Utility-based Regression

Luis Torgo[1,2] and Rita Ribeiro[1,3]

[1] LIAAD-INESC Porto LA, R. Ceuta, 118, 6., 4050-190 Porto, Portugal
[2] FEP, University of Porto, R. Dr. Roberto Frias, 4200-464 Porto, Portugal
[3] FC, University of Porto, R. Campo Alegre, 1021/1055, 4169-007 Porto, Portugal
[ltorgo,rita]@liacc.up.pt

Abstract. Cost-sensitive learning is a key technique for addressing many real world data mining applications. Most existing research has been focused on classification problems. In this paper we propose a framework for evaluating regression models in applications with non-uniform costs and benefits across the domain of the continuous target variable. Namely, we describe two metrics for asserting the costs and benefits of the predictions of any model given a set of test cases. We illustrate the use of our metrics in the context of a specific type of applications where non-uniform costs are required: the prediction of rare extreme values of a continuous target variable. Our experiments provide clear evidence of the utility of the proposed framework for evaluating the merits of any model in this class of regression domains.

1 Introduction

In many real world applications the costs and benefits of using prediction models are non-uniform. These observations have motivated the work on cost-sensitive learning (e.g. [5]) and more generally on utility-based mining [9, 11]. In the context of applying the discovered knowledge under a non-uniform cost setup, most works have focused on classification tasks (e.g. [3–6]). Still, within numeric prediction problems, also know as regression, similar problems arise. As mentioned by Crone et. al. [2] most works on regression assume uniform costs and use some form of average error statistic. In this context, several authors (e.g. [1, 2]) have proposed new cost of error functions that try to address these issues. However, most of these works only consider one particular type of non-uniform costs of errors: the difference between under- and over-predictions, i.e. situations where the predicted values are above or below the true values, respectively.

This paper proposes a framework for evaluating regression models in the context of arbitrarily shaped costs and benefits across the domain of the numeric target variable of regression tasks. We propose two new evaluation metrics that incorporate the notions of costs and benefits and thus are able to provide better feedback on the merits of regression models in the context of the specific biases of any numeric prediction task. These metrics use cost and benefit surfaces that we also formalize, which can be regarded as continuous versions of the well-know notion of misclassification cost matrices. We illustrate the use of our proposed metrics in a particular class of non-uniform costs/benefits application: the prediction of rare extreme values of a continuous variable.

2 Problem Formulation

Predictive learning tries to obtain an approximation of an unknown function $f : \chi \rightarrow \gamma$, based on a training data set D drawn from a distribution with domain $\chi \times \gamma$, where χ is the domain of the set of predictor variables and γ is either a discrete domain in the case of classification tasks, or $\Re$ in the case of regression. The obtained approximation, $\hat{f}_\beta$, is a model with a set of parameters, β, that are obtained by optimizing some preference criterion. For classification, this is usually the error rate, while in the case of regression the most frequent are the mean squared error, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, or the mean absolute deviation, $MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$.

Many authors (e.g. [3, 5]) have noticed the problems arising from the uniform cost assumption of the error rate evaluation criterion, which is unacceptable for many real world domains. The cost matrix formulation overcomes these limitations by allowing the specification of the cost of misclassifying class i by class j, and leads to the criterion of expected cost minimization, $\frac{1}{n}\sum_{i=1}^{n}C(\hat{y}, y)$, where $C(\hat{y}, y)$ is an entry on the pre-specified cost matrix. Regards regression few authors have addressed the issue of differentiated costs. Most of the existing works on having non-uniform costs for regression have been addressing the issue of differentiating the cost of under-predictions ($\hat{y} < y$), from the cost of over-predictions ($\hat{y} > y$) (e.g. [1, 2]). Although these approaches address several important application-specific requirements, they fail to provide a means to specify a cost function across all domain of the target variable, which was shown to be of key importance for this type of applications [8]. In this paper we address this issue by associating to each prediction a cost that is dependent on an user-defined relevance of both the true and predicted values.

3 Utility-Based Regression

As mentioned by Zadrozny [10], research on cost-sensitive learning has traditionally been formalized in terms of costs as opposed to benefits or rewards. However, evaluating a model in terms of benefits is generally preferable because there is a natural baseline from which to measure all benefits whether positive (real benefits of a prediction) or negative (that are in effect costs) [5]. Our proposal follows these lines, by measuring the utility of a regression model through the total balance between the costs and benefits originated by its predictions.

3.1 Relevance Functions

We assume that for some applications the relevance (importance) of the values of the target variable is not uniform across its domain. This domain-dependent information shall be provided through the specification of a relevance function, $\phi(Y) : \Re \rightarrow 0..1$, that maps the domain of the target variable into a 0..1 scale of relevance, where 1 represents maximum relevance. Our proposal is independent of the shape of the $\phi()$ function. We assume this function is specified by the user using his/her domain knowledge. The specification of the relevance function is the step of our proposal that is most challenging for the user. Given the large range of applications where the relevance of the target variable is non-uniform, it is virtually impossible to describe reasonable default relevance functions for all these applications. Still, in many applications relevance is often associated with rarity (e.g. highly profitable customers; high variations on stock prices; extreme weather conditions, etc.). For these applications relevance can be defined as a function that is inversely proportional to the probability density function (*pdf*) of the target variable. Although obtaining the functional form of these *pdf*'s is generally non trivial, reasonable approximations based on the available data sample can be obtained with techniques like kernel density approximators. In Section 4 we propose an even simpler strategy to derive a relevance function for a class of applications where relevance is associated with rarity: the prediction of rare extreme values of a numeric variable.

3.2 Cost and Benefit Surfaces

Generally, the cost of a prediction depends not only on the relevance of the test case value but also on the relevance of the predicted value. In effect, all three following situations are penalizing in a cost-sensitive application:

1. Predict a relevant value for an irrelevant test case (false alarm);
2. Predict an irrelevant value for a relevant test case (opportunity cost);
3. Predict a relevant but very different value for a relevant test case (the most serious mistakes: confusing relevant events).

We capture this notion of relevance of the prediction for a given test case by means of the definition of a bi-variate relevance function, $\Phi(\hat{Y}, Y)$, that depends on the relevance of both the true and predicted values,

$$\Phi(\hat{Y}, Y) = (1 - m) \cdot \phi(\hat{Y}) + m \cdot \phi(Y) \tag{1}$$

This function is a weighted average of the individual relevances of $\hat{Y}$ and Y. It is maximum when both are highly relevant and these are the cases where the cost of the predictions may reach the maximum if they are not accurate enough. The m parameter ($0 \leq m \leq 1$) differentiates between situations 1 (false alarms) and 2 (opportunity costs). Setting $m > 0.5$ makes the latter more important.

The cost of a prediction should also depend on its precision, i.e. how near are $\hat{Y}$ and Y from each other. Moreover, it should also be possible for the user to establish some kind of application-specific measure of cost in whatever units make sense for the domain. In this context, we define the cost of a prediction as,

$$c(\hat{Y}, Y) = \Phi(\hat{Y}, Y) \times C_{\max} \times L(\hat{Y}, Y) \tag{2}$$

where $C_{\max}$ is the maximum cost that is only assigned when the relevance of the prediction is maximum (i.e. $\Phi(\hat{Y}, Y) = 1$); and $L(\hat{Y}, Y)$ is a loss function that measures the prediction error.

The term $\Phi(\hat{Y}, Y) \times C_{\max}$ can be seen as a kind of case-specific maximum cost value. This is the maximum penalty we get if $\hat{Y}$ is the "worst possible" prediction for the test case under consideration. With respect to the loss function we could use any metric function, e.g. the absolute deviation $|\hat{Y} - Y|$. However, in order to make the meaning of the value $c(\hat{Y}, Y)$ more intuitive, we recommend the use of a percentage-type loss function that ranges from 0 to 1. Such function will then represent the proportion of the case-specific maximum cost we get due to our prediction. For maximum error ($L(\hat{Y}, Y) = 1$) we get the full penalty of the particular test case ($\Phi(\hat{Y}, Y) \times C_{\max}$), while a perfect prediction ($L(\hat{Y}, Y) = 0$) would entail no cost as expected. This means the value of $c(\hat{Y}, Y)$ will be expressed in the same units as $C_{\max}$, which is provided by the user, and thus it is more intuitive for him/her. In this context, we propose the following loss function that ranges from 0 to 1:

$$L(\hat{Y}, Y) = |\max_{i \in \hat{Y}..Y} \phi(i) - \min_{i \in \hat{Y}..Y} \phi(i)| \tag{3}$$

The use of the maximum and minimum functions is due to the fact that we want to let the user specify any arbitrarily shaped $\phi()$ function. This means that we can have two quite different Y values with the same value of $\phi()$, which would look like a perfect prediction if we had used the difference of relevances directly. However, these cases are exactly the most serious mistakes we want to avoid (the 3rd case on the list presented before). With our proposal, if both values have high relevance but are quite different then surely there will be values in between with lower relevance and this will result in a higher value of the loss function.

The function $c()$ can be seen as a continuous version of cost matrices, i.e. a cost surface. The total cost of the predictions of a model is defined as,

$$TC = \sum_{i=1}^{n} c(\hat{y}_i, y_i) \tag{4}$$

Our proposal also considers the benefits of the predictions of a model, with the goal of asserting its ability to accurately predict most of the relevant values in a test set. In the case of benefits it is only the relevance of the true value that counts, i.e. we are interested in asserting how well a model predicts the relevant test cases. In this context, the benefit surface is defined as,

$$b(\hat{Y}, Y) = \phi(Y) \times B_{\max} \times (1 - L(\hat{y}_i, y_i)) \tag{5}$$

where $B_{\max}$ is a user-defined maximum reward that is measured in the same units as the $C_{\max}$ constant; and $L()$ is a loss function as before.

Our definition of benefits associates higher rewards with higher relevance. The term $\phi(Y) \times B_{\max}$ calculates the case-specific benefit, while the last term is the proportion of this reward we get. The total benefits are given by,

$$TB = \sum_{i=1}^{n} b(\hat{y}_i, y_i) \tag{6}$$

Finally, we can define the utility of the predictions of a model as the net balance between its total costs and benefits,

$$U = TB - TC \tag{7}$$

4 An Illustrative Application

Modeling extreme data is very important in several application domains, like finance, meteorology, ecology, etc.. Several of these applications involve predicting a continuous variable. For these domains the extreme (high or low) values of the target variable are much more important than the others. Moreover, these extremes are generally quite rare, which turns this into a very hard prediction problem with very clear non-uniform costs and benefits of predictions. In this section we illustrate the use of our proposed framework for utility-based regression, by using it to compare quite diverse modelling techniques on a real world data set where the prediction of rare extreme values is of primary importance.

The application we use to illustrate our proposal concerns stock market forecasting. Namely, the data are about the task of trying to predict the future daily variation in closing prices of the IBM stock, using information regarding the values of these variations on the 10 previous market sessions. The data set consists of information on 8166 daily market sessions (roughly 30 years), each being described by 10 predictor variables (the variations on the 10 previous days) and a target variable (the variation on the next day). This application is a very clear example of non-uniform costs (and benefits) of predictions. In effect, any model that is extremely accurate at predicting small price variations (the most common) is essentially useless for a trader. Profitable trading is based on being able to capitalize on large price changes. Trades carried out over small price changes are usually not able to cover the trading costs and thus are non-profitable or even represent a loss of money. As such, in these applications the accuracy on the relevant (i.e. extreme high or low) changes of prices is the key criterion.

In order to apply our evaluation method we need to specify a relevance function for this domain. In this class of applications relevance is strongly associated with extreme and rare values of the target variable. The distribution of the target variable has a normal-like shape with very marked tails (the rare extreme price variations). From the description of the goals of this application it should be clear that the relevance function should have a shape that is inverse of the *pdf* of the price variations. We propose to use a sigmoid-like function for establishing a smooth relevance function. In order to define this function we use some of the statistics provided by boxplots that summarize the distribution of the target variable. With this strategy we are able to obtain a relevance function without having to deal with computationally complex approximations of the unknown

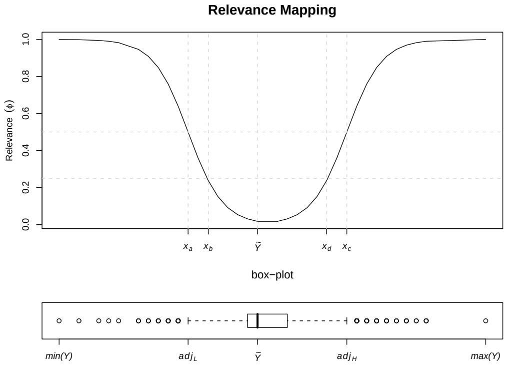

Fig. 1. A sigmoid-based relevance function for rare extreme values prediction.

pdf. Figure 1 provides a graphical illustration of the quantities involved in the derivation of the relevance function we use. This figure shows the box plot of an arbitrary normal-like distribution and the respective sigmoid-based relevance function. The relevance function is defined using distribution properties of the target variable ($\min(Y)$, $adj_L(Y)$, $\tilde{Y}$, $adj_H(Y)$ and $\max(Y)$) that can be easily estimated from the available data sample. This approach can be generally applied on problems where the target variable has a normal-like shape and where relevance is associated to rare extremes.

The other three parameters necessary to use our U metric are $C_{\max}$, $B_{\max}$ and m. Given our absence of domain expertise on stock market trading we have set these parameters using what seemed to us reasonable settings. Namely, we decided that the maximum benefit should be clearly higher than the maximum cost to try to reward proactive models. In our experiments we have used $C_{\max} = 10$ and $B_{\max} = 20$. With respect to the m parameter we have set it to 0.5, i.e. equal importance to false alarms and opportunity costs.

In order to test our proposed metrics under different experimental setups we have applied 3 quite different modeling techniques to the IBM data set. Namely, regression trees, neural networks and support vector regression. For all 3 methods we have used their implementations freely available on the R software environment [7], more specifically the function `rpart()` of the package `rpart`, the function `nnet()` of the package `nnet` and the function `svm()` of the package `e1071`. All 3 methods were used without any extensive parameter tuning as the goal was not to achieve the best possible accuracy but instead to test an evaluation metric under different setups.

All models were evaluated using the MAD, U, TC and TB statistics, which were described in Sections 2 and 3. The MAD statistic was selected as a "representative" of a standard evaluation metric. The values of all statistics were estimated using a 10-fold cross validation process. Statistical significance (95% level) of the differences when compared to the best ranked model were asserted by means of the non-parametric Wilcox test and signaled by "*". The results are shown on Table 1. For each statistic, we provide the ranking of the models and indicate the median and inter-quartile value measured over the 10 repetitions.

Table 1. The results/rankings on the IBM data set.

MAD		**U**		**TB**		**TC**	
dummy		dummy		svm		dummy	
0.01205	(0.00038)	108.18427	(37.76051)	278.41357	(27.94408)	134.12127	(17.29376)
cart	*	rand.forest		dummy	*	cart	*
0.01206	(0.00038)	108.18248	(36.24961)	243.10776	(36.9331)	134.12645	(17.29339)
nnet	*	cart	*	cart	*	nnet	*
0.01209	(0.00042)	108.1688	(37.75762)	243.09674	(36.92891)	134.12665	(17.29379)
rand.forest	*	nnet	*	nnet	*	rand.forest	*
0.01235	(0.00029)	108.1674	(37.75963)	243.09523	(36.93183)	134.14339	(13.72446)
svm	*	svm	*	rand.forest	*	svm	*
0.01447	(0.00047)	21.07661	(30.65326)	242.07468	(36.97128)	248.66908	(32.11022)

The goal of these experiments is not to check if the models are good according to our U metric, as they were obtained optimizing other criteria. Our objective is to check whether by using a metric tunned for giving more weight to rare extreme values, we can spot a method that is better at predicting these cases, particularly if that would not be found by using only standard statistics in the comparitive study.

The results on Table 1 unveil some interesting information that could not be observed from looking at the MAD scores. In effect, we can see that the SVM achieves a much higher score in terms of benefits, clearly indicating that it is able to capture more extreme values. However, this approach also has led to a higher value of TC, resulting of its more risky approach to this prediction problem. This results in a poor score in terms of net balance (U score). Still, given the fact that there was no particular tuning of the model parameters, we can say that the SVM is probably a model where more time should be invested in the context of this application, so that the signals it is producing get more precise. This sort of information is only available due to the use of an evaluation metric that is tunned towards the application goals.

5 Conclusions

This paper has described a new evaluation framework for regression tasks with non-uniform costs and benefits of the predictions. Our proposal is based on the specification of a relevance function over the domain of the target continuous variable. This function is the basis of the definitions of cost and benefit surfaces that can be regarded as continuous versions of cost/benefit matrices used in classification tasks. The use of the relevance function relieves the user from the heavy burden of having to specify a cost (and benefit) for all points in the bi-dimensional space of the predicted and true target values. The total cost and benefit of the predictions of a model provide, either individually or aggregated on an utility measure, important insights on the predictive performance of a model. Moreover, these insights are related to the application goals in terms of what is really relevant.

We have illustrated the use of our evaluation framework in the context of a particular class of applications: the prediction of rare extreme values of a continuous variable. Namely, we have used a data set from stock market prediction and have introduced a general relevance function for rare extremes prediction tasks. The results of our experiments have confirmed that our proposed metric provides a better insight on the ability of the models to accurately predict the cases that are more important for this class of applications.

Acknowledgments

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

References

1. P. Christoffersen and F. Diebold. Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics*, 11:561–571, 1996.
2. S. Crone, S. Lessmann, and R. Stahlbock. Utility based data mining for time series analysis - cost-sensitive learning for neural networks. In G. Weiss, M. Saar-Tsechansky, and B. Zadrozny, editors, *Proceedings of the 1st International Workshop on Utility-Based Data Mining*, pages 59–68, 2005.
3. P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 155–164. ACM Press, 1999.
4. C. Drummond and R. Holte. Exploiting the cost of (in)sensitivity of decision tree splitting criteria. In *Proc. 17th International Conf. on Machine Learning*, pages 239–246. Morgan Kaufmann, San Francisco, CA, 2000.
5. C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of 7th IJCAI'01*, pages 973–978, 2001.
6. W. Fan, S. Stolfo, J. Zhang, and P. Chan. AdaCost: misclassification cost-sensitive boosting. In *Proc. 16th International Conf. on Machine Learning*, pages 97–105. Morgan Kaufmann, San Francisco, CA, 1999.
7. R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2006. ISBN 3-900051-07-0.
8. L. Torgo. Regression error characteristic surfaces. In R. Grossman, R. Bayardo, K. Bennett, and J. Vaidya, editors, *Proc. of the 11th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, pages 697–702. ACM Press, 2005.
9. G. Weiss, M. Saar-Tsechansky, and B. Zadrozny, editors. *Proceedings of the 1st International Workshop on Utility-Based Data Mining*, 2005.
10. B. Zadrozny. One-benefit leaning: Cost-sensitive learning with restricted cost information. In *1st Intern. Work. on Utility-Based Data Mining*, pages 53–58, 2005.
11. B. Zadrozny, G. Weiss, and M. Saar-Tsechansky, editors. *Proceedings of the 2nd International Workshop on Utility-Based Data Mining*, 2006.