WIREs Wiley Interdisciplinary Reviews

In silico toxicology: comprehensive benchmarking of multi-label classification methods applied to chemical toxicity data

Arwa Bin Raies, Vladimir B. Bajic*
King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Centre (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), Thuwal, Saudi Arabia

Supplementary Information

Table of Contents

Methods implementation and software

Most of the methods were implemented in Python 2.7. We adapted the code from reference[1] for the binary relevance and classifier chains approaches, and modified the code from the scikit-multilearn library (http://scikit.ml/) for multi-label *K* nearest neighbours method. The deep learning and multi-label Boolean matrix decomposition methods were applied using Meka[2]. We used scikit-multilearn python interface to Meka. The base classifiers were implemented using Scikit-learn library[3] in Python 2.7. The code for analysis and generation of figures was implemented in Python 3.4. All relevant codes for training and testing the models, analysing the results and creating the figures, and instructions to run the code are available online at www.cbrc.kaust.edu.sa/mlc/index.php.

Dataset sources

The dataset consists of *in vivo* toxicity data of pharmaceutical, environmental and industrial compounds that we compiled from public toxicity databases.

- Carcinogenicity data was gathered from National Toxicology Program Dataset (http://www.predictive-toxicology.org/data/ntp), Carcinogenic Potency Database (https://toxnet.nlm.nih.gov/cpdb/), Food and Drug Administration Carcinogenicity Studies with Rats and Mice (http://www.predictive-toxicology.org/data/fda), and ToxCast Data (https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data).
- Toxicity of environmental compounds was acquired from Cal/Ecotox Database (http://oehha.ca.gov/ecotoxicology/general-info/calecotox-database), Ecological Soil Screening Database (http://www.epa.gov/ecotox/ecossl/), and Ecotoxicology Knowledgebase (http://cfpub.epa.gov/ecotox/).
- Developmental, maternal and reproductive toxicity data was obtained from ToxCast, and ILSL Developmental Toxicity Database (http://www.ilsi.org/ResearchFoundation/RSIA/Pages/DevelopmentlToxicityDatabase.aspx).

Additionally, we used Aggregated Computational Toxicology Resource Database (https://actor.epa.gov/actor/home.xhtml), OECD QSAR Toolbox (http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm), and Echemportal (http://www.echemportal.org/echemportal/) to acquire more data about carcinogenicity, developmental toxicity, eye irritation, genotoxicity, maternal toxicity, reproductive toxicity, and skin sensitization and irritation.

References

1. Montanari F, Zdrazil B, Digles D, Ecker GF. Selectivity profiling of BCRP versus P-gp inhibition: from automated collection of polypharmacology data to multi-label learning. *J Cheminform* 2016, 8:7.
2. Read J, Reutemann P, Pfahringer B, Holmes G. MEKA: A multi-label/multi-target extension to WEKA. *J Mach Learn Res* 2016, 17:1-5.
3. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011, 12.