The research work will be organized around three main tasks.

Task 1: "Hybrid Exploratory KDD" - the principles of combinations of classifiers for mining complex data are defined, and the mining of metabolomic data is considered as a main case study, with the study of discrimination vs. prediction characters. The exploratory aspects rely on interactions with symbolic data mining and machine learning methods. One general idea is to use structured prior knowledge in the form of knowledge graphs for example and to show how this improves performances of different mining approaches.

Following the tracks from meta-learning to meta-mining through Exploratory Data Analysis (EDA) and declarative data mining, there is a need for defining an operational framework for "Hybrid Exploratory Knowledge Discovery". 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, and combining both remains a task rather poorly solved.

The objective of this postdoc research work is to study the Knowledge Discovery in Databases (KDD) process at a strategic and tactic levels for understanding how KDD should be carried out, e.g. for big data and the mining methods at hand. For that, we consider that the KDD process is iterative, interactive, supported by graphical tools, and dependent on main dimensions related to data, domain knowledge and the target task (problem-solving).

We intend to study the potential and quality of a such hybrid exploratory KDD 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 mined complex and possibly large data.

Assignment

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, etc. In [Eriksson et al. ECML-PKDD 2016], 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 analyse data related to an elderly male overweight 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) performing an integrated study to discover the links and correlations between several parameters such as nutritional habits (nutrition is a major health factor), social and economical factors, metabolomic data, functional and clinical parameters which are potentially risk factors of metabolic syndrome development.

In this experiment, numerical mining methods such as Support Vector Machine (SVM) and Random Forest are combined with pattern mining and Formal Concept Analysis (FCA). Discovering new predictive biomarkers in metabolomic data requires to distinguish discriminant and predictive patterns. Discriminant markers should classify the population in homogeneous groups while predictive markers should predict the occurrence of the target disease. Discriminant patterns are more related to supervised classifiers while predictive patterns are more related to supervised classifiers leading to a hybrid exploratory KDD process. In such a reference application, all basic ingredients of hybrid exploratory KDD are present, i.e. data preparation, hybrid mining, visualization, interpretation and replay.

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 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. Moreover, a generic hybrid exploratory KDD process mining such complex data is useful for understanding phenomena involved in diabetes of type 2 (T2D) and cancers. Impacts on society are straightforward as this is related to a public health problem. Other applications of this research work can be planned in chemistry for mining molecular structures and in agronomy for understanding landscape data.

Main activities

The research work will be organized around three main tasks:

1. **Hybrid Exploratory KDD**: the principles of combinations of classifiers for mining complex data are defined, and the mining of metabolomic data is considered as a main case study, with the study of discrimination vs. prediction characters. The exploratory aspects rely on interactions with symbolic data mining and machine learning methods. One general idea is to use structured prior knowledge in the form of knowledge graphs for example and to show how this improves performances of different mining approaches.

2. **Following the tracks from meta-learning to meta-mining through Exploratory Data Analysis (EDA)** and declarative data mining, there is a need for defining an operational framework for "Hybrid Exploratory Knowledge Discovery". 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, and combining both remains a task rather poorly solved.

3. **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) performing an integrated study to discover the links and correlations between several parameters such as nutritional habits (nutrition is a major health factor), social and economical factors, metabolomic data, functional and clinical parameters which are potentially risk factors of metabolic syndrome development.

About Inria

Inria, the French National Institute for computer science and applied mathematics, promotes "scientific excellence for technology transfer and society". Graduates from the world’s top universities, 2200 employees rise to the challenges of digital sciences. With its open, agile model, Inria is able to explore original approaches with its partners in industry and academia and provide an efficient response to the multidisciplinary and application challenges of the digital transformation. Inria is the source of many innovations that add value and create jobs.

The keys to success

**Application deadline**

**June 6th, 2018 (Midnight Paris time)**

**How to apply**

Upload your file on jobs.inria.fr in a single pdf or zip file, and send it as well by email to amedeo.napoli@inria.fr. Your file should contain the following documents:

- CV including a description of your research activities (2 pages max) and a short description of what you consider to be your best contributions and why (1 page max) and 3 contributions max.
- The report(s) from your PhD external committee (if known) and the expected members of your PhD committee.
- The report(s) from your PhD external reviewers, if applicable.
- If you haven't defended yet, the list of expected members of your PhD committee (if known) and the expected date of defense (the defense, not the manuscript submission).

In addition, at least one recommendation letter from your PhD advisor should be sent directly by their author(s) to amedeo.napoli@inria.fr.

Applications are to be sent as soon as possible.

**Conditions for application**
the analyst and depend on domain knowledge, preferences, seed patterns, links with declarative data mining.

Task 2. "An Integrated Platform for Exploratory Hybrid KDD": the implementation of such an operational hybrid exploratory KDD system includes modules for visualization and management of preferences and constraints.

Task 3. "Publications": The preparation and writing of papers complements the practical work on the research subject.

No particular risk is expected as we have data at hand (biomedical and chemical data) and preliminary experiences produced substantial 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.

``Hybrid'' and ``Exploratory'' are two main characteristics of data analysis that did not receive much attention yet in knowledge discovery. Indeed, people working in Deep Learning are trying to integrate knowledge aspects in their systems. We rather work in the opposite direction and take into account data and domain dimensions for guiding the KDD process.

Skills

Keywords: knowledge discovery, mining of complex data, pattern mining, numerical methods, hybrid mining, meta-mining.

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

Benefits package

- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave

Remuneration

Salary: 2653€ gross/month

Defence Security:

This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST). Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy:

As part of its diversity policy, all Inria positions are accessible to people with disabilities.

Warning: you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.