Utility-based Performance Measures for Regression

Rita Ribeiro RPRIBEIRO@LIAAD.UP.PT
LIAAD - INESC PORTO LA, R. Ceuta, 118, 6., 4050-190 Porto, Portugal
FCUP - DCC, University of Porto, R. Campo Alegre, 1021/1055, 4169-007 Porto, Portugal

Luís Torgo LTORGO@LIAAD.UP.PT
LIAAD - INESC PORTO LA, R. Ceuta, 118, 6., 4050-190 Porto, Portugal
FEP, University of Porto, R. Dr. Roberto Frias, 4200-464 Porto, Portugal

Abstract

Costs and benefits are of key importance for most real-world data mining applications. The main work carried out within ML/DM on cost-sensitive learning has been centered on classification tasks. Nevertheless, there are many real-world regression applications where costs and/or benefits of the predictions play an important role. On such applications, the costs and/or benefits of a prediction can vary across the domain of the target variable. For instance, it may be much more relevant to be accurate at a certain range of values than on other parts of the target variable domain. In this context, it is important to address regression tasks from a cost-sensitive perspective. Using as an example the case of rare extreme values prediction, we show why the standard regression error metrics are no longer effective for this type of applications. We propose an utility-based evaluation framework, which allows for differentiated scores to be assigned to the predictions based on their cost/benefit in conformity with the application preference biases. Based on this utility concept, new performance metrics can be developed for a more reliable evaluation/comparison of models and also for model development.

1. Introduction

In many real-world applications, the main focus of interest is a small subset of values. Typically, these applications are related to cost-sensitive decision problems, where different predictions can trigger different decisions in which costs are often involved. Several cost-sensitive techniques (e.g. Domingos (1999); Fan et al. (1999); Elkan (2001); Zadrozny et al. (2006)) have been studied and developed in order to address such kind of applications. Still, in what concerns regression tasks not much has been done.

In effect, there are some real-world applications where the continuous target variable shows non-uniform costs across its domain. The anticipation of a critical phenomenon in domains such as finance, meteorology, ecology, fraud detection, etc. are among this kind of applications. These critical phenomena are usually described by a particular subset of values of the target continuous variable, and usually trigger some sort of alarm or action. So, predictions made for this subset of values should have a differentiated cost/benefit to be in accordance to the application biases.

Some authors (e.g. Christoffersen and Diebold (1996); Crone et al. (2005)) have already proposed new error metrics for handling differentiated costs in regression tasks. However, as it will be explained, they still have some drawbacks in terms of addressing the general problem of cost-sensitive regression.

In this work, we will present a new utility-based framework, based in one of our previous works (Torgo & Ribeiro, 2007), to address such cost-sensitive regression applications.

2. The Inadequacy of Standard Evaluation Approaches

The performance measures more commonly used in regression are the mean squared error, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, and the mean absolute deviation, $MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, where y_i and $\hat{y}_i$ represent

Appearing in *Proceedings of the ICML 3rd Workshop on Evaluation Methods for Machine Learning*, Helsinki, Finland, 2008.

the true and predicted values of the continuous target variable Y, respectively. Both measures take all the prediction errors equally across the domain of the target variable Y, as they only consider the error amplitude. For cost-sensitive regression tasks, this type of performance measures is no longer effective.

Suppose that we have an application problem where Y has a normal-like shape but with very heavy tails. Moreover, let us assume the most relevant values are the ones which lye on these tails of the distribution, that is, the rare and extreme values. We can find several prediction tasks in these conditions: anticipation of ecological or meteorological catastrophes (e.g. harmful algae bloom in a river, extreme weather conditions), financial forecasting (e.g. big stock market price changes), etc.. In all these domains, the most relevant values are often described by rare and extremely high or extremely low values of the target variable.

Imagine that on an example of this type of tasks the values above 40 are considered the most relevant. In Figure 1, we show the predictions of two hypothetical models for a set of test cases of such problem.

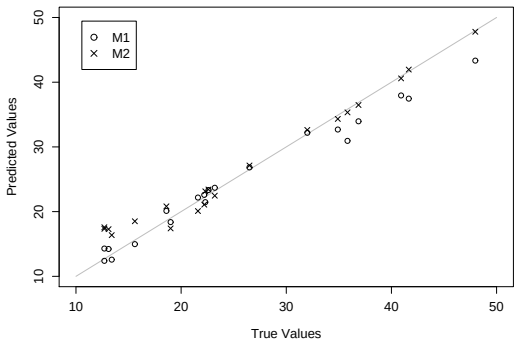

Figure 1. Predictions of two models

If we look at the performance of the two models, we can easily see that model $M2$ should be regarded as the best model, given the application preference biases. With $M2$ most accurate predictions occur on the rare extreme cases, whilst the opposite happens with $M1$, and thus $M2$ is much more useful. However, if we estimate their performance using MAD or MSE, we see that the two models are ranked *exequo* (*c.f.* Table 1).

Table 1. Estimated performance of two different models

	MAD	MSE
M1	1.5978	4.7454
M2	1.5978	4.7454

In effect, these are two artificially generated models designed to illustrate the drawbacks of standard metrics on cost-sensitive applications. $M2$ is obtained from $M1$ in such a way that the smaller errors (in amplitude) are allocated to the cases which have rare extreme values on the target variable. This means that both models have exactly the same error amplitudes but these occur for different test cases, $M2$ having the smaller errors on the most important test cases. This illustrates the problem of using metrics that are insensitive to the concrete values involved in the predictions, only looking at the error amplitudes. This is more serious when the values have non-uniform costs/benefits as it is the case on this small toy problem.

We could try to overcome this problem by considering a weighted error measure, such that higher weights are given to the rare extreme values cases. Nonetheless, this would only partially meet our application requirements. In effect, we want to avoid bad predictions for relevant values (missed events), but also predictions of relevant values for irrelevant test cases (i.e. false alarms). Therefore, the cost/benefit of a prediction should depend on both the true and predicted values.

Christoffersen and Diebold (1996) and Crone et al. (2005) have addressed the issue of differentiated prediction costs in the context of financial applications, proposing asymmetric loss functions. Their main goal was to be able to distinguish two types of errors, and assign costs accordingly, namely, the cost of under-predictions ($\hat{y} < y$) and the cost of over-predictions ($\hat{y} > y$). Although this is a step in the direction of cost-sensitive regression, they only distinguish these two types of situations and moreover, they consider all under-(over-) predictions as equally serious, taking only in consideration the error magnitude as in the standard error metrics. As such, this approach is far from satisfying our goal of having a general framework for handling cost-sensitive regression tasks.

3. An Utility-based Evaluation Framework

Our proposal builds upon the notion of utility, which incorporates both costs and benefits of the predictions of a model. The utility of a prediction $\hat{y}$ for a true value y will be defined by two factors: the relevance of $\hat{y}$ and y according to the application biases; and the relevance of the prediction error, given by a loss function $L(\hat{Y}, Y)$, and by a threshold t within the range of this loss function that establishes a kind of maximum admissible error.

3.1. Relevance Function

We define the relevance of the values of the target variable as in Torgo and Ribeiro (2007). We assume that the end-user provides us with a relevance function ϕ. This function maps the domain of the target variable into 0..1 scale, where 1 represents maximum relevance, i.e. $\phi(Y) : \mathbb{R} \rightarrow [0, 1]$. This step is equivalent to the specification of a cost matrix in the context cost-sensitive classification tasks. Still, on several applications the concept of relevance is often associated to the notion of rarity and extremeness of values. For such applications, if a relevance function is not available, we can obtain a reasonable approximation using the box-plot, as shown in Figure 2. With the statistical indicators returned by the box-plot of the target variable Y, we can setup a sigmoid-like function that grows smoothly as the level of rarity increases.

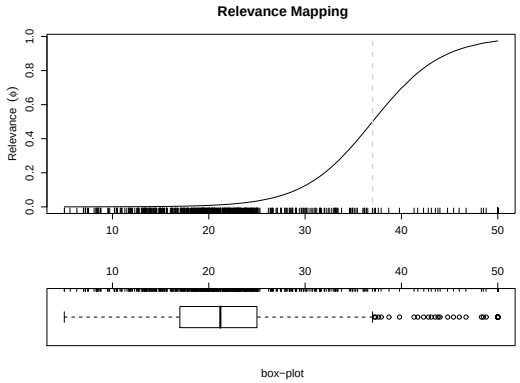

Figure 2. A sigmoid-based relevance function (ϕ)

3.2. From Prediction Errors to Utility Scores

The second factor of our proposal is related to the mapping of the prediction errors into cost/benefits, i.e. into a score of utility. Considering costs as negative benefits, we have single scale of benefits. We define the utility scores through the specification of the function $\zeta : \mathbb{R}_0^+ \rightarrow [-B, B]$, shown in Equation 1, where B is the maximum benefit, t establishes a maximum admissible error, and η is a decay parameter which ensures that the maximum benefit is achieved.

$$\zeta(L(\hat{Y}, Y)) = sgn(t - L(\hat{Y}, Y)) \cdot B \cdot \ (1 - e^{-\eta \cdot |t - L(\hat{Y}, Y)|}) \quad (1)$$

Without any crisp divisions, this function provides a smooth decrease from the maximum benefit B, achieved when the prediction error is zero, to the minimum benefit $-B$ (*c.f.* Figure 3).

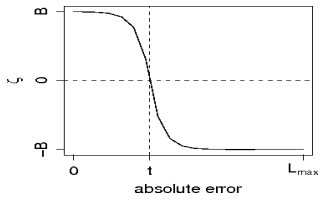

Figure 3. From prediction errors to utility scores.

The final utility of a prediction $\hat{y}$ for a true value of y also takes into consideration the relevance of both $\hat{y}$ and y. This value depends on whether we are facing a positive benefit ($L(\hat{y}, y) \leq t$) or a negative benefit ($L(\hat{y}, y) > t$), as described in Equation 2. In the first case, we are facing an admissible error, i.e. a kind of "accurate" prediction. Our utility value is proportional to the smallest relevance of both $\hat{y}$ and y. The idea is that the smallest of these two relevances should condition how much utility score (i.e. benefits) we get. In the second case, we are facing a kind of prediction "error". The cost we incur is proportional to the weighted average of the relevances of both $\hat{y}$ and y. This process allows us to distinguish the importance of false alarms from missed events (also known as opportunity costs), using a parameter p that sets the relative importance of these two types of errors.

$$U(\hat{Y}, Y) = \begin{cases} \min\left\{\phi(Y), \phi(\hat{Y})\right\} \cdot \zeta(L(\hat{Y}, Y)) & L(\hat{Y}, Y) \leq t \\ ((1-p) \cdot \phi(\hat{Y}) + p \cdot \phi(Y)) \cdot \zeta(L(\hat{Y}, Y)) & L(\hat{Y}, Y) > t \end{cases} \quad (2)$$

The function $U()$ defines an utility surface that can be regarded as a kind of continuous and smooth version of cost matrices. For instance, using this function, the relevance function shown in Figure 2, and a maximum admissible error amplitude of 15 (i.e. $t = 15$) we obtain the utility surface shown in Figure 4.

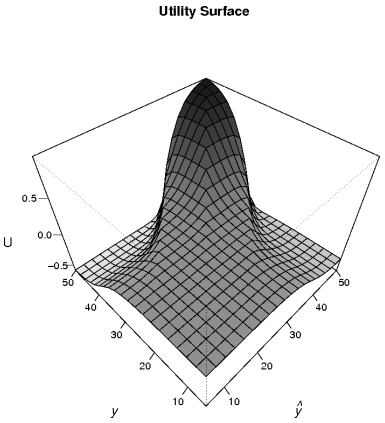

Figure 4. An utility surface

In this figure we have set $p = 0.6$, making missed events slightly more serious than false alarms. The surface reflects the application preference bias on rare high extreme values prediction. Near the diagonal (smaller errors) we have a positive utility which grows fast as we reach higher values of both predicted and true values. On the contrary, corners of the surface are the more serious errors and, as such, lead to costs (i.e. negative utility). Given the p value specified for this example, the higher costs are on the area of missed events, that is, a prediction of an irrelevant value for a true relevant value

3.3. An Utility-based Performance Measure

Once we have a fully specified utility surface, several performance measures can be derived, such as the average utility per test case $\bar{U} = \frac{1}{n} \cdot \sum_i^n U(\hat{y}_i, y_i)$.

Such metric is able to properly evaluate the performance of different models according to the applications preference biases in terms of different costs/benefits across the domain of the target variable. For instance, using a parametrization similar to the one resulting on the surface shown in Figure 4, we can evaluate the two artificial models presented in Figure 1, and obtain a correct ranking of the models as shown on Table 2.

Table 2. Estimated performance of $M1$ and $M2$.

	MAD	MSE	$\bar{U}$
M1	1.5978	4.7454	-0.1332
M2	1.5978	4.7454	0.1968

As it can be observed, contrary to the standard error metrics, with the $\bar{U}$ metric we get a model ranking that is according to what is expected given the application preference bias and, moreover, we get a proper feedback on the performance of model $M1$ that gets a negative value of average utility, which means that its predictions on average lead to a cost.

Moreover, an extensive set of comparative experiments have confirmed us the usefulness of the $\bar{U}$ metric in the context of cost-sensitive regression applications, in terms of being able to correctly rank a set of alternative models. These experiments have also shown that the use of standard error metrics can be misleading in the sense that models that are ranked top according to these metrics may actually be sub-optimal according to the application preference biases.

4. Conclusions and Future Work

In this paper we have presented a new utility-based framework that allows a reliable evaluation and comparison of regression models under a cost-sensitive context. We have illustrated the drawbacks of using standard error metrics and have shown the ability of our proposed evaluation metric to overcome these difficulties. The use of our proposal can provide better results in terms of model comparisons for cost-sensitive regression applications. Moreover, if this metric is plugged into a regression algorithm we can expect to obtain models that are better according to the applications preference biases. Based on this framework of utility scores, we are currently working on regression versions of precision-recall and cost curves for better model comparisons under different setups.

Acknowledgments

The work of the first author is supported by a PhD scholarship of the Portuguese government (SFRH/BD/1711/2004).

References

Christoffersen, P. F., & Diebold, F. X. (1996). Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics*, *11*, 561–571.

Crone, S., Lessmann, S., & Stahlbock, R. (2005). Utility based data mining for time series analysis - cost-sensitive learning for neural networks. *Proc. of the 1st International Workshop on Utility-Based Data Mining* (pp. 59–68).

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. *Proc. of the 5th International Conf. on Knowledge Discovery and Data Mining (KDD-99)* (pp. 155–164). ACM Press.

Elkan, C. (2001). The foundations of cost-sensitive learning. *Proc. of 7th International Joint Conf. of Artificial Intelligence (IJCAI'01)* (pp. 973–978).

Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. (1999). AdaCost: misclassification cost-sensitive boosting. *Proc. 16th International Conf. on Machine Learning* (pp. 97–105). Morgan Kaufmann.

Torgo, L., & Ribeiro, R. (2007). Utility-based regression. *Proc. of the 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007)* (pp. 597–604). Springer.

Zadrozny, B., Weiss, G., & Saar-Tsechansky, M. (Eds.). (2006). *Proc. of the 2nd international workshop on utility-based data mining.*