dislib Documentation

Release 0.7.0

Barcelona Supercomputing Center

Nov 11, 2022

Contents:

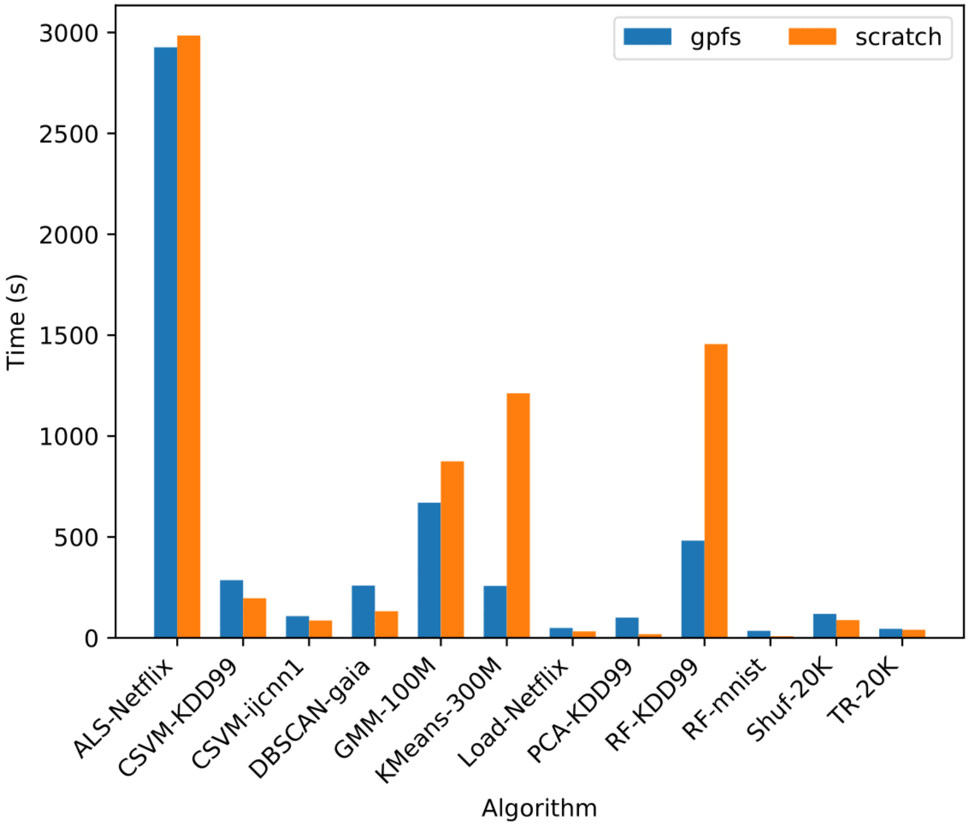
The Distributed Computing Library (dislib) provides distributed algorithms ready to use as a library. So far, dislib is highly focused on machine learning algorithms, and is greatly inspired by scikit-learn. However, other types of numerical algorithms might be added in the future. The main objective of dislib is to facilitate the execution of big data analytics algorithms in distributed platforms, such as clusters, clouds, and supercomputers.

Dislib has been implemented on top of PyCOMPSs programming model, and it is being developed by the Workflows and Distributed Computing group of the Barcelona Supercomputing Center.

CHAPTER 1

Performance

The following plot shows fit time of some dislib models on the MareNostrum 4 supercomputer (using 8 worker nodes):



Labels on the horizontal axis represent algorithm-dataset, where:

- ALS = AlternatingLeastSquares

- CSVM = CascadeSVM
- GMM = GaussianMixture
- Load = `load_svmlight_file`
- RF = RandomForestClassifier
- Shuf = `shuffle`
- TR = `Array.transpose`

and:

- Netflix = The Netflix Prize dataset.
- KDD99 = The KDDCUP 1999 dataset.
- ijcnn1 = The ijcnn1 dataset.
- gaia = The Tycho-Gaia Astrometric Solution dataset[1].
- 100M and 300M = 100 and 300 million random samples, with 100 features each.
- mnist = The mnist dataset.
- 20K = Square matrix of 20 thousand rows and 20 thousand columns, with random values.

[1] Michalik, Daniel, Lindegren, Lennart, and Hobbs, David, "The Tycho-Gaia astrometric solution - How to get 2.5 million parallaxes with less than one year of Gaia data," A&A, vol. 574, p. A115, 2015.

CHAPTER 2

Source code

The source code of dislib is available online at Github.

CHAPTER 3

Support

If you have questions or issues about the dislib you can join us in Slack.

Alternatively, you can send us an e-mail to support-compss@bsc.es.

CHAPTER 4

Citing dislib

If you use dislib in a scientific publication, we would appreciate citations to the following paper:

J. Álvarez Cid-Fuentes, S. Solà, P. Álvarez, A. Castro-Ginard, and R. M. Badia, “dislib: Large Scale High Performance Machine Learning in Python,” in *Proceedings of the 15th International Conference on eScience*, 2019, pp. 96-105

4.1 Bibtex:

```
@inproceedings{dislib,
            title       = {{dislib: Large Scale High Performance Machine Learning in␣
↪Python}},
            author      = {Javier Álvarez Cid-Fuentes and Salvi Solà and Pol Álvarez␣
↪and Alfred Castro-Ginard and Rosa M. Badia},
            booktitle   = {Proceedings of the 15th International Conference on␣
↪eScience},
            pages       = {96-105},
            year        = {2019},
 }
```

4.1.1 Indices and tables

- genindex
- modindex
- search