2018-00723 - PhD Position / Data and Knowledge Representation and Processing / Graduate degree or equivalent

Level of qualifications required: Graduate degree or equivalent
Fonction: PhD Position

Context

Context and motivation:

This proposal aims to explore the process of Knowledge Discovery in Databases (KDD) at a foundational level for understanding how KDD should be carried out given the data and the mining methods at hand. Indeed, we are living in an information society that is being revolutionized by large amount of data available and constantly produced, and the vast amount of methods constantly developed to analyse and explore it. However, this raises two natural questions that haven't been properly addressed, namely, "how to choose pertinent information from the data to be explored" and "what methods to employ for its exploration". Both of these questions can be regarded as decision problems that depend on the nature of data (symbolic, qualitative or numerical), and on the purpose of the study (descriptive or predictive).

Recent developments in KDD have shown that numerical methodologies such as those based on neural networks (NN) provide good predictive results on supervised numerical environments but they perform rather poorly on unsupervised symbolic environments. Likewise, symbolic methodologies such as subgroup discovery (SD) can operate in both supervised and unsupervised environments and provide important descriptive information, however their predictive outcomes remain rather modest when compared to numerical approaches, with a potentially high computational cost. This entails the two main goals of this thesis proposal, namely: (i) to bridge gaps between symbolic and numerical methodologies, and (ii) to develop decision making framework to integrate the best aspects of each world.

This doctoral project is motivated by two research projects about biomedical data, which focused on the tasks of analyzing patient data related to cancer and to metabolomic research. In this context, one of basic problems addressed is that of identifying homogeneous groups (bio-profiles) of patients with certain diseases (types of cancer/diabetes) and thus contribute to better diagnosis and to efficient treatments by identifying the key drivers of the disease. However, metabolomic data are difficult to study as they are obtained through complex physical processes and devices, and comprising a rather small number of instances (hundreds of individuals) described by a high number of features (thousands) of several types (numerical, ordinal, categorical). This asks for a decision making framework for hybrid KD that integrates numerical approaches for mining tasks with predictive purposes, as well as symbolic approaches for descriptive purposes and thus enhance visualization and user/practitioner interpretation. Such a hybrid framework should also provide explanations that contribute to the understanding of the various biomedical interrelated phenomena involved in such diseases.

Main objectives:

This doctoral work proposal originates from the above mentioned questions in the context of biomedical data, but the hybrid KDD framework that we seek goes beyond this case study, and should provide a reusable methodology that will impact other data sciences. Firstly, we aim to study KDD systems from a strategic and a qualitative point of view and to understand how the knowledge discovery process should be carried out given the data and the mining methods available. There is usually a variety of algorithms to choose from and some criteria to guide our choice. However, there is no clear strategy for selecting and combining them that takes into account the relationship between datasets and methods at work. This strategic information should be extracted from data, analysed and evaluated, used for descriptive purposes and reused to guide the decision making process in the strategic combination of mining methods.

The main tasks of the successful candidate are thus (i) to formalize the key components of the hybrid KD framework...
and (ii) to design an operational work flow. The candidate will first get acquainted with aggregation and decision making tools for the strategic combination of base classifiers. The underlying motivation for this is that not every classifier can be an expert in classifying all unknown samples, although each base classifier is an expert in a different local region of the feature space. Following the tracks of meta-learning and dynamic classifier selection, the second task of the candidate is to design an operational decision making framework for combining different classifiers based on their performances and their areas of expertise. In the first stages, we will consider simple aggregation models for combining the selected classifiers such as voting or weighted aggregation of classifier outputs. In later stages of the doctoral thesis, the doctoral student should propose more evolved decision and aggregation models for combining the selected classifiers that should also be learned during the KDD process.

**Assignment**

The ideal candidate is a graduated student of computer science and/or applied mathematics, familiar with knowledge discovery techniques (preferably, both symbolic and numerical) and/or decision making tools.

He/she should be acquainted with programming languages, preferably Python, as part of the work entails the implementation of the frameworks developed in course of this doctoral work.

**Main activities**

Main activities (5 maximum):

- Propose theoretical solutions for bridging the gap between numerical and symbolical computing
- Develop programs/applications/interfaces
- Write and publish research papers
- Write reports

**Skills**

The ideal candidate is a graduated student of computer science and/or applied mathematics, be familiar with knowledge discovery techniques (preferably, both symbolic and numerical) and/or decision making tools.

He/she should be acquainted with programming languages, preferably Python, as part of the work entails the implementation of the frameworks developed in course of this doctoral work.

**Benefits package**

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

**Remuneration**

Gross Salary per month: 1982€ brut per month (year 1 & 2) and 2085€ brut/month (year 3)

**General Information**

- **Theme/Domain**: Data and Knowledge Representation and Processing
- **Statistics (Big data) (BAP E)**
- **Town/city**: Villers-lès-Nancy
- **Inria Center**: CRI Nancy - Grand Est
- **Starting date**: 2018-10-01
- **Duration of contract**: 3 years
- **Deadline to apply**: 2018-05-01

**Contacts**

- **Inria Team**: ORPAILLEUR
The keys to success

Application deadline

May 1st, 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 chedy.raissi@inria.fr and miguel.couceiro@loria.fr. Your file should contain the following documents:

- Your CV.
- A cover/motivation letter describing your interest in this topic.
- A short (max one page) description of your Master thesis (or equivalent) or of the work in progress if not yet completed.
- Your degree certificates and transcripts for Bachelor and Master (or the last 5 years).
- Master thesis (or equivalent) if it is already completed and publications if any (it is not expected that you have any). Only the web links to these documents are preferable, if possible.

In addition, one recommendation letter from the person who supervises(d) your Master thesis (or research project or internship) should be sent directly by his/her author to chedy.raissi@inria.fr and miguel.couceiro@loria.fr.

Applications are to be sent as soon as possible.

Conditions for application

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.