Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Doc-Start

Subject Section

koios: machine learning for big biological data

Simon Dirmeier [1] and Niko Beerenwinkel [1,*]

[1] Department, Institution, City, Post Code, Switzerland

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Analysing biological data is becoming increasingly difficult due to its ever increasing volumnes and dimensionality. Since desktop solutions are not sufficient anymore to analyse big biological data sets comprising terabytes of data, such as features extracted from images or gene expression of millions of cells, we propose a novel machine learning command line tool called *koios* for a distributed highly-parallel analysis of biological data.
Availability and implementation: *koios* is available from Github (https://github.com/cbg-ethz/koios) and soon to be released on Bioconda.
Contact: niko.beerenwinkel@bsse.ethz.ch

1 Introduction

The not-so-recent advent of high-dimensional data is still posing not only statistical, but primarily methodological problems for researchers across fields such as computational biology, astronomy or the social sciences. (**?**). Recently, approaches such as (**?**) have been introduced to tackle analysing high dimensional data sets. However, scanpy only scales up to a (few) million observations rendering it unsuitable for analysis of, e.g., microscopy imaging data that often has images of billions of cells. Approaches that also rely on Apache Spark, such as, (**?**) lack ease of use and a capability of effortless data analysos.
To summarize, the contributions of this paper are:

- A command line tool for analysis of billions of observations.
- Clustering and regression using MLLib from Apache Spark.
- A proper landing page.
- Customizable and automatic builds for single analyses

2 Methods

koios is a Python package callable as executable from the command line that implements various machine learning algorithms for automated data analysis. When using koios the user merely needs to specify input and ouput data and the algorithms to be used for analysis as config file. Running koios will automatically work through the different algorithms. On the and ouput data. Snakemake itself calls Apache's Spark which distributes the data on several workers and runs the respective models.

koios comes with a variety of already implemented algorithms (Figure **??**):

- Dimensionality reduction: PCA, kPCA, factor analysis,
- Supervised-learning: regression and random forests,
- Unsupervised-learning: Gaussian mixture models and k-means clustering.

.

2.1 Regression example

Table **??**

2.2 Clustering example

For instance, when working with single-cell features extracted from microscopy images, one usually wants to first eliminate highly-correlated features, for instance, by embedding into a lower uncorrelated space using PCA, kPCA or factor analysis. Following up clustering

TODO mention not on desktop computer -> needs drivers and so.

3 Conclusion

koios is a commandline tool for machine learning for big biological data

Acknowledgements

Funding

This work has been supported by the ... Text Text Text Text.

References