JOSS
The Journal of Open Source Software

q2-sample-classifier: machine-learning tools for microbiome classification and regression

Nicholas A Bokulich[1], **Matthew R Dillon**[1], **Evan Bolyen**[1], **Benjamin D Kaehler**[2], **Gavin A Huttley**[2], **and J Gregory Caporaso**[1, 3]

1 The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA **2** Research School of Biology, Australian National University, Canberra, Australia **3** Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA

DOI: 10.21105/joss.00934

Software

- Review
- Repository
- Archive

Submitted: 10 August 2018
Published: 23 October 2018

Summary

q2-sample-classifier is a plugin for the QIIME 2 microbiome bioinformatics platform that facilitates access, reproducibility, and interpretation of supervised learning (SL) methods for a broad audience of non-bioinformatics specialists.

Microbiome studies often aim to predict outcomes or differentiate samples based on their microbial compositions, tasks that can be efficiently performed by SL methods (Knights et al., 2011). The goal of SL is to train a machine learning model on a set of samples with known target values/class labels, and then use that model to predict the target values/class membership of additional, unlabeled samples. The ability to categorize new samples, as opposed to describing the structure of existing data, extends itself to many useful applications, e.g., the prediction of disease/susceptibility (Pasolli, Truong, Malik, Waldron, & Segata, 2016; Schubert, Sinani, & Schloss, 2015; Yazdani et al., 2016), crop productivity (Chang, Haudenshield, Bowen, & Hartman, 2017), wine chemical composition (Bokulich et al., 2016b), or sample collection site (Bokulich, Thorngate, Richardson, & Mills, 2013); the identification of mislabeled samples in microbiome data sets (Knights et al., 2011); or tracking microbiota-for-age development in children (Bokulich et al., 2016a; Subramanian et al., 2014).

We describe q2-sample-classifier, a QIIME 2 plugin to support SL tools for pattern recognition in microbiome data. This plugin provides several SL methods, automatic parameter tuning, feature selection, and various learning algorithms. The visualizations generated provide portable, shareable reports, publication-ready figures, and integrated decentralized data provenance. Additionally, integration as a QIIME 2 plugin streamlines data handling and supports the use of multiple user interfaces, including a prototype graphical user interface (q2studio), facilitating its use for non-expert users. The plugin is freely available under the BSD-3-Clause license at https://github.com/qiime2/q2-sample-classifier.

The q2-sample-classifier plugin is written in Python 3.5 and employs pandas (McKinney, 2010) and numpy (Walt, Colbert, & Varoquaux, 2011) for data manipulation, scikit-learn (Pedregosa et al., 2011) for SL and feature selection algorithms, scipy (Jones, Oliphant, Peterson, & others, 2001) for statistical testing, and matplotlib (Hunter, 2007) and seaborn (Waskom et al., 2017) for data visualization. The plugin is compatible with macOS and Linux operating systems.

The standard workflow for classification and regression in q2-feature-classifier is shown in Figure 1. All q2-sample-classifier actions accept a feature table (i.e., matrix of feature counts per sample) and sample metadata (prediction targets) as input. Feature observations for q2-sample-classifier would commonly consist of microbial counts (e.g., amplicon

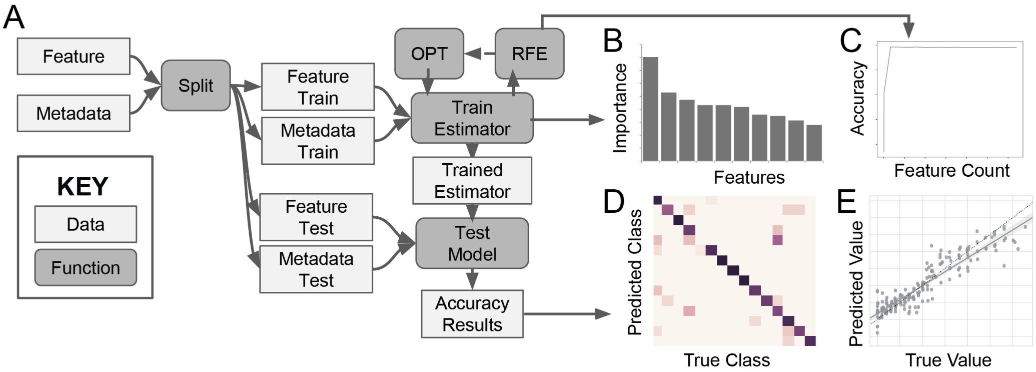

Figure 1: Workflow schematic (A) and output data and visualizations (B-E) for q2-feature-classifier. Data splitting, model training, and testing (A) can be accompanied by automatic hyperparameter optimization (OPT) and recursive feature elimination for feature selection (RFE). Outputs include trained estimators for re-use on additional samples, lists of feature importance (B), RFE results if RFE is enabled (C), and predictions and accuracy results, including either confusion matrix heatmaps for classification results (D) or scatter plots of true vs. predicted values for regression results (E).

sequence variants, operational taxonomic units, or taxa detected by marker-gene or shotgun metagenome sequencing methods), but any observation data, such as gene, transcript, protein, or metabolite abundance could be provided as input. Input samples are shuffled and split into training and test sets at a user-defined ratio (default: 4:1) with or without stratification (equal sampling per class label; stratified by default); test samples are left out of all model training steps and are only used for final model validation.

The user can enable automatic feature selection and hyperparameter tuning, and can select the number of cross-validations to perform for each (default = 5). Feature selection is performed using cross-validated recursive feature elimination via scikit-learn's RFECV to select the features that maximize predictive accuracy. Hyperparameter tuning is automatically performed using a cross-validated randomized parameter grid search via scikit-learn's RandomizedSearchCV to find hyperparameter permutations (within a sensible range) that maximize accuracy.

The following scikit-learn (Pedregosa et al., 2011) SL estimators are currently implemented in q2-sample-classifier: AdaBoost (Freund & Schapire, 1997), Extra Trees (Geurts, Ernst, & Wehenkel, 2006), Gradient boosting (Friedman, 2002), and Random Forest (Breiman, 2001) ensemble classifiers and regressors; linear SVC, linear SVR, and non-linear SVR support vector machine classifiers/regressors (Cortes & Vapnik, 1995); k-Neighbors classifiers/regressors (Altman, 1992); and Elastic Net (Zou & Hastie, 2005), Ridge (Hoerl & Kennard, 1970), and Lasso (Tibshirani, 1996) regression models.

Acknowledgments

The authors thank Jai Ram Rideout for his input and assistance integrating q2-sample-classifier with QIIME 2. This work was supported by the National Science Foundation [1565100 to JGC], and by the National Institutes of Health / National Cancer Institute Partnership for Native American Cancer Prevention [U54CA143924 and U54CA143925 to JGC].

References

Altman, N. S. (1992). An introduction to kernel and Nearest-Neighbor nonparametric regression. *Am. Stat.*, *46*(3), 175. doi:10.1080/00031305.1992.10475879

Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., D Lieber, A., et al. (2016a). Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.*, *8*(343), 343ra82. doi:10.1126/scitranslmed.aad7121

Bokulich, N. A., Collins, T. S., Masarweh, C., Allen, G., Heymann, H., Ebeler, S. E., & Mills, D. A. (2016b). Associations among wine grape microbiome, metabolome, and fermentation behavior suggest microbial contribution to regional wine characteristics. *MBio*, *7*(3). doi:10.1128/mBio.00631-16

Bokulich, N. A., Thorngate, J. H., Richardson, P. M., & Mills, D. A. (2013). Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. *Proceedings of the National Academy of Sciences*, *111*(1), E139–E148. doi:10.1073/pnas.1317377110

Breiman, L. (2001). Random forests. *Mach. Learn.*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Chang, H.-X., Haudenshield, J. S., Bowen, C. R., & Hartman, G. L. (2017). Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. *Front. Microbiol.*, *8*, 519. doi:10.3389/fmicb.2017.00519

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, *20*(3), 273–297. doi:10.1007/BF00994018

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, *55*, 119–139. doi:10.1006/jcss.1997.1504

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.*, *38*(4), 367–378. doi:10.1016/S0167-9473(01)00065-2

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.*, *63*(1), 3–42. doi:10.1007/s10994-006-6226-1

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi:10.1080/00401706.1970.10488634

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, *9*(3), 90–95. doi:10.1109/MCSE.2007.55

Jones, E., Oliphant, T., Peterson, P., & others. (2001). SciPy: Open source scientific tools for Python. Retrieved from http://www.scipy.org/

Knights, D., Kuczynski, J., Koren, O., Ley, R. E., Field, D., Knight, R., DeSantis, T. Z., et al. (2011). Supervised classification of microbiota mitigates mislabeling errors. *ISME J.*, *5*(4), 570–573. doi:10.1038/ismej.2010.148

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, 51–56.

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., & Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLoS Comput. Biol.*, *12*(7), e1004977. doi:10.1371/journal.pcbi.1004977

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, *12*, 2825–2830.

Schubert, A. M., Sinani, H., & Schloss, P. D. (2015). Antibiotic-induced alterations of the murine gut microbiota and subsequent effects on colonization resistance against clostridium difficile. *MBio*, *6*(4), e00974. doi:10.1128/mBio.00974-15

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M. A., Benezra, A., et al. (2014). Persistent gut microbiota immaturity in malnourished bangladeshi children. *Nature*, *510*(7505), 417–421. doi:10.1038/nature13421

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, *58*(1), 267–288.

Walt, S. van der, Colbert, S. C., & Varoquaux, G. (2011). The numpy array: A structure for efficient numerical computation. *Comput. Sci. Eng.*, *13*(2), 22–30. doi:10.1109/MCSE.2011.37

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., et al. (2017, September). Mwaskom/seaborn: V0.8.1 (september 2017). doi:10.5281/zenodo.883859

Yazdani, M., Taylor, B. C., Debelius, J. W., Li, W., Knight, R., & Smarr, L. (2016). Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. In *2016 IEEE international conference on big data (big data)*. doi:10.1109/BigData.2016.7840731

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, *67*, 301–320. doi:10.1111/j.1467-9868.2005.00503.x