MIPS: A graph mining library

Thomas Fannes* Ashraf Kibriya* Kurt De Grave* Jan Ramon*

Many practical datasets (e.g., biological, social, economic, ... networks) can be elegantly represented with graphs. In the MiGraNT project[1] we aim to develop a sound theoretical understanding of mining and learning with graphs. The MIgrant Prototype System `MIPS` is a library of effective algorithms, based on this theory. This is an ongoing project, which aims to integrate a larger number of results. Here, we present the basic system and a first set of algorithms.

Principles and basic system. `MIPS` is written in `C++` and strongly benefits from the meticulous use of `C++` templates, which allows to unite flexibility with efficiency. The library utilizes the `C++` boost library, especially the Boost Graph Library to represent flexibly graphs of different types ((un)directed, (un)labeled, ...) with the same code. The documentation is doxygen-based.

Frequent pattern mining. Mining frequent patterns is a data mining task often used in machine learning for feature generation. Depending on the application, homomorphism or subgraph isomorphism is the matching operator of preference, even though the latter one is more popular. For even a simple path, subgraph isomorphism is NP-complete, and classical mining algorithms become intractable for patterns of a very modest size—a lot of work studies frequency counting of patterns between 3 and 5 nodes. We use recent advances on fixed parameter tractability to construct (randomized) algorithms capable of deciding subgraph isomorphism of a pattern in a network in $O(k^2 \log^2(k) m^w 2^k)$, with m the number of network edges, k the number of pattern vertices and w the pattern treewidth. See [1] for (a part of) the relevant theory. Our algorithm can mine frequent trees up to size 17–18, and is, to the best of our knowledge, the first tractable tree pattern miner under subgraph isomorphism for large, dense networks. Currently, we are empirically studying the behavior for non-tree graphs.

Supervised learning. Although many libraries contain decision tree and random forest learning algorithms, `MIPS` includes a new implementation where the novelty lies in the aforementioned graph-based approach and exploitation of the templating mechanism. `MIPS` can efficiently learn from training data that does not fit in memory. These capabilities were successfully applied to the field of proteomics [2].

Future development. We are adding further components to the system, amongst which algorithms to estimate the effective sample size of a set of networked (and hence non-independent) examples, kernel regression and decision tree learners for dependent examples, algorithms to learn dynamic models for time-evolving graphs, self-compiling graph algorithms, and algorithms to let the previous parts work on graph databases (not fitting in memory). We are also improving documentation and the integration of the several components.

License. `MIPS` is GPLv3 licensed and available at `https://dtai.cs.kuleuven.be/software/mips`[2].

References

[1] A. Kibriya and J. Ramon. Nearly exact mining of frequent trees in large networks. *Data Mining and Knowledge Discovery*, 27(3):478–504, November 2013.

[2] T. Fannes, E. Vandemarliere, L. Schietgat et al. Predicting tryptic cleavage from proteomics data using decision tree ensembles. *Journal of Proteome Research*, 12(5):2253–2259, April 2013.

*Department of Computer Science, KU Leuven, Belgium
{firstname}.{lastname}@cs.kuleuven.be

[1] This research was supported by ERC Starting Grant 240186 'MiGraNT: Mining Graphs and Networks: a Theory-based approach'.

[2] It is also available at `http://mloss.org/software/view/608/`