Towards Intelligent Clinical Decision Support Systems

Inês Dutra
Departamento de Ciência de Computadores
Faculdade de Ciências, Univerisdade do Porto
CRACS / INESC TEC LA

Host Research Unit: CRACS
External Researcher: Prof. Ashwin Srinivasan, IIIT-Delhi, India
Alternative External Researcher: Dr. Nuno A. Fonseca, EMBL, UK

Abstract

Inductive Logic Programming (ILP) is a machine learning technique that: (1) can handle multi-relational data, (2) can produce classifiers that are interpretable by specialists and (3) can help uncovering non-trivial knowledge that is also interpretable by specialists. One of the main drawbacks of ILP systems is their lack of efficiency in producing new knowledge when searching for good classifiers in a large space of hypotheses. Despite the various efforts towards making ILP systems efficient, it is still not possible to exploit a vast portion of the search space in order to find better classifiers. A notable exception is the very recent work by Srinivasan and Bain [12], which used a data stream technique to be able to process millions of data items. We argue that paralelization techniques can further improve data processing by ILP systems and would like to pursue this research path. Several works in the literature have shown that parallelization can help improving performance without loosing the classification quality, but to the best of our knowledge works in this area only deal with hundreds or thousands of data items. We would like to be able to process millions of data items in parallel in an ILP system. Our main application is in the context of the FCT project ABLe whose main objective is to build a system that can integrate radiologists' expertise with Inductive Logic Programming (ILP).

1 State-of-the-Art and Originality

Inductive Logic Programming (ILP) machine learning systems are arguably the most useful in the medical domain but they face the challenge of dealing with huge search spaces caused by the ever growing amount of data produced in

medical routines daily. Nevertheless, ILP systems have already been used to construct predictive models for data drawn from diverse domains, for over a decade. These include the sciences, engineering, language processing, environment monitoring, and software analysis.

Previous studies about the breast cancer condition have demonstrated that relevant clinical information can be automatically extracted from mammographic data. For example, there are indications that high mass density, an attribute that is usually not considered relevant by most doctors, plays an important role on malignancy of findings [6, 7, 14]. Another example [4] uncovers that seven out of 435 women had an incorrect diagnostic leading to a tumour becoming malignant in a two-year screening period. Yet other two works found a classifier to predict undecisive biopsies [5, 9].

Despite these successful histories, an ILP system's search space can grow very quickly in ILP applications, very often preventing the systems from finding the best possible set of rules. Several techniques have therefore been proposed to improve search efficiency. Such techniques include improving computation times at individual nodes [2, 10], better representations of the search [1], sampling the search space [11, 15], and parallelism [8, 3]. The latter can be obtained from very different alternative approaches, such as dividing the search tree, dividing the examples, or even through performing cross-validation in parallel [13]. Very recently, a stream-based ILP system has proven to work with millions of data items [12]. Research on parallelization of ILP systems only exhibit results for hundreds or thousands of data items. We would like to improve on that and be able to process millions of data items in parallel using an ILP system.

It is thus proposed to address two issues in this work; the first is the development and test of novel ILP parallel algorithms to be applied on several large sets of mammographic data with the purpose of extracting non-trivial clinical information. There is a performance concern regarding the ILP engine when extracting rules from a large dataset of past mammographic data. We aim at developing and testing different algorithms to achieve a comparatively fast ILP rule inference procedure; this process will require dealing with very large amounts of data, and so tackling database issues that will surely arise can be considered a specific goal of this work as well.

The second issue we propose to address - and that is orthogonal to the performance of the ILP system - is the integration of experts' advice with the previously inferred rules from the past mammographic data. One of the state-of-the-art techniques to improve efficiency of ILP algorithms is to somehow restrict the search space over which the system's resources must span [2]. Our work proposes to develop restrictions of the search space based on the validation of rules by experts; this methodology allows the ILP system to focus its search on parameters which are empirically more likely to be clinically pertinent both for diagnosis and treatment of the breast cancer condition.

References

[1] H. Blockeel, L. Dehaspe, B. Demoen, G. Janssens, J. Ramon, and H. Vandecasteele. Executing query packs in ILP. In J. Cussens and A. Frisch, editors, *Proceedings of the 10th International Conference on Inductive Logic Programming*, volume 1866 of *Lecture Notes in Artificial Intelligence*, pages 60–77. Springer-Verlag, 2000.

[2] H. Blockeel, B. Demoen, G. Janssens, H. Vandecasteele, and W. Van Laer. Two advanced transformations for improving the efficiency of an ILP system. In J. Cussens and A. Frisch, editors, *Proceedings of the Work-in-Progress Track at the 10th International Conference on Inductive Logic Programming*, pages 43–59, 2000.

[3] Rui Camacho, Ruy Ramos, and Nuno A. Fonseca. AND Parallelism for ILP: the APIS system. In *International Conference on Inductive Logic Programming*, 2013.

[4] J. Davis, E. S. Burnside, I. C. Dutra, D. Page, and V. Santos Costa. Knowledge discovery from structured mammography reports using inductive logic programming. In *American Medical Informatics Association 2005 Annual Symposium*, pages 86–100, 2005.

[5] Inês Dutra, H. Nassif, David C. Page, J. Shavlik, R. Strigel, Y. Wu, E. M. Elezaby, and Elizabeth S. Burnside. Integrating machine learning and physician knowledge to improve the accuracy of breast biopsy. In *American Medical Informatics Association 2011 Annual Symposium*, 2011.

[6] P. Ferreira, I. Dutra, N. A. Fonseca, R. Woods, and E. Burnside. Studying the relevance of breast imaging features. In *Proc. of the international Conference on Health Informatics (HealthInf)*, Jan 2011.

[7] Pedro Miguel Ferreira, Nuno A. Fonseca, Inês Dutra, Ryan Woods, and Elizabeth S. Burnside. Predicting malignancy from mammography findings and surgical biopsies. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2011)*. IEEE, IEEE, November 2011.

[8] J. Graham, D. Page, and A. Wild. Parallel inductive logic programming. In *Proceedings of the Systems, Man, and Cybernetics Conference*, 2000.

[9] Finn Kuusisto, Inês Dutra, Houssam Nassif, Yirong Wu, Molly Klein, Heather Neuman, Jude Shavlik, and Elizabeth Burnside. Using machine learning to identify benign cases with non-definitive biopsy. In *15th IEEE International Conference on e-Health Networking, Application & Services (HEALTHCOM 2013)*, Portugal, 2013. IEEExplore, IEEExplore.

[10] V. Santos Costa, A. Srinivasan, and R. Camacho. A note on two simple transformations for improving the efficiency of an ILP system. In J. Cussens and A. Frisch, editors, *Proceedings of the 10th International Conference on*

Inductive Logic Programming, volume 1866 of *Lecture Notes in Artificial Intelligence*, pages 225–242. Springer-Verlag, 2000.

[11] A. Srinivasan. A study of two sampling methods for analysing large datasets with ILP. *Data Mining and Knowledge Discovery*, 3(1):95–123, 1999.

[12] Ashwin Srinivasan and Michael Bain. Relational Models with Streaming ILP. In *International Conference on Inductive Logic Programming*, 2013.

[13] J. Struyf and H. Blockeel. Efficient cross-validation in ILP. In Céline Rouveirol and Michèle Sebag, editors, *Proceedings of the 11th International Conference on Inductive Logic Programming*, volume 2157 of *Lecture Notes in Artificial Intelligence*, pages 228–239. Springer-Verlag, September 2001.

[14] R. W. Woods, L. Oliphant, K. Shinki, C. D. Page, J. Shavlik, and E. S. Burnside. Validation of results from knowledge discovery techniques: mass density as a predictor of breast cancer. *J Digit Imaging*, 2009.

[15] F. Zelezny, A. Srinivasan, and D. Page. Lattice-search runtime distributions may be heavy-tailed. In *Proceedings of the 12th International Conference on Inductive Logic Programming*. Springer Verlag, July 2002.