Postdoc Subject LUE 2018
Hybrid Knowledge Discovery
Application to Biomedical Data (GeenAge Project)

Marie-Dominique Devignes and Amedeo Napoli
LORIA / Inria Nancy Grand Est
BP 239, 54506 Vandoeuvre les Nancy
{Marie-Dominique.Devignes;Amedeo.Napoli}@loria.fr

1 Context, positioning and objectives of the proposal

Human agents have the remarkable capability to learn a large variety of concepts, often with very few examples, whereas current state-of-the-art data mining algorithms require hundreds or thousands of data points and struggle with problems such as ambiguity, validity, overfitting etc. Another characteristic of human agents is the ability to acquire knowledge about the world and to draw subsequent inferences. Following this line, we are interested in investigating the combination of numeric and symbolic data mining and as well the use of domain knowledge for improving the performances and capabilities of different data mining approaches. Indeed, data mining methods can be either symbolic or numerical, and applying one or the other to a given dataset does not provide the same output. Combining numeric and symbolic data mining methods remains a tasks still poorly investigated [5, 2].

Accordingly, the objectives of this postdoc research work are to study how the Knowledge Discovery (KD) should be carried out, given the data at hand and a collection of data mining methods. In this way, we consider that the knowledge discovery process is iterative, interactive, supported by graphical tools, and dependent on various dimensions related to data, domain knowledge and the target task (problem-solving). We intend to study the potential and the characteristics of a such an hybrid KD process, and establish an operational and reusable methodology. Hybrid means that symbolic and numerical methods, as well as supervised and non supervised methods, can be combined for mining complex and possibly large data.

This study takes place within the GeenAge Research Project. The ambition of the GeenAge project[1] is to design new strategies of diagnosis and management of healthy and pathological ageing by targeting the functional consequences of the interplay between genes, epigenome and environment. Our teams are particularly involved in the Axis "New transcriptomics avenues in diagnosis and therapy of age-related diseases". In this axis, biologists are characterizing circulating non coding RNAs in the blood of patients from a longitudinal cohort, the Stanislas cohort. The used methodology allows to distinguish various forms of RNAs –trapped in exosomes, complexed with proteins– and to identify various nucleotide modifications that could affect their function. These specific features will be integrated with all other patient descriptors available from the cohort database, and with relevant domain knowledge such as the known functions and interactions of non coding RNAs, in view of identifying biological ageing signatures.

[1] GeenAge is one of the so-called "IMPACT Projects" funded by "Lorraine Université d'Excellence".

Thus, the mission of the post-doctoral fellow will be to design a framework for hybrid knowledge discovery and to implement the related algorithms in the context of the GeenAge project. Interactions with biologists will help to identify the potential domain knowledge to be reused. Visualization and data mining tools will be deployed in this hybrid and knowledge-based approach.

2 Related Work

A knowledge discovery problem requires most of the time an interdisciplinary collaboration, involving researchers in data science and researchers in other domains such as biology, chemistry, medicine... In [4], we experienced a preliminary hybrid framework for mining metabolomic data within the so-called Diapason project (Diapason for "Diet-Health Interaction Along Life-Predictive Biomarkers of Life Transition Outcome Linked to Retirement"). The objective of Diapason is to analyze data related to an elderly male over-weighted population whose members are selected w.r.t. biological criteria related to "metabolic syndrome". The latter is considered as a risk factor for biochemical and physiological abnormalities associated with the development of type 2 diabetes and cardiovascular diseases. Two main tasks are investigated: (i) identification of biomarkers predicting the evolution of the health status towards metabolic syndrome of an individual five years before its occurrence, (ii) discovery of links and correlations between nutritional habits, social and economical factors, metabolomic data, functional and clinical parameters.

In this experiment, metabolomic data have the usual characteristics of life science data: a rather small number of individuals (hundreds) and a high number of features (thousands). Numerical mining methods such as Support Vector Machine (SVM) and Random Forest [1] are combined with pattern mining [6] and Formal Concept Analysis (FCA [3]). Numerical mining methods with feature selection are well suited to the mining of metabolomic data, while symbolic methods can provide missing explanations and a basis for visualization and interpretation, establishing a bridge between data and domain knowledge. Actually, most of the elements supporting the view of hybrid knowledge discovery are present, i.e. data preparation, hybrid mining, visualization, interpretation and replay.

3 The Organization of the Research Work

The research work will be organized around three main tasks.

- Task 1. "Hybrid Knowledge Discovery": the principles of combinations of classifiers for mining complex data are defined, and the mining of biomedical data is considered as a main case study. The Knowledge Discovery process relies on interactions with the analyst and depends on domain knowledge, preferences, and seed patterns.
- Task 2. "An Integrated System for Hybrid Knowledge Discovery": the implementation of the associated combination of data mining algorithms. This includes also an integration of modules for visualization and management of domain knowledge and related constraints.
- Task 3. "Publications": The preparation and writing of papers complements the practical work on the research subject.

No particular risk is expected as data are available (biomedical data). Some preliminary

experiences produced encouraging results. One risk would be to design a system which is not generic and efficient enough for working with the diversity of data available nowadays.

The knowledge discovery process is expected to be hybrid, exploratory, and also knowledge-based. Actually, in such an integrated approach, we try to take into account the multiple dimensions of knowledge discovery, i.e. data, knowledge and problem solving. By contrast, researchers working with numerical data mining methods (including deep learning) do not sufficiently pay attention and integrate domain knowledge in their systems.

4 Details for Application

- **Project Teams:**
  Capsid and Orpailleur (LORIA/Inria Nancy Grand Est)

- **Supervisors:**
  Marie-Dominique Devignes (CR CNRS, `Marie-Dominique.Devignes@loria.fr` ; `https://members.loria.fr/MDDevignes/`)

  Amedeo Napoli (DR CNRS, `Amedeo.Napoli@loria.fr.` ; `https://members.loria.fr/ANapoli/`)

- **Keywords:**
  knowledge discovery, mining of complex data, pattern mining, numerical data mining, meta-mining, biomedical data.

- **Skills and profile of the candidate:**
  A PhD Thesis in Computer Science or in Applied Mathematics. Prior research work on knowledge discovery, data mining and or machine learning will be highly appreciated. Additional knowledge about biomedical data and processes will be welcome too.

- **Job location, terms and duration:**
  This two-year position is based in LORIA/Inria Nancy Grand Est Lab in Nancy (Capsid and Orpailleur teams).

  Applicants will be interviewed by a commission in October 2018.

  The duration of the postdoc is 24 months with a planned starting date in November 2018 (this date is flexible).

- **How to apply:**
  Applicants are requested to submit the following elements:

  - A full and pudated CV including the list of publications of the applicant.
  - A motivation letter related to the position.
  - Recommendation letters.
  - Academic transcripts (if possible).

  The Deadline for applications is September 16th, 2018.

  Applications are only accepted through email. All documents must be sent to `marie-dominique.devignes@loria.fr` and `amedeo.napoli@loria.fr`.

References

[1] Charu C. Aggarwal. *Data Mining – The Textbook*. Springer, 2015.

[2] Hendrik Blockeel. Data mining: From procedural to declarative approaches. *New Generation Comput.*, 33(2):115–135, 2015.

[3] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis*. Springer, Berlin, 1999.

[4] Dhouha Grissa, Blandine Comte, Estelle Pujos-Guillot, and Amedeo Napoli. A hybrid knowledge discovery approach for mining predictive biomarkers in metabolomic data. In *Proceedings of ECML-PKDD Conference (Vol. I)*, pages 572–587, 2016.

[5] P. Nguyen, Melanie Hilario, and Alexandros Kalousis. Using Meta-mining to Support Data Mining Workflow Planning and Optimization. *Journal of Artificial Intelligence Research*, 51:605–644, 2014.

[6] Jilles Vreeken and Nikolaj Tatti. Interesting patterns. In Charu C. Aggarwal and Jiawei Han, editors, *Frequent Pattern Mining*, pages 105–134. Springer, 2014.