FAIR multi-label classification

Dragi Kocev, Jasmin Bogatinovski, Ana Kostovska, Panče Panov

ECML PKDD 2021 Tutorial
17 September 2021

Presenting...

- Dragi Kocev, Jožef Stefan Institute, Ljubljana, Slovenia
- Jasmin Bogatinovski, Technische Universität Berlin, Germany
- Ana Kostovska, Jožef Stefan Institute, Ljubljana, Slovenia
- Panče Panov, Jožef Stefan Institute, Ljubljana, Slovenia

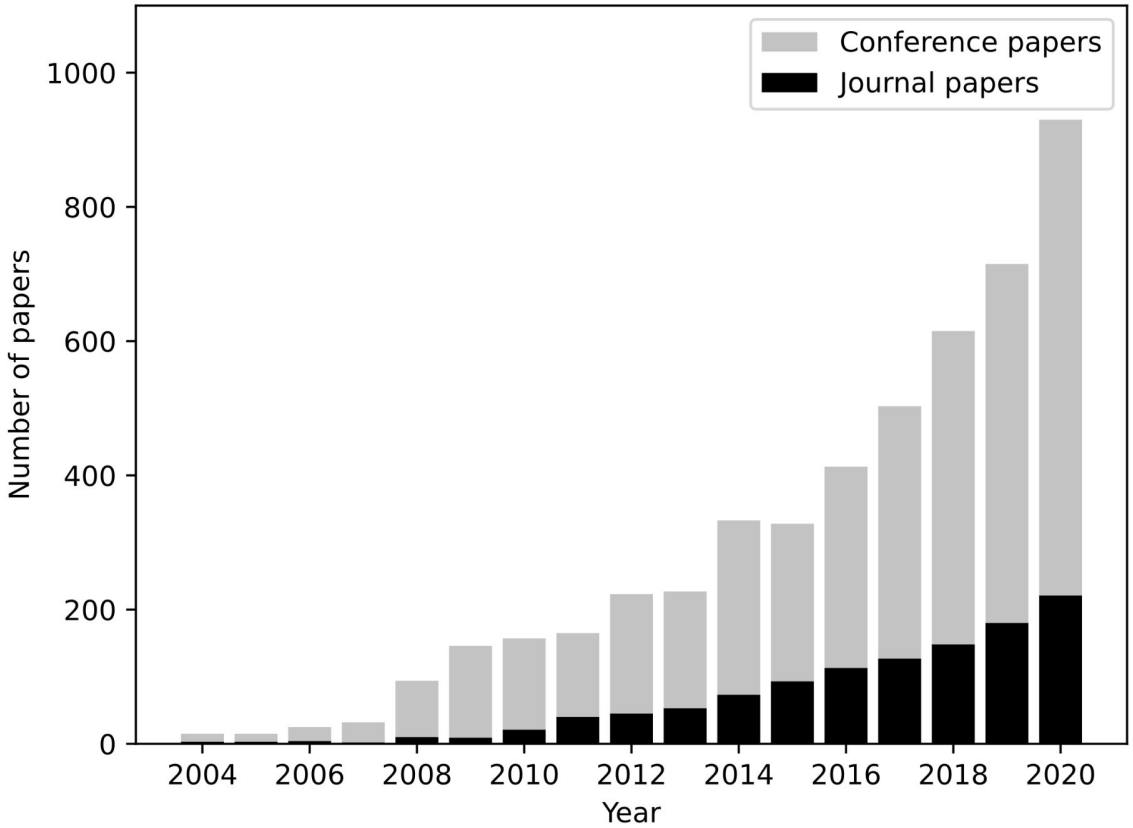
Multi-label classification

Multi-label classification (MLC) is a machine learning task where the goal is to label an example with multiple labels simultaneously

MLC offers a positive predictive edge on many practical problems and it is widely used in a variety of application domains:

- **life sciences and medicine** (e.g., gene and protein function prediction, drug modes of action, drug repurposing, patient status prediction),
- **environmental sciences** (e.g., habitat models, compounds' toxicity),
- **text classification** (e.g., tag suggestion, news articles, web pages, patents, e-mails, bookmarks, recipes),
- **semantic annotation of images and videos** (e.g., image annotation, news clips, movies, species identification from recordings, music evoking emotions),
- **marketing** and others.

Increase of the number of MLC papers in SCOPUS!



There is a need for FAIR MLC!

- The exponential explosion of MLC papers requires
  - proper **benchmarking**,
  - **reusability** of previous results and
  - better **understanding** of the proposed novel methods and the problems addressed with them.
- The papers introducing novel MLC methods typically
  - evaluate the performance of the models on a (often arbitrary) selected subset of data,
  - against some (often arbitrary) selected competing methods,
  - and using some (often arbitrary) evaluation measures (there are more than a dozen of measures to choose from).

The roadmap to FAIR MLC

Ensuring a proper, correct, robust, trustworthy benchmarking is of utmost importance for the further development of the field.

We believe that this can be achieved through

1. Managing all digital artifacts (e.g., datasets, methods, performance data and their meta descriptors) according to the cutting edge standards of data management such as the FAIR (Findable, Accessible, Interoperable, and Reusable) principles.

2. Landscaping of the space of data and methods for MLC

FAIR MLC: Tutorial outline

- 09.15-10.00: Knowledge and data representation, and FAIR principles (Panče Panov)
- 10.00-10.45: FAIR MLC data representation and repositories (Ana Kostovska)
- 10.45-11.15: Coffee break
- 11.15-12.00: Comprehensive empirical study of MLC methods (Jasmin Bogatinovski)
- 12.00-12.45: Landscaping MLC through meta learning on the empirical results (Dragi Kocev)
- 12.45-13.00: Summary remarks outlining the open challenges in MLC

All materials will be made available at

http://mlc.ijs.si/fair-mlc-ecml-2021/