Easy Execution of Data Mining Models through PMML

Alex Guazzelli*, Wen-Ching Lin and Michael Zeller

Zementis, Inc., San Diego, CA, USA
* Contact author: Alex.Guazzelli@zementis.com

Keywords: PMML, ADAPA, Model Deployment, Cloud Computing, Model Execution

PMML (Predictive Modeling Markup Language) is an XML-based language used to define data mining models. It was specified by the Data Mining Group, an independent group of leading technology companies. By providing a uniform standard to represent predictive models, PMML allows for the exchange of predictive solutions between different applications and various vendors. Many statistical packages already support the PMML standard; these include, for example, SAS and SPSS. In an effort to broaden the scientific workbench available to data mining scientists and to support the open source community, Zementis recently contributed code to the R project. In particular, we implemented the export of neural network models built with the *nnet* R package available through the *VR bundle* package (Venables and Ripley, 2002) as well as Support Vector Machines built with the *kernlab* R package (Karatzoglou et al., 2008) for objects of class *ksvm*. The same PMML exporter package (Williams et al., 2009) can also produce decision trees built with *rpart* (Therneau and port by Brian Ripley, 2008) and linear regression models as well as binary logistic regression models for objects of class *lm* and *glm* from *stats*. The PMML exporter package is currently available through CRAN (the Comprehensive R Archive Network).

All of the R exported PMML 3.2 models are readily available to be uploaded into an execution engine for scoring or classification. For example, the ADAPA engine, which can be used for testing and exploration, can be downloaded as a gadget and added to a personalized iGoogle console. This service is available free of charge and leverages the Amazon Elastic Compute Cloud (Amazon EC2).

Our aim here is to show how one can quickly build a data mining model in R, such as a Support Vector Machine, and use the PMML package to produce a model file which can be uploaded and executed in a different application. We demonstrate how one can use data containing expected results to verify correct model deployment. If all computed and expected values match, the model can be considered ready for production, i.e. available for generating predictions on incoming data as part of an overall enterprise decision management strategy. From R to ADAPA, we use PMML as an effective way to express and execute data mining models.

Our work shows how PMML can be effectively used to allow for model exchange between different applications. Also, it highlights how one can benefit from an open-source statistical package such as R to easily export models into PMML and upload them into ADAPA, a light-weight scoring engine which consumes several PMML 3.2 models and data transformations. The ease of model expression and execution allows data mining scientists to concentrate on the important tasks: data analysis and model building. Real-time, scalable execution is handled through software tools which communicate through a common language, PMML.

References

Data Mining Group (2009). *PMML version 3.2*,
http://www.dmg.org/pmml-v3-2.html.

A. Karatzoglou, A. Smola, and K. Hornik (2008). *The kernlab package*.
http://cran.R-project.org/web/packages/kernlab. R package version 0.9-8.

T. M. Therneau and B. A. R. port by Brian Ripley (2008). *Rpart: Recursive Partitioning*.
http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm. R package version 3.1-42.

W. N. Venables and B. D. Ripley (2002). *Modern Applied Statistics with S.* Statistics and Computing. Springer, New York, 4th edition.
http://cran.R-project.org/web/packages/VR. R package version 7.2-45.

G. Williams, M. Harshler, A. Guazzelli, M. Zeller, W. Lin, H. Ishwaran, U. B. Kogalur, and R. Guha. (2009). *PMML: Generate PMML for various models*.
http://rattle.togaware.com/. R package version 1.2.7.