The PMMLTransformations package

Tridivesh Jena[1,*] , Alex Guazzelli[1,] , Wen Ching Lin[1] , Michael Zeller[1]

1. Zementis, Inc.
*Contact author: tridivesh.jena@zementis.com

Keywords: R, Predictive Analytics, PMML, Data Transformations, Standards

PMML, the Predictive Model Markup Language, is the de facto standard to represent predictive analytic models [1,2,3]. With PMML, is is extremely easy to move the model from the scientist's desktop to the operational IT environment for real-time execution or batch scoring, since there is no recoding necessary.

Typically, the analytic process involves quite a bit of data pre-processing before a predictive model is build; the R **pmmlTransformations** package was designed to cover just such a need [4,5]. This package not only enables a data scientist to transform input data but also, when used along with the R **pmml** package [6,7], makes it possible to represent the transformation steps in PMML. The produced PMML file will then contain the combination of the predictive model itself together with all the steps necessary to pre-process incoming data. This is remarkable, since it allows for systems and applications to connect directly to the raw input data and leave it up to the PMML consumer to deliver the expected predictions.

The **pmmlTransformations** package implements many of the commonly used transformation operators used by data scientists, among them the Z-transform, linear transformation, data discretization, data normalization and value mapping. The result is not only the transformed data itself but also information to represent the transformation operators in PMML format. In this work, we illustrate the steps necessary to: 1) acquire raw data; 2) pre-process the raw data through available operators; 3) gain access to the manipulated data; and 4) output all the performed operations in PMML format. We also describe the various settings and options associated to each transformation operator and available to the data scientist. After the entire predictive workflow is exported into PMML, the predictive model or solution can easily be operationally deployed and executed in a variety of platforms including Hadoop, in-database or cloud computing.

References

[1] The Data Mining Group (DMG) website: www.dmg.org

[2] A. Guazzelli, W. Lin, T. Jena (2010). *PMML in Action: Unleashing the Power of Open Standards for Data Mining and Predictive Analytics (2nd Edition)*. CreateSpace (available on Amazon.com).

[3] A. Guazzelli (2010). What is PMML? Explore the power of predictive analytics and open standards. IBM developerWorks website.

[4] T. Jena, A. Guazzelli, W. Lin, M. Zeller (2013). The R *pmmlTransformations* Package. *In Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

[5] The R pmmlTransformations package: http://cran.r-project.org/web/packages/pmmlTransformations/index.html>

[6] A. Guazzelli, M. Zeller, W. Lin, G. Williams (2009). **PMML: An Open Standard for Sharing Models**. *The R Journal*, Volume 1/1.

[7] The R pmml package: http://cran.r-project.org/web/packages/pmml/index.html