SciForum
Mol2Net

Using the RRegrs R package for Automating Predictive Modelling

Georgia Tsiliki [1,*], Cristian R Munteanu [2], Jose A Seoane [3], Carlos Fernandez-Lozano [2], Haralambos Sarimveis [1] and Egon L Willighagen [4]

[1] School of Chemical Engineering, National Technical University of Athens, 15780, Greece; E-Mails: gtsiliki@central.ntua.gr (G.T.); hsarimv@central.ntua.gr (H.S.)

[2] RNASA-IMEDIR Group, Computer Science Faculty, University of A Coruna, 15071 A Coruña, Spain; E-Mails: crm.publish@gmail.com (CR.M.); carlos.fernandez@udc.es (C.F.-L.)

[3] Stanford Cancer Institute, Stanford University, C. J. Huang Building, 780 Welch Road, Palo Alto, CA 94304, USA; E-Mail: seoane@stanford.edu

[4] Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, P.O. Box 616, UNS50 Box 19, 6200 MD Maastricht, The Netherlands.; E-Mail: egon.willighagen@gmail.com

* Author to whom correspondence should be addressed; E-Mails: gtsiliki@central.ntua.gr; Tel.: +30-210-7723-236; Fax: +30-210-7723-138.

Published: 4 December 2015

Abstract: Cheminformatics and bioinformatics are extensively using predictive modelling and exhibit a need for standardization of methodologies such as data splitting, cross-validation methods, best model criteria and Y-randomization. RRegrs is a new R package, available at https://www.github.com/enanomapper/RRegrs (0.05 release), which suggests an integrated framework to assist model selection and speed up the process of predictive model development. The tool proposes a fully validated scheme by employing repeated 10-fold and leave-one-out cross-validation for ten linear and non-linear regression methods. Standardized reports are produced to compare the output of modelling algorithms and assess cross-validation results for selected models. Here, we demonstrate RRegrs capabilities in terms of performance using five well-established data sets.

Keywords: Multiple regression; QSAR; cross-validation; model selection

1. Introduction

RRegrs introduces an integrated framework for producing reliable and fully validated regression models in an automated way [1]. In its current release 0.05 (DOI:

10.5281/zenodo.32580), ten simple and complex regression methods are implemented, particularly: Multiple Linear regression (LM), Generalized Linear Model with Stepwise Feature Selection (GLM), Partial Least Squares regression (PLS), Lasso regression, Elastic Net regression (ENET), Support vector machine using radial functions (SVRM), Neural Networks regression (NN), Random Forest (RF), Random Forest-Recursive Feature Elimination (RF-RFE) and Support Vector Machines Recursive Feature Elimination (SVM-RFE). The methodology was implemented as an open source R package, available at https://github.com/enanomapper/RRegrs, by reusing and extending on the caret R package [2].

A single RRegrs function call is needed to run the entire workflow and obtain the produced validated models in a reproducible format. RRegrs suggests an easy way to explore the models' search space of linear and non-linear models with special parameters specifications and cross-validation (CV) schemes. Furthermore, model outputs are easily accessible and readable, organized by methods, centralized and averaged by multiple reproducible data set splits. Summary files are also produced helping the user to easily access all methodologies results, which can then be prioritized based on various statistics. A main feature of the package is its exhaustive validation scheme which introduces multiple random data splits. For each algorithm and data split, the model is produced based on training and validation sets, however, the test set is used to select the final best model. Parallel processing is enabled for accelerating the process.

2. Results and Discussion

Although the primary applications of RRegrs are aimed at finding Quantitative Structure – Activity Relationships (QSAR) models [3] under the settings of cheminformatics and nanotoxicology, here we demonstrate its efficiency for five standard data sets from UC Irvine Machine Learning Repository [4], using RRegrs current release 0.05. The five data sets considered, which are derived from diverse disciplines such as environmental economics and medical research, are the Housing [5], Computer Hardware, Wine Quality [6], Automobile [7] and Parkinsons Telemonitoring [8] data sets.

In Table 1 we present two statistic values for the five data sets, namely the R^2_{Test} and $RMSE_{Test}$ values, averaged over 10 different data splits and employing 10-fold repeated CV and 10 Y-randomizations. For all data sets, advanced methods such as RF-RFE and RF give the highest R^2_{Test} values. PLS is providing the poorest results in terms of both R^2_{Test} and $RMSE_{Test}$ values, whereas LM, GLM and LASSO are performing better in all cases but the Parkinson Telemonitoring data set. Very low $RMSE_{Test}$ values are observed, for instance SVRM method exhibits low $RMSE_{Test}$, although the corresponding R^2_{Test} values are generally lower compared to alternative methods.

Table 1. Averaged R^2_{Test} and $RMSE_{Test}$ values for the five data sets.

RRegrs	Housing		Computer h/w		Red wine		Automobile		Parkinson t/m	
method	R^2_{Test}	$RMSE_{Test}$	R^2_{Test}	$RMSE_{Test}$	R^2_{Test}	$RMSE_{Test}$	R^2_{Test}	$RMSE_{Test}$	R^2_{Test}	$RMSE_{Test}$
LM	0.707	0.111	0.822	0.056	0.355	0.131	0.824	0.085	0.154	0.217
GLM	0.709	0.111	0.825	0.056	0.353	0.131	0.824	0.085	0.153	0.217
PLS	0.660	0.120	0.793	0.064	0.331	0.133	0.784	0.098	0.121	0.221
LASSO	0.704	0.112	0.828	0.055	0.354	0.131	0.831	0.084	0.154	0.217
ENET	0.706	0.112	0.825	0.056	0.355	0.131	0.828	0.085	0.154	0.217
SVRM	0.845	0.080	0.765	0.066	0.396	0.127	0.853	0.075	0.637	0.142
NN	0.844	0.081	0.882	0.043	0.367	0.130	0.795	0.095	0.535	0.161
RF	0.874	0.074	0.909	0.045	0.501	0.115	0.915	0.059	0.972	0.040
RF-RFE	0.876	0.074	0.894	0.046	0.503	0.115	0.915	0.058	0.900	0.084
SVMRFE	0.717	0.120	0.692	0.124	0.378	0.129	0.728	0.151	0.479	0.173

3. Materials and Methods

In order to run RRegrs with full functionality a call to the RRegrs() function is required. All parameters have default values; a detailed list of parameters and functions' descriptions is given in the RRegrs package tutorial available online at https://github.com/enanomapper/RRegrs/blob/master/RRegrs-package-tutorial.pdf. Within the default values a default location for the output files is set, execution of all modelling steps (removal of NA, and near zero variance features, and of correlated features), normalization of the data set, ten splits, ten Y-randomization steps, and running of all ten regression methods. RRegrs function calls can be integrated into complex desktop and web tools for QSAR modelling.

A simple call to the function for a data set file named "MyDataSet.csv" and an output repository "MyResultsFolder" is the following:

```
>library(RRegrs)
>RRegrsResults<- RRegrs(DataFileName="MyDataSet.csv",
PathDataSet="MyResultsFolder")
```

4. Conclusions

RRegrs integrates results of individual models and decides on the best model given the data set and the user specified parameters. We have demonstrated its performance with five well-established data sets and showed that good performance results are produced in all cases. Its efficiency suggests that RRegrs can be used as a reliable fully-validated and automated predictive modelling framework, and a baseline for comparable results across various studies.

Acknowledgments

This work was supported by the eNanoMapper project, funded by the European Union's Seventh Framework Programme for research, technological development and demonstration (FP7 NMP-2013-SMALL-7) under grant agreement no 604134. The authors acknowledge the support by the Galician Network of Drugs R+D REGID (Xunta de Galicia R2014/025) and by "Collaborative Project on Medical Informatics (CIMED)" P I 13/00280 funded by the Carlos III Health Institute from the

Spanish National plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Tsiliki, G.; Munteanu, C.R.; Seoane, J.A.; Fernandez-Lozano, C.; Sarimveis, H.; Willighagen, E.L. RRegrs: an R package for computer-aided model selection with multiple regression models. *Journal of cheminformatics* **2015**, 7(1): 1-16.
2. Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software* **2008**, 28(5): 1-26.
3. Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR and Combinatorial Science* **2007**, 26(5): 694.
4. UC Irvine Machine Learning Repository. Available online: http://archive.ics.uci.edu/ml/ (accessed on 6th November 2015).
5. Harrison, D.; Rubinfeld, D.L. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management* **1978,** 5(1): 81-102.
6. Cortez, P.; Cerdeira, A.; Almeida, F.; Matos, T.; Reis, J. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* **2009**, 47(4): 547-553.
7. Kibler, D.; Aha, D.W.; Albert, M.K. Instance-based prediction of real-valued attributes. *Computational Intelligence* **1989,** 5(2): 51-57.
8. Tsanas, A.; Little, M.; McSharry, P.E.; Ramig, L.O. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering* **2010,** 57(4): 884-893.