2012 IEEE 12th International Conference on Data Mining Workshops

Towards Utility Maximization in Regression

Rita P. Ribeiro
LIAAD - INESC TEC / DCC
Faculdade de Ciências - Universidade do Porto
Porto, Portugal
rpribeiro@dcc.fc.up.pt

***Abstract*—Utility-based learning is a key technique for addressing many real world data mining applications, where the costs/benefits are not uniform across the domain of the target variable. Still, most of the existing research has been focused on classification problems. In this paper we address a related problem. There are many relevant domains (e.g. ecological, meteorological, finance) where decisions are based on the forecast of a numeric quantity (i.e. the result of a regression model). The goal of the work on this paper is to present an evaluation framework for applications where the numeric outcome of a regression model may lead to different costs/benefits as a consequence of the actions it entails. The new metric provides a more informed estimate of the utility of any regression model, given the application-specific preference biases, and hence makes more reliable the comparison and selection between alternative regression models. We illustrate the objective of our evaluation methodology on a real-life application and also carry a set of experiments over a subset of our target regression tasks: the prediction of rare and extreme values. Results show the effectiveness of our proposed utility metric for identifying the models that perform better on this type of applications.**

***Keywords*-Cost-sensitive learning, regression, utility-based performance estimate.**

I. Introduction

Many real-world applications are related to cost-sensitive decision problems, where different predictions lead to different decisions involving costs/benefits (e.g. credit approval, insurance contracts, targeted marketing). In effect, these applications have motivated the work on cost-sensitive learning (e.g. [1], [2]) and, more recently, on utility-based mining (e.g. [3]). However, most of these studies focus on classification tasks even though similar problems arise within regression tasks. This is the case of event-based applications like ecological/meteorological catastrophes, fraud detection, etc. Some ranges of values of the target numeric variable are frequently associated with critical phenomena, which usually triggers some sort of alarm or action. Thus, predictions made for this subset of values should have a differentiated cost/benefit to be in accordance with the application goals.

The majority of the work in regression assumes uniform costs [4]. The estimated performance of a regression model is given by an average statistic over the magnitude of all the prediction errors that are given equal importance. To overcome the limitations of this assumption several authors (e.g. [4], [5]) have proposed new loss functions for regression. Nevertheless, the proposed functions are only capable of distinguishing under-predictions from over-predictions, i.e. situations where the predicted values are below or above the true values, respectively. Nevertheless, these loss functions, as well as other alternative evaluation measures, do not fully address our target applications, which are described by regression tasks with non-uniform costs/benefits.

In this context, our goal is to address two main research challenges: (i) express the preference bias of the users of these applications; (ii) provide a reliable evaluation/comparison/selection of models in this scenario.

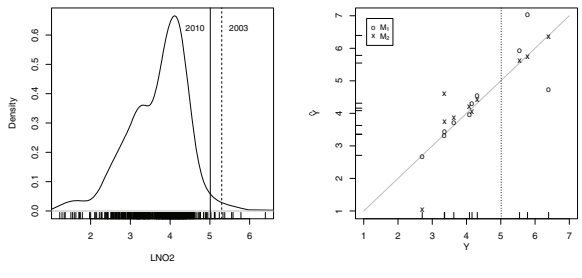
This paper is organized as follows: in Section II we introduce the problem of non-uniform costs in regression; in Section III, we explain how different benefits/costs can be embedded into a regression task; in Section IV we provide an illustration of an utility surface for real-world regression problem; in Section V we carry some experiments with the proposed utility-based framework; and finish with the main conclusions in Section VI.

II. Non-uniform Costs in Regression

Our goal is to address regression tasks, i.e. learn models that are able to predict a continuous target variable Y based on the values of predictor variables $X_1, X_2, \ldots, X_p$. Our target regression tasks differ from standard regression because not all values of the target variable are equally relevant, as for some values it is more important to be accurate, while for other values the accuracy may be completely irrelevant. Our objective is to handle such numeric prediction tasks with differentiated costs/benefits of predictions.

Standard error measures, such as the Mean Squared Error, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, or the Mean Absolute Deviation, $MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, are not suitable for this type of regression tasks. The reasoning is that all standard error estimates are based on additive measures of error and have as objective to find the model that minimizes the expected error within a uniform cost error framework. Nevertheless, there are real-world applications where different ranges of the target variable have different importance attached and hence so should have the errors committed by the predictions made at those ranges, as we will illustrate next.

978-0-7695-4925-5/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDMW.2012.82



(a) (b)

Figure 1. Two sets of predictions for outdoor air pollution registered values, given by the log-transformed hourly concentration of NO_2

Keeping the concentration of the pollutants at admissible levels has been a major concern to the World Health Organization (WHO). To reduce the emission of the pollutants, the air quality standards have been changing over the recent years making the annual average concentrations stricter. Nitrogen dioxide (NO_2) is an important atmospheric gas, not only because of its bad health effects but also because it absorbs visible solar radiation and contributes to the constitution of smog. In this context, it can also have a potential role in global climate change if its concentration becomes too high. Therefore, the control of concentration values of NO_2 is a very important task. In Figure 1a we have the probability density function (pdf) of log-transformed concentration values of NO_2 measured in $\mu g/m^3$ on each hour, at Alnabru in Oslo, Norway, between October 2001 and August 2003 [1]. In the same figure we also have the indication of the hourly average concentration value to achieve in 2010 according to WHO directives. The higher concentration values, are the most harmful to the public health and, as so, the most important from the application perspective. Nevertheless, these values are also the less frequent ones and thus the harder to predict with standard regression techniques. To better illustrate this fact, we have produced two different sets of predictions, M_1 and M_2, for `LNO2` values (cf. Figure 1b). They both have the same MSE (0.460) and MAD (0.402). Still, from the application perspective they should considerably different. In this context, it is perceptible that the *relevance* of the values of the target variable `LNO2` is *not uniform*. In effect, both M_2 and M_1 have exactly the same error amplitudes but these occur in different ranges of the target variable, M_2 having the smaller errors on the most important cases. A cost-sensitive classification approach (e.g. [2]) could have been taken if the interest was only to make accurate predictions between what is a harmful concentration level and an acceptable concentration level. This approach, would have the advantage of considering the location/class of true and predicted value and thus assign each prediction a cost or a benefit. Still, this constitutes a discrete approach to inherent continuous problem, and hence there will be always loss of information. In regression, other alternatives could have been taken to address this evaluation issue namely, assigning higher powers to the errors, using case weights, asymmetric loss functions [4]. Nevertheless, none of these hypothesis are sensitive to the concrete values involved in the predictions as they are, mostly, focus on the error amplitudes. This may lead to some counter-intuitive indications concerning the comparison of models for regression problems where the application preference bias is not uniform across the domain of the continuous target variable.

[1]Available on the StatLib Datasets Archive: `http://lib.stat.cmu.edu/datasets/`

III. Utility in Regression

Research on cost-sensitive learning has traditionally been formalized in terms of costs as opposed to benefits or rewards. However, according to Elkan [2], evaluating a model in terms of benefits is generally preferable because there is a natural baseline from which to measure all benefits whether positive (as real benefits of predictions) or negative (as costs of the predictions). Some studies (e.g. [3], [6]) also refer that performance measures should adopt the 'business' objectives. According to these studies, driving the data-mining process by these objectives is determinant to achieve potentially useful results.

In standard regression setups, the usefulness of a prediction is inversely proportional to its loss value, and thus to the difference between true and predicted values. In this context, one can define that in standard regression scenarios, utility is a function of the error of the prediction. In this scenario, having L as a loss function and U as an utility function, one can derive two properties of utility in standard regression:

(a) equal accuracy predictions have the same utility, i.e. $L(\hat{y}_1, y_1) = L(\hat{y}_2, y_2) \implies U(\hat{y}_1, y_1) = U(\hat{y}_2, y_2)$;
(b) more accurate predictions are always preferable, i.e. $L(\hat{y}_1, y_1) < L(\hat{y}_2, y_2) \implies U(\hat{y}_1, y_1) > U(\hat{y}_2, y_2)$.

For the applications we target with our proposal, these otherwise reasonable properties, can be counter-intuitive. According to the general framework of utility-based learning as proposed by Elkan [2] and Zadrozny [7], the utility of a prediction is the net balance between its benefits and costs (i.e. negative benefits). Hence, in non-uniform benefits/costs scenarios, utility is a function of both the error of the prediction (i.e. $L(\hat{y}, y)$) and the relevance (importance) of both $\hat{y}$ and y. In this context, the properties of utility in standard regression no longer hold. This means that, considering again L as some loss function and U as an utility function the following cases can occur:

(a) equal accuracy predictions do not have always equal usefulness, that is, $L(\hat{y}_1, y_1) = L(\hat{y}_2, y_2) \wedge U(\hat{y}_1, y_1) \neq U(\hat{y}_2, y_2)$;
(b) more accurate predictions are not always preferable, that is, $L(\hat{y}_1, y_1) < L(\hat{y}_2, y_2) \wedge U(\hat{y}_1, y_1) \leq U(\hat{y}_2, y_2)$.

If we take the Outdoor Air Pollution prediction problem presented in the previous section, we see that the notion of usefulness in standard regression metrics, no longer applies. For instance, with the set of predictions M_1 we get $L(2.68, 2.71) = 0.03$ and with the set of predictions M_2 we get $L(6.37, 6.40) = 0.03$. Still, from the perspective of this application, the usefulness of the first prediction should be much lower (i.e. $U(2.68, 2.71) \ll U(6.37, 6.40)$). In effect, the second situation forecasts a high NO_2 concentration value that will probably trigger some alarm. On the contrary, the first prediction forecasts an average NO_2 concentration value and, in this sense, we can say that it is much less useful. This small example contradicts the property of equal accuracy predictions being equivalent as established in standard regression Another example can be found for M_1 where $L(2.68, 2.71) < L(4.72, 6.40)$ but, from the application perspective, we should have $U(2.69, 2.71) < U(4.72, 6.40)$.

A. Relevance of Target Values

Relevance is the crucial property that distinguishes non-uniform benefit/cost regression problems from standard regression problems. For our target applications, not all values of the target variable are equally relevant for the user. There may exist a range of so-called relevant values where it is particularly useful to be accurate. It is the relevance of the target variable that expresses the domain-specific biases concerning the different importance of the values.

We propose a continuous *relevance function* $\phi\colon \mathcal{Y} \rightarrow [0, 1]$ to express the application-specific bias concerning the target variable domain $\mathcal{Y}$ by mapping it into a [0,1] scale of relevance, where 0 represents the minimum and 1 represents the maximum relevance.

As the objective of the relevance function is to represent domain knowledge, we do not impose any particular shape to the ϕ function. We assume that the specification is provided by the end-user. Still, specifying such function in an analytical way may not always be easy. It is also virtually impossible to describe reasonable default relevance functions for all non-uniform utility applications. Our proposal is to use a smooth interpolation method, such as piecewise cubic Hermite interpolation [8], at some user-defined points to define the relevance function. There will be regions of target variable where the end-user emphasizes the relevance compared to neighbouring regions. These way it expresses his interest on these regions, which might be associated with some key actions/decisions regarding the target application. At these ranges, the relevance function exhibits the shape of a bump. Bumps correspond, in fact, to intervals of the target variable where the relevance function is quasiconcave and this leads to the notion of *bumps of relevance*.

Quasiconcave functions are widely used in many research fields due to some of their interesting properties. Quasiconcave functions are characterized as *single-peaked* functions like the probability density function of distributions such as uniform, normal, exponential, logistic, *Weibull*, *Gamma*, among other distributions. Nevertheless, we should stress that our goal is to address applications where the relevance of the target variable may not be uniform, and, in particular, may not be *single-peaked*. Still, in many applications, relevance is often associated with specific ranges or with extremeness and rarity (e.g. highly profitable customers; high variations on stock prices; extreme weather conditions). Our proposal is to locally address these ranges as a way to achieve the global applications' objectives. Our ultimate goal is to maximize utility, and that can only be achieved by minimizing the error on the most relevant values. Some of the properties of quasiconcave functions are at the partial fulfillment of this goal, namely, the Local-Global Property of the Maximum [9]. This property ensures the existence of an unique maximum for these shape-like functions. This allow us to give a more rigorous definition of bump of relevance.

Let $\mathcal{Y} \subseteq \Re$ be the domain of a continuous target variable and ϕ be the continuous relevance function associated to Y. A bump of relevance $\mathcal{B}_i$ is an interval of $\mathcal{Y}$ defined by $\mathcal{B}_i := \langle b_i^-, b_i^*, b_{i+1}^- \rangle$, with $b_i^- \leq b_i^* < b_{i+1}^-$ and $b_i^-, b_i^*, b_{i+1}^- \in \mathcal{Y}$, such that:

(a) $\phi : [b_i^-, b_{i+1}^-[\rightarrow [0, 1]$ is quasiconcave;

(b) b_i^* is the average value at which the target variable reaches the maximum relevance of the bump, that is $b_i^* = \overline{\mathcal{B}_M}$, where $\mathcal{B}_M = \left\{ y \in \mathcal{B}_i \mid \operatorname{argmax}_y \phi(y) \right\}$

(c) b_i^- is the average value at which the target variable attains the left minimum relevance, i.e. before it grows to its maximum, that is $b_i^- = \overline{\mathcal{B}_{m_L}}$, where $\mathcal{B}_{m_L} = \left\{ y \in \mathcal{B}_i \mid \operatorname{argmin}_{y \leq b_i^*} \phi(y) \right\}$

(d) b_{i+1}^- is the average value at which the target variable attains the right minimum relevance, i.e. after it grows to its maximum, that is $b_{i+1}^- = \overline{\mathcal{B}_{m_R}}$, where $\mathcal{B}_{m_R} = \left\{ y \in \mathcal{B}_i \mid \operatorname{argmin}_{y > b_i^*} \phi(y) \right\}$

From the above definition of bumps of relevance, we can state that the family of all bumps of relevance, defined across $\mathcal{Y}$ with respect to the relevance function ϕ, is a bumps partition of $\mathcal{Y}$, $\mathcal{P}_\phi(\mathcal{Y})$. Figure 2 illustrates an example of a bumps partition defined across the domain of the target variable Y.

B. From Prediction Errors to Utility Values

Our main motivation in the formulation of utility-based regression is to address applications where the target prediction variable has non-uniform relevance for the user. This non-uniform relevance usually results from the fact that predictions may be actionable. In such domains, taking the right action results in positive benefits, while taking a wrong action results in costs (i.e. negative benefits). As the set of possible actions is limited, this could lead to a typical classification setup. However, the other key distinction of our target applications is that we are interested in degrees of action and that is the key that takes us apart from

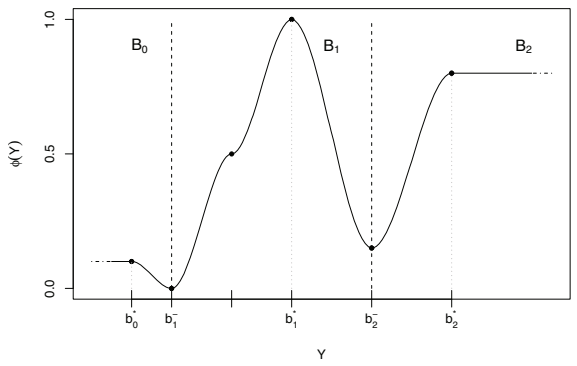


Figure 2. Identification of bumps of relevance in the target variable Y

classification approaches. Thus, in order to calculate the utility value of a prediction we must consider two aspects: (i) does the predicted value leads to the correct action/decision? - or equivalently, do y and $\hat{y}$ belong to the same bump? (ii) what is the accuracy of the predicted value? - or equivalently, what is the value of $L(\hat{y}, y)$?

First notice that each bump may have different sensitiveness to the prediction errors. In particular, "narrow" bumps are supposed to be more intolerant to prediction errors, while "wider" bumps are supposed to accept larger prediction errors. We established the maximum admissible loss in a bump $\mathcal{B}_i$ as the double of its smallest half-amplitude, which is given by the difference between each one of its bounds, b_i^- and b_{i+1}^-, and its maximum value b_i^*, i.e.

$$b_i^{\Delta} = 2 * \min\{|b_i^- - b_i^*|, |b_i^* - b_{i+1}^-|\} \tag{1}$$

The concrete value of utility results from the assessment of benefits and costs to each prediction, as we describe next.

The first rationale for assessing a benefit is that we are entailing a correct action. The resulting reward should then be dependent on the accuracy of the predicted value and on the relevance of the true value. In this sense, we define the benefit of a prediction to be a proportion of the relevance of the true value, i.e. $\phi(y)$, which constitutes its maximum value. The *benefit function* B_ϕ is defined as follows,

$$B_\phi(\hat{y}, y) = \phi(y)\,(1 - \Gamma_B(\hat{y}, y)) \tag{2}$$

where Γ_B is the *bounded loss function for benefits*. The bounded loss function Γ_B asserts the proportion of the maximum benefit a prediction should get. It should be 0 if we have a perfect prediction thus leading to the maximum benefits in the current situation, i.e. $\phi(y)$. The function should increase up to 1 as we move away from a perfect prediction or if we cross the boundaries of the bump to which y belongs.

Formally, the *bounded loss* of a prediction $\hat{y} \in \mathcal{Y}$ for a true value $y \in \mathcal{Y}$ is given by the following function:

$$\Gamma_B(\hat{y}, y) = \begin{cases} L(\hat{y}, y)/\dot{L}_B(\hat{y}, y), & \text{if } L(\hat{y}, y) < \dot{L}_B(\hat{y}, y) \\ 1, & \text{if } L(\hat{y}, y) \geq \dot{L}_B(\hat{y}, y). \end{cases} \tag{3}$$

where L is a "standard" loss function (e.g. absolute deviation) and $\dot{L}_B$ is a *benefit threshold function*.

The benefit threshold function determines when the predicted value stops leading to a benefit. This may occur due to two conditions: (i) overcoming the maximum admissible loss of the bump; or (ii) being on a different bump. The following function implements these criteria.

$$\dot{L}_B(\hat{y}, y) = \min\{b^{\Delta}_{\gamma(y)},\ \ddot{L}_B(\hat{y}, y)\} \tag{4}$$

where $b^{\Delta}_{\gamma(y)}$ is the maximum admissible loss established for the bump of y, i.e. the bump index $\gamma(y)$ [2] (cf. Equation 1), and $\ddot{L}_B$ is defined as follows,

$$\ddot{L}_B(\hat{y}, y) = \begin{cases} |y - b^-_{\gamma(y)}|, & \text{if } \hat{y} < y \\ |y - b^-_{\gamma(y)+1}|, & \text{if } \hat{y} \geq y. \end{cases} \tag{5}$$

Regards the behaviour of the proposed definition of benefits (cf. Equation 2) we have that when $L(\hat{y}, y) = 0$, $B_\phi(\hat{y}, y) = \phi(y)$. This property follows from the definition of the bounded loss used in the definition of benefits. Moreover, when the loss value increases, our benefits will reach zero as shown below.

Lemma 1: For sufficient large values of $L(\hat{y}, y)$, we have $B_\phi(\hat{y}, y) = 0$.

Proof: We want to prove that there exists an $t_L \in [0, 1]$, such that $\forall \hat{y}, y\ L(\hat{y}, y) \geq t_L$ we have $B_\phi(\hat{y}, y) = 0$. If we assign $t_L = \dot{L}_B(\hat{y}, y)$, then by the boundary loss function definition, we have $\Gamma_B(\hat{y}, y) = 1$ and hence $1 - \Gamma_B(\hat{y}, y) = 0$. Therefore, $B_\phi(\hat{y}, y) = 0$. ■

Regarding costs, the first rationale behind them is that we are making a too inaccurate prediction or/and entailing a wrong action. Besides these two facts, to really assess the cost of the made prediction, we must inspect how relevant is the impact of such wrong action. While the benefits of a prediction depend on the usefulness of its associated action (i.e. $\phi(y)$), costs depend not only on the action associated with the true value but also on the action of the predicted value. This means that costs should be proportional to the relevance of both the true and predicted values. The *joint relevance function* (cf. Equation 6) captures this notion by calculating a weighted average of these two factors.

$$\phi^p(\hat{y}, y) = (1 - p)\,\phi(\hat{y}) \ + \ p\,\phi(y) \tag{6}$$

where $p \in [0, 1]$ is a factor differentiating the types of errors. In this context of actionable predictions there are three different types of incorrect actions: (i) *false alarms* where the prediction leads the user to a relevant event/action

[2] $\gamma(y) = \left\{ i \mid y \in \mathcal{B}_i \wedge \mathcal{B}_i \subset \mathcal{P}_\phi(\mathcal{Y}) \right\}$

when the true value is rather irrelevant (i.e. we act when we should not); (ii) *missed opportunities* where the model predicts an irrelevant value but the true value is highly relevant (i.e. we did not do anything when we should have acted); or (iii) *confusing events* where the prediction leads to a wrong action (i.e. we ought to act but we carry out the wrong action). The third scenario is the most serious type of mistake.

We define the cost of a prediction to be a proportion of the joint relevance of both true and predicted values, which gives its maximum value. Hence, the *cost function* C^p_ϕ is defined as follows,

$$C^p_\phi(\hat{y}, y) = -\phi^p(\hat{y}, y)\,\Gamma_C(\hat{y}, y) \tag{7}$$

where ϕ^p is the joint relevance function and Γ_C is the bounded loss (cf. Equation 3) calculated using the *cost threshold function* given below.

The cost threshold function determines when the costs reach the maximum value. As with the benefit threshold function this may occur due to two conditions: (i) overcoming the maximum admissible loss of the bump; or (ii) predicting a value that has the maximum relevance of a different bump (i.e. a different action). The next function implements these conditions.

The cost threshold of a prediction $\hat{y} \in \mathcal{Y}$ for a true value $y \in \mathcal{Y}$, is given by the following function:

$$\dot{L}_C(\hat{y}, y) = \min\{b^\Delta_{\gamma(y)},\, \ddot{L}_C(\hat{y}, y)\} \tag{8}$$

where $b^\Delta_{\gamma(y)}$ is the maximum admissible loss established for the bump of y (cf. Equation 1) and $\ddot{L}_C$ is given by,

$$\ddot{L}_C(\hat{y}, y) = \begin{cases} |y - b^*_{\gamma(y)-1}|, & \text{if } \hat{y} < y \\ |y - b^*_{\gamma(y)+1}|, & \text{if } \hat{y} \geq y. \end{cases} \tag{9}$$

The first term in the $\min$ function is the maximum admissible error amplitude of the bump of the true value. The second term in the $\min$ function checks if the predicted value has reached the maximum relevance value of a neighbouring bump. Regarding the proposed definition of the cost of a prediction, we know that when $L(\hat{y}, y) = 0$, $C^p_\phi(\hat{y}, y) = 0$ because $\Gamma_C(\hat{y}, y) = 0$, but when the loss value increases our costs should be proportional to the relevance of the true and predicted value.

Lemma 2: For sufficient large values of $L(\hat{y}, y)$, we have $C^p_\phi(\hat{y}, y) = -\phi^p(\hat{y}, y)$.

Proof: We want to prove that there exists an $t_L \in [0, 1]$, such that $\forall \hat{y}, y\ L(\hat{y}, y) \geq t_L$ we have $C^p_\phi(\hat{y}, y) = -\phi^p(\hat{y}, y)$. If we assign $t_L = \dot{L}_B(\hat{y}, y)$, then by the boundary loss function definition, we have $\Gamma_C(\hat{y}, y) = 1$. Therefore, $C^p_\phi(\hat{y}, y) = -\phi^p(\hat{y}, y)$ ■

Having defined how to obtain the benefits and costs associated with any prediction we can now propose a method to calculate the utility of a prediction. Our proposal implements the following general principle: *"The utility of a prediction is the net balance between its benefits and costs."*

Definition 1: The *utility* of a prediction $\hat{y} \in \mathcal{Y}$ for a true value $y \in \mathcal{Y}$ is given by:

$$\begin{aligned} U^p_\phi(\hat{y}, y) &= B_\phi(\hat{y}, y) + C^p_\phi(\hat{y}, y) \\ &= \phi(y)\,(1 - \Gamma_B(\hat{y}, y)) - \phi^p(\hat{y}, y)\,\Gamma_C(\hat{y}, y) \end{aligned}$$

In the above setting, it is also possible to prove that for any given prediction $\hat{y}$ of a true value y there is only a bounded range of utility values to assign.

Theorem 1: Let $U^p_\phi \colon \mathcal{Y} \times \mathcal{Y} \to [-1, 1]$ be an utility function defined for a continuous target variable with domain $\mathcal{Y}$ and with a relevance function $\phi \colon \mathcal{Y} \to [0, 1]$. Then whatever is the predicted value $\hat{y} \in \mathcal{Y}$ for a true value $y \in \mathcal{Y}$, we have:

(a) the maximum of $U^p_\phi(\hat{y}, y)$ is $\phi(y)$;
(b) the minimum of $U^p_\phi(\hat{y}, y)$ is $p\,(1 - \phi(y)) - 1$.

Proof:

(a) According to our utility function formulation, for the maximum value of utility to be achieved it is necessary that the prediction matches the true value, i.e. $\hat{y} = y$. This means that the loss value is zero, i.e. $L(\hat{y}, y) = 0$. From the definition of bounded loss function with respect to benefits threshold function $\dot{L}_B$, we have that if $L(\hat{y}, y) = 0$ then $\Gamma_B(\hat{y}, y) = 0$. Similarly, regarding bounded loss function with respect to cost threshold function $\dot{L}_C$, we have that if $L(\hat{y}, y) = 0$ then $\Gamma_C(\hat{y}, y) = 0$. Thus, in these conditions and considering that the relevance function ϕ only takes values in $[0, 1]$, it can be shown that,

$$\begin{aligned} U^p_\phi(\hat{y}, y) &\leq U^p_\phi(y, y) \\ &= \phi(y)\,(1 - \Gamma_B(y, y)) - \phi^p(y, y)\,\Gamma_C(y, y) \\ &= \phi(y) \end{aligned} \tag{10}$$

(b) From the zero loss prediction, the utility value decreases according to benefits and cost criteria that rely on the action that is implied by the prediction and on the loss value. In this sense, the worst scenario is achieved when the benefit is zero and the cost is maximum. Lemma 1 and Lemma 2 have established this scenario. When loss tends to infinity, i.e. $L(\hat{y}, y) \to +\infty$, we have $B_\phi(\hat{y}, y) = 0$ and $C^p_\phi(\hat{y}, y) = -\phi^p(\hat{y}, y)$. In these conditions and considering that both the relevance function ϕ and the penalizing cost factor p only take values in $[0, 1]$, it can be shown that,

$$\begin{aligned} U^p_\phi(\hat{y}, y) &= B_\phi(\hat{y}, y) + C^p_\phi(\hat{y}, y) \\ &= \phi(y)\,(1 - \Gamma_B(\hat{y}, y)) - \phi^p(\hat{y}, y)\,\Gamma_C(\hat{y}, y) \\ &\geq -\phi^p(\hat{y}, y) \\ &\geq p\,(1 - \phi(y)) - 1 \end{aligned} \tag{11}$$

■

C. Relationship with Standard Regression

A standard regression problem can be recast as an utility-based regression problem. In effect, we can see standard regression as a problem where all values have equal and maximal relevance, i.e. $\phi(y) = 1, \forall y \in \Re$. Hence, we have a single bump. Provided that the relevance function is constant, we have no indications on the benefits or loss threshold to define the utility loss functions Γ_B and Γ_C. In such conditions, we can assign the maximum admissible loss of the only existing bump with a standard error estimate ε of the target variable. This estimate can be based on the *Gaussian level* of noise and on the number of observations. This means that only the predictions with a loss below this estimated standard error ε are subject to benefit/cost analysis. Moreover, there are no actions attached to the target variable domain. Thus, the maximum admissible loss is the only criterion for a prediction to be considered as a benefit or a cost, and it is the same for both of them. In these conditions, our utility function is reformulated as follows,

$$\begin{aligned} U^p_\phi(\hat{y}, y) &= \phi(y)\,(1 - \Gamma_B(\hat{y}, y)) - \phi^p(\hat{y}, y)\,\Gamma_C(\hat{y}, y) \\ &= 1 - \Gamma_B(\hat{y}, y) - \Gamma_C(\hat{y}, y) \\ &= 1 - \Gamma_\varepsilon(\hat{y}, y) - \Gamma_\varepsilon(\hat{y}, y) \\ &= 1 - 2\,\Gamma_\varepsilon(\hat{y}, y) \end{aligned} \tag{12}$$

With this utility function, a standard loss function L is scaled into a $[-1, 1]$ interval of utility values. The maximization of U is equivalent to the minimization of the *bounded loss function* Γ_ε, and thus the minimization of L, the goal of a standard regression task. Through this approach, any standard regression problem can be addressed in our utility-based regression framework.

IV. Illustrative Example of an Utility Surface

Suppose that following WHO directives, we have defined a set of control points, derived the relevance function ϕ and identified the respective bumps partition (cf. Figure 3a). In Figures 3b and 3c, we present the utility surface obtained for the NO_2 Emissions prediction problem, with the penalization costs factor $p = 0.5$. At the left, Figure 3b, we have the utility isometrics of the surface, i.e. the lines that share the same value of utility, while at the right, Figure 3c, we have a 3D representation of how the utility values evolve as a function of the true and predicted values. Through this surface, we obtain a better understanding regarding the costs and benefits associated to this application. Near the diagonal where true and predicted values are equal, we have a positive utility that grows fast as we reach higher values of both the predicted and true values (top right corner). These are the values with higher relevance. At the top left and bottom right corners we get costs, i.e. negative utility values. These areas correspond to inaccurate predictions leading to incorrect actions or predictions where the loss value goes beyond the maximum admissible loss (≈ 8.8). As the domain is characterized by having one action (bump) only, that is alarm high `LNO2` concentration values, and the admissible loss is large, costs never get too high for the range of values considered in the graphs. This explains why the utility surface is mainly positive.

V. Experimental Analysis

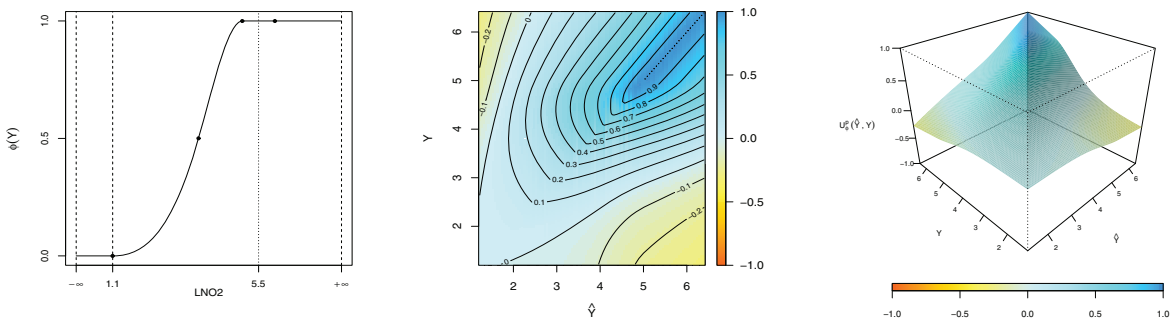
We have presented an utility-based evaluation methodology that is able to handle applications with non-uniform costs across the domain of a continuous target variable. In this section, we test the sensitiveness of such methodology in the task of identifying the best models for this kind of applications, when compared with a standard error metric. We claim that without our proposed utility-based metrics, we can only obtain suboptimal models comparisons in the context of our target applications. Through a set of experiments we will show that standard evaluation metrics are not always able to choose the best models.

Our experiments were conducted on the NO_2 concentration data set we have been using. For the experiments we used all the domain knowledge that we had. The used relevance function ϕ was the same shown in Figure 3a. Concerning the utility surface, we established $p = 0.75$ to make *opportunity costs* more serious than *false alarms*.

With the goal of avoiding any bias on our conclusions concerning the used modelling techniques, we have selected four quite different approaches. For all of these techniques we have used the implementations available on the R software environment [10]. The selected methods were:

- regression trees (cart) - we used the implementation available through package `DMwR` [11], with default setting `se=1`.
- support vector machines (svm) - we used the implementation available in the package `e1071` [12], width default setting `cost=1` and radial-basis kernel.
- multivariate adaptive regression splines (mars) - we used the implementation available in the package `earth` [13], with default setting `degree=1` and `thresh=0.001`.
- random forests (randomF) - we used the implementation available in `randomForest` package [14], with default setting `ntree=500` and `nodesize=5`.

No extensive parameter tuning was carried out, as top performance is not the goal here - our goal is to compare the model rankings produced by different evaluation metrics under different setups. None of the alternative models we are considering in our experiments optimizes any of our utility-based metrics. In this context, we incur the danger of concluding that all models perform equally bad in terms of the applications goals, which would not allow us to conclude anything concerning the eventual advantages of our metrics. To avoid this problem we wanted to ensure that among the candidate models there were alternatives that were clearly better in terms of being able to optimize



(a) (b) (c)

Figure 3. An utility surface for Outdoor Air Pollution

utility. These models should come up as the best models if the used metrics have any value. To obtain these "optimal models" we generated artificial sets of predictions based on the predictions of the "standard" models. Namely, we created these artificial predictions by tweaking the prediction errors of our original four modelling techniques in such a way that the errors are artificially re-allocated to the test cases that improve the overall utility score. By proceeding this way, we reach to a set of artificial set of predictions that have exactly the same set of error amplitudes as the base models, but with a higher utility value because the smallest errors were re-allocated to more relevant true values.

In our experiments we named these artificial sets of predictions improv.⟨*base_technique*⟩ (e.g. improv.cart). From the perspective of the goals of our target applications, these improved models should appear ahead of the base models (or at most at the same position in the unlikely event that the base model already achieves this "improved" performance). A failure to rank these improved models on top would mean the failure of our proposal. The function U^p_ϕ gives us a base to evaluate models for regression tasks with non-uniform benefits/costs across the target variable. Through it, we can estimate the expected utility of a model. Given a regression model, we can estimate its Mean Utility (MU) by,

$$MU = \frac{1}{n}\sum_{i=1}^{n} U^p_\phi(\hat{y}_i, y_i) \tag{13}$$

where ϕ is the relevance function and p is the penalizing costs factor. This metric reflects the model performance regarding the application-specific biases. Positive values indicate that the model is useful on average, while negative values indicate that the model usually issues predictions that represent costs. An equivalent metric is obtained if we map the values of MU to the interval $[0, 1]$, thus leading the Normalized Mean Utility (NMU) as follows,

$$NMU = \frac{\sum_{i=1}^{n} U^p_\phi(\hat{y}_i, y_i) + n}{2\,n} \tag{14}$$

where ϕ is the relevance function and p is the penalizing costs factor. For comparison purposes, we have evaluated the performance of the models by these utility-based metrics and by the $NMAD$ evaluation metric, selected as a "representative" of a standard evaluation metric that consists of a normalized version of MAD to the interval $[0, 1]$, as follows,

$$NMAD = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} |\widetilde{Y} - y_i|} \tag{15}$$

where $\widetilde{Y}$ is the median of the target variable Y. All the three evaluation metrics use, as loss function, the absolute deviation. Given that our data set has a small number of cases, we chose to run a stratified 1x10-fold cross validation process to obtain the performance estimates. The statistical significance of the difference between the score of each modelling technique and the one ranked as the best, was measured using the *Student paired t-test*. For each statistic we provide the mean (μ) and the standard deviation (σ) obtained from the cross validation process.

The results are presented in Table I. From a first analysis of the results, it is possible to notice that our utility-based metrics (MU and NMU) and $NMAD$ do not agree on the top ranked modelling technique. Moreover, as expected, the $NMAD$ statistic is not able to distinguish between the improved versions of the models and the original ones. According to $NMAD$ randomF is the best model for this problem. However, randomF is not a well positioned modelling technique according to both our utility metrics, with a performance that is significantly worse than the top model according to these two metrics. Still, according to MU, the performance of randomF is positive, which means that on average its predictions lead to a benefit. In effect, in this application all modelling techniques achieved a globally good utility score. This is explained by the overall positive utility surface of this problem (cf. Figure 3c). Independently of these observations, our experiments show that our metrics are able to identify the best models (which were artificially made the best) from the perspective of the user preference

Table I
UTILITY PERFORMANCE ESTIMATES FOR THE PREDICTION OF NO_2.

$MU : \mu \pm \sigma$	$NMU : \mu \pm \sigma$	$NMAD : \mu \pm \sigma$
improv.randomF	improv.randomF	randomF
0.51417 ± 0.03604	0.75709 ± 0.01802	0.61803 ± 0.07345
improv.svm	improv.svm	improv.randomF
0.51264 ± 0.03573	0.75632 ± 0.01787	0.61803 ± 0.07345
improv.mars	improv.mars	mars ∗
0.51222 ± 0.03617	0.75611 ± 0.01809	0.65046 ± 0.10225
improv.cart ∗∗	improv.cart ∗∗	improv.mars ∗
0.50609 ± 0.0358	0.75305 ± 0.0179	0.65046 ± 0.10225
svm ∗∗	svm ∗∗	svm ∗
0.48409 ± 0.03248	0.74205 ± 0.01624	0.66435 ± 0.05996
randomF ∗∗	randomF ∗∗	improv.svm ∗
0.48344 ± 0.03268	0.74172 ± 0.01634	0.66435 ± 0.05996
mars ∗∗	mars ∗∗	cart ∗∗
0.48009 ± 0.03302	0.74005 ± 0.01651	0.75887 ± 0.05341
cart ∗∗	cart ∗∗	improv.cart ∗∗
0.46883 ± 0.03303	0.73441 ± 0.01651	0.75887 ± 0.05341

biases for this application. Moreover, according to our metrics, the best model is improv.randomF. Still, more important than this observation is the fact that our metrics were able to rank at the top all improved variants of the base models, which provides strong evidence concerning their ability to identify models that perform better in terms of the goals of this application.

VI. CONCLUSION

In this paper we have argued for the existence of real-world data mining applications that represent regression tasks with non-uniform costs and/or benefits across the target variable. To properly evaluate the performance of regression models in this context, we have proposed an utility-based function that assesses benefits/costs to models predictions. For this purpose, we take into account the relevance of the true and predicted values, defined by a continuous function that maps the domain of the target variable into a $[0, 1]$ scale of relevance, and consider two different aspects: (i) the accuracy of the action associated with the predicted value; and (ii) the numeric accuracy of the predicted value. Through this utility-based evaluation framework, we are now able to cope with non-uniform benefits/costs in regression, likewise it happened in classification, and, at the same time, incorporate the standard error metrics (e.g. mean squared error) as special cases. Our experiments have confirmed the risks of using standard evaluation metrics when comparing models in applications with non-uniform costs. Namely, there are problems where standard metrics are unable to identify the best models, which can be critical for some real-world applications. Moreover, we have shown that these risks can be overcome by the use of our metric, which is able to provide a model ranking that is in accordance to the preference biases of the applications. Given the exploratory character of this paper, there are many interesting future perspectives. Namely, we believe that the integration of this metric into the learning phase of regression algorithms can lead to better models from the perspective of the application preference biases.

ACKNOWLEDGMENT

This work is funded by the ERDF - European Regional Development Fund through the COMPETE Programme.

REFERENCES

[1] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *KDD'99: Proc. of the 5th Int. Conf. on Knowledge Discovery and Data Mining*, 1999, pp. 155–164.

[2] C. Elkan, "The foundations of cost-sensitive learning," in *IJCAI'01: Proc. of 17th Int. Joint Conf. of Artificial Intelligence*, vol. 1, 2001, pp. 973–978.

[3] G. Weiss, B. Zadrozny, and M. Saar-Tsechansky, "Guest editorial: special issue on utility-based data mining," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, pp. 129–135, 2008, springer.

[4] S. Crone, S. Lessmann, and R. Stahlbock, "Utility based data mining for time series analysis - cost-sensitive learning for neural networks," in *Proc. of the 1st Int. Workshop on Utility-Based Data Mining*, 2005, pp. 59–68.

[5] P. F. Christoffersen and F. X. Diebold, "Further results on forecasting and model selection under asymmetric loss," *Journal of Applied Econometrics*, vol. 11, pp. 561–571, 1996.

[6] S. Daskalaki, I. Kopanas, and N. M. Avouris, "Evaluation of classifiers for an uneven class distribution problem," *Applied Artificial Intelligence*, vol. 20, no. 5, pp. 381–417, 2006.

[7] B. Zadrozny, "One-benefit learning: cost-sensitive learning with restricted cost information," in *Proc. of the 1st Int. Workshop on Utility-Based Data Mining*, 2005, pp. 53–58.

[8] R. L. Dougherty, A. Edelman, and J. M. Hyman, "Nonnegativity-, monotonicity-, or convexity-preserving cubic and quintic Hermite interpolation," *Mathematics of Computation*, vol. 52, no. 186, pp. 471–494, 1989.

[9] E. M. T. Hendrix and B. G. Toth, *Introduction to Nonlinear and Global Optimization*, 1st ed., ser. Springer Optimization and Its Applications. Springer-Verlag, 2010, vol. 37.

[10] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[11] L. Torgo, *Data Mining with R, learning with case studies*. Chapman and Hall/CRC, 2010. [Online]. Available: http://www.liaad.up.pt/ ltorgo/DataMiningWithR

[12] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, , and A. Weingessel, *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2010.

[13] S. Milborrow, T. Hastie, and R. Tibshirani., *earth: Multivariate Adaptive Regression Spline Models*, 2010.

[14] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News: The Newsletter of the R project*, vol. 2, no. 3, pp. 18–22, 2002.