Hybrid Architecture with Deep Learning for Data Mining Applications

Thursday 5th April, 2018 (13:13)

João Loureiro
Informatics and Computing Engineering
Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n, 4200–465
Email: ei08101@fe.up.pt

Abstract

The health sciences are applied sciences used in the delivery of health care to human beings. These sciences are increasingly using the field of *Bioinformatics* [1] which is responsible for developing methods and applying tools of computation and analysis to extract knowledge of biological data. [2]

In the globalizing world, knowledge and information are seen as keys of power and a new trading good. [3] Information is widely spread and public, the main challenge to extract usable and needed knowledge is in the diversity, volume and complexity of the available data. Particularly in *Bioinformatics* and computational biology we face not only increased volume and a diversity of highly complex, multidimensional and often weakly-structured and noisy data, but also the growing need for integrative analysis and modeling. [4] [5] Some examples of the data that we are addressing include: X-ray; Mammography; Electrocardiography; Clinical Analysis; Patient's Medical History.

Machine Learning algorithms or Data Mining have some limitations on the analysis of a wide range of multidimensional data. Deep Learning for instance is well suited for the classification of images and numeric attributes. [6] Decision trees truly shine on conditional probabilities with categorical data. [7] However neither Deep Learning nor Decision trees can be used to a direct analysis of relational data.

The goal of our work is to test if an Hybrid Architecture [8] can be an advantage to process and extract knowledge from this diverse data taking into account the specificity's of a wide range of Machine Learning and Data Mining algorithms. This hybrid approach must be able to deal with a heterogeneous set of information picking the right algorithm for each subset of data and combine the results of each subset to output a good result.

We test the various Machine Learning algorithms, such as Random Forest, SVMs, Decision Trees, Neural Networks and others, against the existing subsets that can be images, tabular data, reports *et al* and do a performance measures of these algorithms to see which algorithms are better for each specific subset of data. After this phase we apply the previous knowledge and combine the multiple subsets results and verify how many times were we correct.

At the end we expect that our approach can produce better results in comparison to what already exists and be used to improve the health care system.

Keywords

Data Mining, Machine Learning, Hybrid Architecture, Bioinformatics, Health Sciences.

REFERENCES

[1] Ardeshir Bayat. "Science, medicine, and the future: Bioinformatics". In: *BMJ: British Medical Journal* 324.7344 (2002), p. 1018.

[2] Vasant Dhar. "Data science and prediction". In: *Communications of the ACM* 56.12 (2013), pp. 64–73.

[3] Tan Yigitcanlar. *Knowledge-Based Urban Development: Planning and Applications in the Information Era: Planning and Applications in the Information Era*. IGI Global, 2008.

[4] Andreas Holzinger. "Weakly structured data in health-informatics: The challenge for human-computer interaction". In: *Proceedings of INTERACT 2011 Workshop: Promoting and supporting healthy living by design.* . 2011.

[5] Mourad Ouzzani, Paolo Papotti, and Erhard Rahm. "Introduction to the special issue on data quality". In: *Information Systems* (2013).

[6] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), p. 115.

[7] S Rasoul Safavian and David Landgrebe. "A survey of decision tree classifier methodology". In: *IEEE transactions on systems, man, and cybernetics* 21.3 (1991), pp. 660–674.

[8] Dongyan Xu et al. "Analysis of a CDN–P2P hybrid architecture for cost-effective streaming media distribution". In: *Multimedia Systems* 11.4 (2006), pp. 383–399.