# SMOTE for Regression

Luís Torgo[1,2], Rita P. Ribeiro[1,2], Bernhard Pfahringer[3], and Paula Branco[1,2]

[1] LIAAD - INESC TEC
[2] DCC - Faculdade de Ciências - Universidade do Porto
[3] Department of Computer Science - University of Waikato
{ltorgo,rpribeiro}@dcc.fc.up.pt, bernhard@cs.waikato.ac.nz,
paobranco@gmail.com

**Abstract.** Several real world prediction problems involve forecasting rare values of a target variable. When this variable is nominal we have a problem of class imbalance that was already studied thoroughly within machine learning. For regression tasks, where the target variable is continuous, few works exist addressing this type of problem. Still, important application areas involve forecasting rare extreme values of a continuous target variable. This paper describes a contribution to this type of tasks. Namely, we propose to address such tasks by sampling approaches. These approaches change the distribution of the given training data set to decrease the problem of imbalance between the rare target cases and the most frequent ones. We present a modification of the well-known SMOTE algorithm that allows its use on these regression tasks. In an extensive set of experiments we provide empirical evidence for the superiority of our proposals for these particular regression tasks. The proposed SMOTER method can be used with any existing regression algorithm turning it into a general tool for addressing problems of forecasting rare extreme values of a continuous target variable.

## 1 Introduction

Forecasting rare extreme values of a continuous variable is very relevant for several real world domains (e.g. finance, ecology, meteorology, etc.). This problem can be seen as equivalent to classification problems with imbalanced class distributions which have been studied for a long time within machine learning (e.g. [1–4]). The main difference is the fact that we have a target numeric variable, i.e. a regression task. This type of problem is particularly difficult because: i) there are few examples with the rare target values; ii) the errors of the learned models are not equally relevant because the user's main goal is predictive accuracy on the rare values; and iii) standard prediction error metrics are not adequate to measure the quality of the models given the preference bias of the user.

The existing approaches for the classification scenario can be cast into 3 main groups [5, 6]: i) change the evaluation metrics to better capture the application bias; ii) change the learning systems to bias their optimization process to the goals of these domains; and iii) sampling approaches that manipulate the

L. Correia, L.P. Reis, and J. Cascalho (Eds.): EPIA 2013, LNAI 8154, pp. 378–389, 2013.

training data distribution so as to allow the use of standard learning systems. All these three approaches were extensively explored within the classification scenario (e.g. [7, 8]). Research work within the regression setting is much more limited. Torgo and Ribeiro [9] and Ribeiro [10] proposed a set of specific metrics for regression tasks with non-uniform costs and benefits. Ribeiro [10] described system ubaRules that was specifically designed to address this type of problem. Still, to the best of our knowledge, no one has tried sampling approaches on this type of regression tasks. Nevertheless, sampling strategies have a clear advantage over the other alternatives - they allow the use of the many existing regression tools on this type of tasks without any need to change them. The main goal of this paper is to explore this alternative within a regression context. We describe two possible methods: i) using an under-sampling strategy; and ii) using a Smote-like approach.

The main contributions of this work are: i) presenting a first attempt at addressing rare extreme values prediction using standard regression tools through sampling approaches; and ii) adapting the well-known and successful Smote [8] algorithm for regression tasks. The results of the empirical evaluation of our contributions provide clear evidence on the validity of these approaches for the task of predicting rare extreme values of a numeric target variable. The significance of our contributions results from the fact that they allow the use of any existing regression tool on these important tasks by simply manipulating the available data set using our supplied code.

## 2 Problem Formulation

Predicting rare extreme values of a continuous variable is a particular class of regression problems. In this context, given a training sample of the problem, $\mathcal{D} = \{\langle \mathbf{x}, y \rangle\}_{i=1}^{N}$, our goal is to obtain a model that approximates the unknown regression function $y = f(\mathbf{x})$. The particularity of our target tasks is that the goal is the predictive accuracy on a particular subset of the domain of the target variable $Y$ - the rare and extreme values. As mentioned before, this is similar to classification problems with extremely unbalanced classes. As in these problems, the user goal is the performance of the models on a sub-range of the target variable values that is very infrequent. In this context, standard regression metrics (e.g. mean squared error) suffer from the same problems as error rate (or accuracy) on imbalanced classification tasks - they do not focus on the rare cases performance. In classification the solution usually revolves around the use of the precision/recall evaluation framework [11]. Precision provides an indication on how accurate are the predictions of rare cases made by the model. Recall tells us how frequently the rare situations were signalled as such by the model. Both are important properties that frequently require some form of trade-off. How can we get similar evaluation for the numeric prediction of rare extreme values? On one hand we want that when our models predict an extreme value they are accurate (high precision), on the other hand we want our models to make extreme value predictions for the cases where the true value is an extreme (high recall).