examples) with essentially no loss of accuracy.

GPU-processing has already been applied to learning problems. Learning classifier systems, applicable to classification and reinforcement learning, have been modeled for General Purpose GPU (GP-GPU) computing and have shown runtime speedup by 2x to 32x (Loiacono, 2011). In a life-span prediction task, given data of cancer patients, a parallel genetic algorithm using GPUs offered a speedup of about 7.5x on fitness function calculation, also managing to outperform - in terms of accuracy - other algorithms on the given dataset. Several research works have focused on optimization, Support Vector Machine (SVM) and kernel problems. Linear optimization over very large data with "trillions of features (the number of non-zero entries in the data matrix), billions of training examples and millions of parameters in an hour using a cluster of 1000 machines" was described, using a hybrid online-batch algorithm (Agarwal et al., 20011). The description also refers to the AllReduce communication infrastructure, as opposed to map-reduce, as another Hadoop-enabled alternative. Map-reduce-based classification has been used for SVM large scale spam filtering, with a speedup in training time of almost 4.5x, but a somewhat significant reduction in accuracy (about 2% loss). Since 2009 there exists a toolkit, CUSVM, that implements support vector classification and regression on CUDA, allowing GPU processing. The speedup reported by CUSVM is in the range of 12x to 147x over CPU-based algorithms. An approach on parallel linear support vector training using MPI/OpenMP is offering very competitive results over other algorithms on the PASCAL Large-scale learning challenge (Woodsend & Gondzio, 2010).

Other works, related to data mining and classification are focused on dimensionality reduction, e.g., map-reduce based LSI using k-means clustering and GPU-based manifold learning (Campana-Olivo, 2011). In the manifold learning case, there was an achieved

speedup of up to 26x, over the CPU case. In recent works, there is an efficient distributed implementation of spectral clustering applicable to graphs. There exist also open-source projects aiming at distributed machine learning - e.g., Apache Mahout- or GPU learning- e.g., GPUMLib. Furthermore, a number of data analysis programming models and execution frameworks, are also available, such as Nephele/PACT, GridBatch which run over the cloud, and Dremel which is an interactive analysis tool for large data. The domain of data analysis using map-reduce, which is a fully upcoming research area, is surveyed in a review paper by Lee et al., while a less focused and in-depth view of computational solutions for large scale data management and analysis in genetics can be found in a publication in Nature Reviews - Genetics journal (Lee et al., 2011; Schadt, 2010).

Similarity spaces

A plausible solution for the problems of both, apparently orthogonal, domains is the research on similarity spaces. Note that a similarity space is a space where at least a similarity function between every pair of instances in the space has been defined. For example, in the case of genomic sequences, this similarity can be related to the (possibly fuzzy) edit distance between pairs of sequences. Thus, the similarity space is the space of sequences, given this edit distance as similarity. On the other hand, in the case of protein study, the structural similarity is the direct candidate to form the similarity space of the problem. Thus, proteins will be described based on their respective structural similarity to other proteins.

The focus of the scientific community is turning into similarity spaces because they include any metric space (by trivial extension), allowing applicability to both vectorial and non-vectorial data. Non-vectorial data can be easily found, in addition to bioinformatics and sensor networks, in such domains as social network analysis (graphs and networks, multi-modal graphs), user modeling (multi-modal user models, context-aware modeling), image and video processing and other cases. This approach is highly supported by research on kernel and similarity functions. Recent works have showed that powerful classifiers, such as the Support Vector Machine and Kernel Perceptron, can rely on simple similarity functions to improve their performance. Adding to this finding the scalability offered by General Purpose Graphical Processing Unit (GP-GPU) computing and distributed or cloud infrastructures, one can aim to use the existing computation infrastructures for a high-efficiency calculation of similarities in similarity spaces. The outcome of this process will then feed parallel learning algorithms to support reactive, high-throughput analysis. Essentially, the measurement of similarities between pairs of instances is a highly parallelizable problem in itself, offering good scalability properties.

Overall, combining the power of similarity-based learning and mining with distributed (e.g., map-reduce paradigm) and highly parallel (e.g., GPU-based) processing over responsive storage mechanisms will allow large scale analysis on the fly. This will, in turn,

empower data-intensive research efforts and applications on big data to extract and infer knowledge. Such knowledge is the key to a competitive advantage in the era of prolific data generation and sharing.

Conclusions

A new approach, that could address the aforementioned issues under a common system, is the use of similarity spaces. This allows the high-throughput analysis of both vectorial and non-vectorial data, which has been a hindrance for both sequencing data and protein studies. Furthermore, such an approach can directly benefit from both modern parallelization research, as well as existing computation infrastructures and offers a highly scalable singular framework allowing next-generation data analysis.

References

Agarwal A et al. A reliable effective terascale linear learning system. arXiv:1110.4198 (2011).

Campana-Olivo et al.. Parallel implementation of nonlinear dimensionality reduction methods applied in object segmentation using CUDA in GPU. Proc. SPIE 8048, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVII, 80480R8048, doi:10.1117/12.884767 (2011).

Chen, Y. K., Li, W. & Tong, X. Parallelization of AdaBoost algorithm on multi-core processors. 2008 IEEE Workshop on Signal Processing Systems, doi:10.1109/SIPS.2008.4671775 (2008).

Lee, K. H., Lee, Y. J., Choi, H., Chung, Y. D. & Moon, B. Parallel data processing with MapReduce: A survey. SIGMOD Record40, 11–20, doi:10.1145/2094114.2094118 (2011).

Loiacono, D. Fast prediction computation in learning classifier systems using CUDA. (Association for Computing Machinery, 2011).

Panda, B., Herbach, J. S., Basu, S. & Bayardo, R. J. PLANET: Massively Parallel Learning of Tree Ensembles with MapReduce. Proc. of the 35th Int'l Conf. on Very Large Data Bases (2009).

Schadt E et al. Computational solutions to large-scale data management and analysis. Nat Rev Genet11, 647–657, doi:10.1038/nrg2857 (2010).

Woodsend, K. & Gondzio, J. Hybrid MPI/OpenMP parallel linear support vector machine training. J Mach Learn Res10 (2009).

Papakonstantinou E (2022) Using game theory and similarity spaces in biological data; salvation or damnation? *JStructBioinf* 1(1):15–18

Volume 1: Issue 1

pages 15–18