Information Sciences 374 (2016) 65–84

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Improving regression predictions using individual point reliability estimates based on critical error scenarios

Elia Yathie Matsumoto*, Emilio Del-Moral-Hernandez

Electronic Systems Engineering Department, University of Sao Paulo, Brazil

ARTICLE INFO

Article history:
Received 29 April 2015
Revised 10 September 2016
Accepted 12 September 2016
Available online 13 September 2016

Keywords:
Improvement of computed regression predictions
Individual reliability estimates
Machine Learning
Pattern recognition
Imbalanced datasets
Artificial Neural Networks

ABSTRACT

We present a method to improve computed regression prediction values for unseen data. It aims at obtaining more accurate results by adjusting the calculated predictions instead of by constructing a different regression model. As a result, it can be helpful to improve the prediction of a specific observation provided by an existing benchmark regression model or predictor system. The proposed methodology uses individual point reliability estimates that indicate if a single regression prediction is likely to produce an error considered critical by the user of the regression. We tested the method in two sets of experiments, one using synthetically produced data, and the other using data from the public data repository UCI Machine Learning. The experiments with synthetic data were performed to verify the efficiency of the method under controlled situations. In this case, the method produced superior results improving predictions for cleaner data with progressive worsening with the increase of the noise level. Experiments with ten databases from the UCI data repository were executed to investigate the applicability of the methodology using real world data. The method was able to correctly adjust regressions prediction values in experiments with all the ten databases, achieving statistically significant improvement in eight of them.

1. Introduction

This work proposes a method to improve the predictions provided by an existing regression model without the need for changing either its architecture or its calibrated parameters. In other words, we are interested in obtaining more accurate results by correcting the calculated regression prediction values instead of by constructing a different regression model.

The proposed methodology uses an individual point reliability estimate constructed by applying a refined concept of a method first drafted in our previous works [29,30] in which we presented a procedure to create point reliability estimates for individual regression predictions based on critical error scenario definitions determined by the user of the regression.

The basic idea is that these individual reliability estimates, which are potentially able to correctly identify these critical error scenario cases, can be used to adjust and to improve computed regression prediction values. To investigate this presumption, the experiments herein were carried out using two types of datasets: artificially produced synthetic data and real data extracted from the public data repository UCI Machine Learning.

The experiments with synthetic data were performed to verify the behavior of the proposed adjustment procedure in the case of artificially created clean data and noisy data with different levels of noise. The response of the proposed methodology using real data was observed in experiments performed using ten databases from the UCI Machine Learning Repository.

* Corresponding author.
E-mail addresses: elia.matsumoto@usp.br (E.Y. Matsumoto), emilio@lsi.usp.br (E. Del-Moral-Hernandez).

http://dx.doi.org/10.1016/j.ins.2016.09.034
0020-0255/

In addition to this introduction, this work is composed of five other sections. The next section (Section 2) contains a review of the relevant research currently under development related to the improvement of computed individual regression predictions. The proposed methodology is detailed in Section 3. In Section 4, we describe how the experiments were conducted. The numerical results of the experiments are presented and discussed in Section 5. Conclusions are presented in Section 6, followed by the list of references.

2. Related works

In the extensive literature available on regression modeling techniques contemplating improving accuracy, there are relatively few works about methods to correct computed regression prediction values, and this specific research area is strongly related to the field of reliability estimation of individual regression predictions [9,36].

2.1. Reliability estimation of individual regression prediction

In regression analysis, average error measures, for instance, mean squared error or mean absolute error values are frequently adopted as general purpose prediction evaluation metrics [48]. In Statistics and Machine Learning [13], these average metrics are particularly suitable for use in cross-validation techniques to compare the outcome of models throughout processes such as model formulation choice, features selection, and model parameters calibration. However, these averaged error metrics are less appropriate for estimating the quality of regression predictions for single unseen observations.

The availability of additional information about a specific regression prediction is very desirable in decision-making processes, mainly in risk-sensitive areas. For this reason, research in the field of reliability estimation of individual predictions has increased in the last few decades. The methods addressing this matter are usually divided into two groups in the technical literature [36]. The first group consists of the methodologies that work with model-specific approaches. In this case, the concepts that support these methods are based on the mathematical definition of the regression model and the probabilistic properties of the data; and often analytical solutions are provided. The second group covers model-independent methods that essentially handle the regression model as a “black-box” object, considering just its inputs and outputs. As a result, these methods can be more widely applied, but rarely provide analytical solutions.

The classic linear regression confidence interval method can be considered one of the most conventional examples of the model-specific approach. This method assumes that, given a restrictive set of assumptions, the prediction errors produced by the linear model closely follow a known normal distribution; this information can be then used to construct a constant interval in which the regression prediction errors for unseen observations are supposed to fall within a certain probability, named confidence degree [31].

Another example of the model-specific approach is the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models in time series analysis. Differently from the classic linear regression method, in the case of GARCH models, constant variance (homoscedasticity) is not an assumption; on the contrary, these kinds of models estimate the individual variance of the prediction errors of the time series regression model as a function of the individual variance of the time series variable, and the error analysis is approached using autoregressive models [14].

In the case of nonlinear regressions, there are studies developed considering specific properties of predictive model architectures, such as Regression Trees [32], Support Vector Machines [42], and Neural Networks [11], among others, and a relatively smaller number of individual prediction reliability estimation methods belong to the model-independent group, which is also usually divided into two groups: (a) prediction interval estimate methods and (b) point estimate methods [36].

Similarly to the linear regression confidence interval, the prediction interval estimate methods (a) concern estimating an interval in which the regression prediction error for unknown or future observations are supposed to fall, within a certain probability, but without being necessarily attached to a parametric distribution; as a result, the two methods outcomes have different conceptual interpretations. In the case of a confidence interval with, for instance, 90% degree of confidence, it means that, considering all regression prediction errors, asymptotically, 90% of them are supposed to fall inside the interval [48]. On the other hand, the prediction interval methods estimate an interval with a certain probability of containing the regression prediction errors. Prediction intervals are frequently wider than confidence intervals and they can vary for each of the observations [22].

Particularly regarding the construction of prediction intervals for nonlinear regression models, there are studies exploring traditional approaches, such as empirical distributions [24], bootstrap techniques [39], maximum likelihood properties [34], Bayesian inference theory [23], and other distinct frameworks; for example, Lower Upper Bound Estimation [25], Conformal Prediction [35], and non-parametric approaches using supervised learning framework [37,38], among others.

The other group of reliability estimates for individual regression predictions afore mentioned, (b) the point estimates class provides a value, instead of an interval, as additional information about individual prediction reliability to help the user of the regression gain auxiliary insight about the individual future predictions. Most point estimate methods described in the literature are based on estimates generated by the analysis of how modifications in the data affect the outcomes of the models [36]. The previously cited bootstrap technique and its variants, such as bagging and boosting, are popular examples of this kind of methodology [12].The basic principle is to repeatedly modify the initial set of the known data (the training or learning dataset) applying bootstrap techniques, and then creating one regression model for each of these variations. This collection of regression models is then combined so as to produce a more accurate model, and also to provide the reliability