Mach Learn (2012) 86:169–207
DOI 10.1007/s10994-011-5258-3

RE-EM trees: a data mining approach for longitudinal and clustered data

Rebecca J. Sela · Jeffrey S. Simonoff

Received: 9 February 2010 / Accepted: 14 June 2011 / Published online: 13 July 2011

Abstract Longitudinal data refer to the situation where repeated observations are available for each sampled object. Clustered data, where observations are nested in a hierarchical structure within objects (without time necessarily being involved) represent a similar type of situation. Methodologies that take this structure into account allow for the possibilities of systematic differences between objects that are not related to attributes and autocorrelation within objects across time periods. A standard methodology in the statistics literature for this type of data is the mixed effects model, where these differences between objects are represented by so-called "random effects" that are estimated from the data (population-level relationships are termed "fixed effects," together resulting in a mixed effects model). This paper presents a methodology that combines the structure of mixed effects models for longitudinal and clustered data with the flexibility of tree-based estimation methods. We apply the resulting estimation method, called the RE-EM tree, to pricing in online transactions, showing that the RE-EM tree is less sensitive to parametric assumptions and provides improved predictive power compared to linear models with random effects and regression trees without random effects. We also apply it to a smaller data set examining accident fatalities, and show that the RE-EM tree strongly outperforms a tree without random effects while performing comparably to a linear model with random effects. We also perform extensive simulation experiments to show that the estimator improves predictive performance relative to regression trees without random effects and is comparable or superior to using linear models with random effects in more general situations.

Keywords Clustered data · Longitudinal data · Panel data · Mixed effects model · Random effects · Regression tree · CART

Editor: Johannes Fürnkranz.

R.J. Sela (✉)
J.P. Morgan Chase & Co., Columbus, OH, USA
e-mail: rebeccapaul@yahoo.com

J.S. Simonoff
Statistics Group, Information, Operations, and Management Sciences Department, Leonard N. Stern School of Business, New York University, New York, NY, USA