Predicting Rare Extreme Values

Luis Torgo and Rita Ribeiro

LIACC-FEP, University of Porto, R. de Ceuta, 118, 6., 4050-190 Porto, Portugal
[ltorgo,rita]@liacc.up.pt,
WWW home page: http://www.liacc.up.pt/~[ltorgo,rita]

Abstract. Modelling extreme data is very important in several application domains, like for instance finance, meteorology, ecology, etc.. This paper addresses the problem of predicting extreme values of a continuous variable. The main distinguishing feature of our target applications resides on the fact that these values are rare. Any prediction model is obtained by some sort of search process guided by a pre-specified evaluation criterion. In this work we argue against the use of standard criteria for evaluating regression models in the context of our target applications. We propose a new predictive performance metric for this class of problems that our experiments show to perform better in distinguishing models that are more accurate at rare extreme values. This new evaluation metric could be used as the basis for developing better models in terms of rare extreme values prediction.

1 Introduction

In several applications the main focus of interest is a small proportion of the available data. These unusual cases have a large importance, and as such, anticipating them is a critical task for these domains. An example of such applications is the prediction of the future returns of a stock. Unusually high (low) returns are rare, but they are the most interesting values for investors and thus they should be the target of any financial prediction model.

A related problem has been receiving great attention in the data mining community: the construction of classification models based on samples with unbalanced class distributions (e.g. [1]). Predicting extreme values of a continuous variable can be handled through a classification approach by means of a discretization process (e.g. [2]) off the continuous target variable. This would have the advantage of using all work that has been around in the areas of unbalanced classification problems and evaluation under differentiated misclassification costs. However, this approach would require to establish the number of classes and, moreover, would lead to an undesirable crisp division between what is an extreme and what is a "normal" case. These are some of the major drawbacks of handling regression as a classification problem[1].

The problem of predicting rare extreme values is a particular case of multiple regression where a target continuous variable Y is being modelled using a set

[1] More details on this argument can be found on [4],an extended version of this work.

of predictor or input variables $X_1, X_2, \cdots, X_p$. Any modelling method tries to find the model parameters that minimise an error function over the training sample. Standard functions used in regression setups are the Mean Squared Error, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, or the Mean Absolute Deviation, $MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. Both these measures take all errors equally (with the same cost) and thus can be regarded as less adequate for our target applications, where errors on extreme values are more important.

One possible method for giving more weight to the errors on extreme values is to use case weights. Some algorithms allow the user to attach a weight to each case of the training sample. Model parameters can then be obtained by minimising a criterion that takes into account these weights. Using case weights that depend on the respective Y value being an extreme allows us to bias the obtained model to correctly predict these extreme cases. The main drawback of this approach is that it only sees one side of the problem, the true values. In effect, this method does not try to avoid (or penalise) the cases where an extreme value is predicted by the model, but the truth value is "normal", i.e. false positives according to the classification terminology. This drawback stems from the fact that the weights are dependent solely on the true value of the cases, y_i, instead of being dependent on both y_i and $\hat{y}_i$. Our proposal builds upon this idea by trying to eliminate this drawback through the use of a weight function that depends on both y_i and $\hat{y}_i$.

2 Our Proposal

The overall goal of this work is to have an evaluation metric that is biased towards valuating more the predictions of rare extreme values. Our proposal was developed with the following requirements in mind: i) the cost of a prediction error should depend on both the predicted and the true values, i.e. we should penalise both false positives and false negatives; ii) the cost of the errors should vary smoothly (no crisp divisions between extremes and non-extremes); iii) the method should have reasonable default costs (according to the overall goal) for applications where knowledge about the costs is not available.

We propose an evaluation metric that is basically a weighted average of the errors. Our key contribution lies on the form of calculating the weights. We use a weight function that depends on both the true and predicted values. We propose to use a smooth cost surface, $w(Y, \hat{Y})$, that can be seen as a continuous version of cost matrixes used in classification tasks. Summarising, our proposed Rare Extremes Error metric is defined as,

$$RExE = \frac{1}{n}\sum_{i=1}^{n} w(y_i, \hat{y}_i) \times L(y_i, \hat{y}_i) \tag{1}$$

where $L(y_i, \hat{y}_i)$ can be any loss function, e.g. the squared error.

In order to make the use of smooth cost surfaces practical we need to devise an easy way of specifying them. Our proposal consists of requiring the specification of the cost values at a small set of properly selected points and then using

a function approximation method to interpolate the complete surface. The axes of the surface are the true, Y, and predicted, $\hat{Y}$, values of the target variable. These range from low extreme values ($extr_L$) to high extreme values ($extr_H$). The points selected for specifying the cost surface should be related to the most relevant areas of the surface. These are the areas of lower cost (the model accurately predicts and extreme as such), and of the worse performance (the model predicts an extreme high for a true extreme low, or vice versa).

For applications where no cost information is available but still extremes are more important, we need to describe means to setup the costs for the key points used for surface approximation. The critical question is to define what is a rare extreme value. We use the same definition as in Torgo and Ribeiro [3]. This means that we set $extr_L = adj_L$ and $extr_H = adj_H$, where $adj_L(adj_H)$ is the smallest observation that is greater or equal to the 1st quartile minus $1.5r$, with r being the interquartile range. After having defined these two extreme values we artificially create n grid points by diving the interval $extr_H - extr_L$ in n equally spaced bins. This means that we will have a $(n+2)$x$(n+2)$ matrix to fill in with costs. We use an arithmetic progression to setup the costs from the lowest to the highest cost. Full details and illustrative examples can be found in [4].

3 An Experimental Evaluation of the Proposal

We have carried out a series of experiments with the goal of checking the validity of our proposed metric in the task of identifying the models that are better from the perspective of being more accurate at rare extreme values. With this purpose we have designed the following experimental setup for each data set:

1. Draw a stratified test sample with 50% of the cases;
2. Randomly generate a set of prediction errors with the same size as the test sample. The errors are drawn from a normal distribution. We then pick the n largest errors, where n is the number of extreme values of the distribution of Y, and increase these errors by a constant k. The overall objective of this step is to obtain a set of credible prediction errors for a standard model when making predictions for a problem with some extremes. For this type of problems we expect (we have confirmed this experimentally using several modelling techniques and several real world data sets), the models to achieve a performance of this type: normal-shape distribution of the error with some extreme errors typically occurring on test cases with extreme values of Y.
3. We then artificially allocate this set of generated errors to each case on the test set in two different ways, leading to the "artificial performance" of models A and B. For model A, the smallest errors are allocated to the extremes in the test set, thus leading to what could be considered to be an ideal model for our target applications. On the contrary, model B has the largest errors on the extreme values of the target, in what could be considered a "normal" behaviour of a model in this type of tasks.

Table 1. The results in terms of percentage difference between Models A and B.

Data Set	SigMetric (avg±sd)	NRExE (avg±sd)	Data Set	SigMetric (avg±sd)	NRExE (avg±sd)
algae1	52.6±4.8	**82.9±1**	deltaAilerons	**55.2±0.6**	5±0.4
algae2	55±4.6	**79.3±1.6**	ibm	**71.9±0.3**	7.4±0.4
algae3	71.1±2.4	**88.1±1**	abalone	**70.5±1.1**	6.5±0.5
algae4	73.8±14.1	**87.3±5.1**	cpuSmall	63±1	**81.1±0.4**
algae5	56.1±4.6	**83.9±1.5**	servo	74.4±5.8	**85.2±0.9**
algae6	84.2±0.8	**91.1±0.5**	cwDrag	**57.4±1.6**	2.8±2.6
algae7	52.4±11.1	**82.7±2.3**	co2Emission	**58.4±0.6**	17.8±6.5
Boston	**65±1.6**	21±25.3	availablePower	69.7±1.5	**71.5±0.7**
machineCpu	76.5±3.4	**77.9±1**	china	68.9±2.8	**71.8±1.4**
bank8FM	55.3±0.5	**63.6±0.8**	add	**56.6±0.3**	5.5±0.7

A performance metric that is biased towards accurate predictions on extremes, should clearly indicate that the performance of Model A is better than the performance of Model B. Notice that, given that the errors of the two models are exactly the same (only occurring at different test cases), metrics like the MSE or the MAD will show both models as having exactly the same score.

As we are testing on a large set of domains with a quite different range of target variable values, we have used a normalised version of our performance statistic to allow comparisons across domains,

$$NRExE = \frac{\sum_{i=1}^{n_{test}} w(y_i, \hat{y}_i) \cdot |y_i - \hat{y}_i|}{\sum_{i=1}^{n_{test}} w(y_i, \widetilde{Y}) \cdot |y_i - \widetilde{Y}|} \tag{2}$$

where $\widetilde{Y}$ is the sample median.

The goal of our experiments is to assert the score difference between models A and B, when evaluating them using our proposed metric and an alternative measure. With this purpose we have measured the percentual difference of scores for all data sets. Positive values of this difference indicate that our metric is able to identify Model A as performing better than Model B. We obviously want the difference to be as high as possible, as Model A has an "ideal" performance. We have compared our proposed metric, $NRExE$, against the score obtained by the most similar alternative, an error measure using case weights as mentioned in Section 1. For this competitor we have setup the case weights such that more weight is given to cases with extreme values of the target (details on [4]). Notice that contrary to our approach the weights of this measure only consider the true value of the target, thus not taking into account the predictions of the models.

For each data set we have repeated the experiment outlined above 10 times. The results shown on Table 1 are the average and standard deviation of the observed percentual differences between Model A and B, when using $NRExE$ and the metric with sigmoid-based case weights. The best scores for each data set are indicated in bold. The used datasets are real world problems with a diverse set of rare extreme values types. For instance, some include both low and high

extremes, while others include only one type of extremes. Due to space reasons we are not able to present the full characteristics of these problems.

The results reported in Table 1 show the advantages of our proposed metric for domains where the main objective is to be accurate at rare extreme values. In effect, in most problems our metric correctly signals model A as being significantly better than model B, in spite of being compared against a competitor metric that also take extremes into account. Notice that standard measures, like MSE, would signal both models as being equal (difference equal to zero).

4 Conclusions

In this paper we have described the particular features of a class of problems with high practical importance: the prediction of rare extreme values. We claim that existing metrics for evaluating the performance of different models have several drawbacks and perform poorly on identifying the best models in terms of predictive accuracy on the most important cases for these applications. We have presented a new metric that is particularly suited for these applications.

In a set of experiments using real world data we have shown that this measure is able to identify the best model in terms of accuracy on the rare extreme values, even on the most difficult scenario where both models have exactly the same error distribution and thus have the same score in "standard" metrics like MSE.

One of the main impacts of the results of this work is that our metric can be used to compare different existing models on tasks where the main goal is the accuracy on rare extreme values. The use of our metric should provide better information concerning the merits of alternative models for these important tasks. Another important side effect of this work is the possibility of using the described metric in the search process of any modelling technique, so as to develop models that are built for maximising the predictive performance on extreme values.

Acknowledgements

This work was supported by FCT project MODAL (POSI/SRI/40949/2001) co-financed by POSI and by the European fund FEDER and by a PhD scholarship of FCT (SFRH/BD/1711/2004) to Rita Ribeiro.

References

1. G.Weiss and F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *JAIR*, 19:315–354, 2003.
2. L. Torgo and J. Gama. Regression using classification algorithms. *Intelligent Data Analysis*, 1(4), 1997.
3. L. Torgo and R. Ribeiro. Predicting outliers. In N. Lavrac, D. Gamberger, L. Todorovski, and H. Blockeel, editors, *Proceedings of Principles of Data Mining and Knowledge Discovery (PKDD'03)*, number 2838 in LNAI, pages 447–458. Springer, 2003.
4. L. Torgo and R. Ribeiro. Predicting rare extreme values. Technical Report 2006-01, LIACC-NIAAD. University of Porto, 2006.
(`http://www.liacc.up.pt/~ltorgo/Papers/PREVext.pdf`).