Interactive structured pattern mining

December 5, 2020

Paid 5 or 6-month master's thesis, GREYC, CoDaG Team, Université de Caen Normandie, France

Contact: Albrecht Zimmermann (albrecht.zimmermann@unicaen.fr)

Supervisors: Albrecht Zimmermann, Bertrand Cuissart, Abdelkader Ouali

Start date: February/March 2021

Salary: the master's student will be paid according to rules governing French student internships (560 euro/month)

1 Context

This master's thesis will be performed in the context of the project InvolvD, financially supported by a grant of the french national research agency (ANR).

Pattern mining is the task of finding regularities or unexpected patterns in large databases. Structured pattern mining performs this task on structured data like sequences, trees, or, of particular importance for InvolvD, graphs. Until recently, pattern mining correspond to an iteration of the following pipeline: the data user specifies and parametrizes constraints, then he explores a large set of resulting patterns and adjusts the constraints, and restarts the process.

Recently, different researchers have proposed involving user feedback to more directly shape mining constraints. The feedback typically consists of rejecting or accepting individual patterns, or of ranking a small set of patterns. In addition to the pattern language itself, patterns have a second representation in this setting, which characterizes them w.r.t. observed statistics, covered instances etc. The latter representation is used together with the user feedback to learn a preference function, e.g. via an SVM or a regression learner. Optimizing this preference function then guides the mining process towards areas in the search space that are expected to contain interesting patterns to the user, and away from those containing uninteresting patterns.

Existing work on interactive pattern mining is mainly limited to unstructured patterns, i.e. itemsets, which can be both more easily distinguished, and for which ad-hoc pattern representations can be constructed without much effort. As an example, an itemset $\{i_1, i_2, i_3, i_4\}$ differs from $\{i_1, i_2, i_3, i_5\}$ by the *presence* of i_4/i_5 and a pattern representation that involves presence indicators can be used to learn this. A sequence $i_1 \rightarrow i_2 \rightarrow i_3 \rightarrow i_4$ differs from

$i_2 \rightarrow i_4 \rightarrow i_3 \rightarrow i_1$ due to the *order* of the elements and therefore requires a representation that can efficiently capture this to learn a preference function over sequential patterns. Yet adding all possible pairwise orders (let alone higher-order ones) will quickly make the representation unwieldy, and randomly picking only some risks losing valuable information.

2 Research topic

This master's thesis aims specifically at:

- Producing a review of the state-of-the-art on interactive structured pattern mining
- Proposing a static representation that would allow using methods developed for unstructured patterns
- Proposing and testing dynamic representations that are constructed during the feedback process, cutting down on the size of the initial (and overall) presentation

The funding for InvolvD includes funding for a three-year PhD thesis, which will further explore those topics, and for which the master student would therefore be well positioned.

3 Application

Required skills:

- Experience in machine learning, data mining, computer programming or applied mathematics is highly appreciated.
- French and/or English are the working languages.

Candidates are encouraged to contact us as soon as possible. Begin of the thesis is expected in February or March 2021. The complete application consists of the documents below, which should be sent as a single PDF file to Albrecht Zimmermann (albrecht.zimmermann@unicaen.fr), Bertrand Cuissart (bertrand.cuissart@unicaen) and Abdelkader Ouali (abdelkader.ouali@unicaen.fr).

- CV
- One-page cover letter (clearly indicating available start date as well as relevant qualifications, experience and motivation)
- University certificates and transcripts (both B.Sc and M.Sc degrees marks)
- Contact details of up to three referees
- Possibly an English language certificate and a list of publications
- **Attention**: all documents should be in English or in French.