Predicting Harmful Algae Blooms

No Author Given

No Institute Given

Abstract. In several applications the main interest resides in predicting rare and extreme values. This is the case of the prediction of harmful algae blooms. These rare phenomena consist of an unusual occurrence of certain algae in water samples. The occurrence of these blooms has a strong impact in river life forms and water quality and turns out to be a serious ecological problem. Being able to predict these blooms is of extreme importance. In this paper, we describe a data mining method whose main goal is to predict accurately this kind of rare extreme values, as well as to understand under which conditions these values occur. We propose a new splitting criterion for regression trees that enables the induction of trees achieving these goals. We carry out an analysis of the obtained results with our method on this application domain and compare them to those obtained with standard regression trees. We conclude that this new method achieves better results in terms of the evaluation statistics that are relevant for this kind of applications.

1 Introduction

In most prediction problems, the main interest resides in predicting accurately the most frequent cases. Nevertheless, there are some applications where it would be of major importance to predict rare situations. This is the case of problems like fraud detection, finding hackers in telecomunications and ecological catastrophes. Rare events are usually considered outliers from a statistical perspective. As such, in these application we are in effect targeting at modeling outliers. An intuitive definition of outliers was given by Hawkins [6]. He defines them as observations that deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism.

Sometimes, outliers are associated with extreme values. In this paper we focus exactly on predicting these rare and extreme values. Harmful algae blooms rivers are one of these prediction problems. Algae blooms consist in the occurrence of unusually high values of certain algae. Given that the target variable (the occurrence) is usually measured as a continuous value, we are facing a regression problem. However, the main difference to standard regression tasks is that our main interest is being accurate at the prediction of occurrences of rare high values of the target variable. Another similar real world application is the prediction of stock market returns, where small and highly frequent returns are irrelevant for investors, while large but rare movements of the market are the key events where accurate prediction pays off.

The goal of our proposal is not only to anticipate the occurrence of an extreme value but also to be accurate at predicting its concrete value. Besides that, the interpretabilty is another key point in our application. We want to provide a better understanding of the conditions that lead to these algae blooms so as to enable taking preventive actions.

In this paper we propose a splitting criterion for regression trees which enables the induction of models that meet our applications requirements. We start with a brief overview of our target application in section 2. We then describe some related work in section 3. In section 4, we formalize our target problems and propose evaluation criteria that should guide the search for the best models. Section 5 describes the details of our proposal. The results obtained with this proposal are presented in section 6. We finish with the conclusions of this work and future research directions.

2 Application Description

In recent years there have been some studies concerning the impact man has on environment and subsequent biological processes. Nowadays it is well known that toxic waste from a wide variety of manufacturing processes, farming land run-off and sewage water treatment have a serious effect on the state of rivers. During one of such studies, numerous reports revealed an excessive summer algae growth in temperate climates across the world. High concentration of certain harmful algae in rivers is a serious ecological problem. The blooms of these algae reduce the water clarity and the oxygen levels, causing a massive death of river fish and decreasing the water quality. Therefore, the early forecast of these blooms is of extreme importance. From the analysis of a range of measurable chemical concentrations the objective of this application is to identify the crucial chemical control variables to infer the biological state of the river, in this case the frequence of occurrence of certain algae communities.

The data we used was obtained during a related research study. Water quality samples were taken from different European rivers during a period of approximately one year. These water samples were then submitted to chemical analysis and the resulting measures along with other characteristics like the season of the year,the river size and the river speed were associated with the frequency of seven different harmful algae found in the water[1]. The chemical parameters measured in these water samples[2] were: maximum PH value (mxPH), minimum oxygen value (mnO2), mean value of Chloride (Cl), Nitrates (NO3), Ammonium (NH4), Orthophosphat (oPO4), Phosphat (PO4) and Chlorophyll (Chla). Each water sample is then described by eleven variables.

[1] This data was used in an international data analysis competition (http://www.erudit.de/erudit/competitions/ic-99/)

[2] Actually, each analysed water sample is an aggregation of several water samples carried out over a period of three months, in the same river, and during the same season of the year.

The long-term objective of this modelling task is to anticipate the rare occurrence of high concentrations of these seven harmful algae, given the chemical analysis measurements. This would avoid the need of trained manpower to infer the biological state of the water, specifically the frequence of occurrence of these algae communities. Instead, if the obtained models are able to accurately predict these frequencies based on certain chemical factors (that can be calculated by cheap and automated means, like water probes), one can avoid the microscopic analysis that is slower and more expensive.

The data available for this application consists of two sets of water samples. The first one consists of 200 water samples together with the respective frequency of occurrence of the seven harmful algae. The second set of data consists of 140 water samples, with no information about the frequency of occurrence of these algae. These samples can be regarded as a kind of test set with the goal of predicting the frequency of the seven harmful algae for each water sample.

3 Background

Applications where the main modelling objective are outliers, i.e. rare events, abound in recent data mining literature.

When modelling outliers, one faces two main problems that make this a nontrivial task. Firstly, we need to define what are outliers in our application domain. They can be already indicated as such, or it may be necessary to identify them. In some domains, being an outlier is not just a binary property, as referred by Breunig et al [11]. It depends on how isolated the observation is with respect to its surrounding structure. The process of outlier detection is an important data mining task and is usually referred as outlier mining. There are several methods used for outlier detection: *distance-based* [9], *density-based* [11], *statistical-based*, among others. The second problem with outlier mining is that we need to be aware that what we want to model is rare and as such there are few cases in which we can base our modelling.

Most of the existing work in learning the concept behind the occurrence of outliers uses a classification approach. In these works the main objective is to build a model that is able to distinguish rare from "normal" occurrences, using a discrete target variable. These works include topics like activity monitoring [4], prediction of rare events [20, 21], anticipation of surprising patterns [8], novelty detection, anomaly detection, among others. Most of this research is also linked to applications where a data stream is being monitored with the goal of anticipating rare events, that is time-dependent data.

The importance and impact of rare cases has also been the topic of research on small disjuncts (e.g. [7, 22]). This research is again mainly focused on classification tasks and is also strongly related to the study of applications with unbalanced class distributions (e.g. [5]).

A frequent strategy to bias the models towards being accurate in particular types of cases is the use of differentiated misclassification costs (e.g. [19]). This is a common practice in classification tasks and was also used in solving regression problems through a classification approach [18].

All these classification approaches do not meet our main goal. Firstly, they are not able to accurately predict the specific value of outliers and secondly, their aim is just to detect outliers, not a specific subset of them: those having an extreme value. As such, they are particularly inadequate when these spread over a wide range of values. If the amplitude of the extreme values is relevant for the user, for instance for taking different actions, all these approaches based on classification are not applicable. Obviously, one could further divide the classes representing the extreme values into more specific classes to differentiate their importance but that would mean that we would partition an already low populated class into several classes, thus making our modeling task even more difficult. As such, for this kind of applications only a regression model can handle the problem properly.

Buja and Lee [2] have recently presented a series of new splitting criteria for both classification and regression trees that address related problems. Regarding regression, they propose two different splitting criteria with two objectives: identifying extreme buckets of the data; and identifying pure (low variance) buckets. The first objective is particularly related to ours. The goal of Buja and Lee is to identify areas of the regression surface where the target variable shows a high or low mean value. Although our goal is related to this, we are particularly interested in applications where these extreme values are rare, which demands for specific criteria.

4 Problem Formulation

In this section we present a general formalization of our problem. Let D be a data set, consisting of n cases $\{\langle \mathbf{x}_i, y_i\rangle\}_{i=1}^{n}$, where $\mathbf{x}_i$ is a vector of p discrete or continuous variables, and y_i is a continuous target variable value. As we have mentioned before, we are interested in models that are able to predict accurately rare extreme values of Y. To achieve this goal we need to formalize the notion of rare extreme values. We use the statistical notion of outlier with this purpose. Box plots are visualization tools that are often used to identify extreme-valued outliers. Extreme values are defined in these plots as values above or below the so-called adjacent values [3]. Let r be the interquartile range defined as the difference between the 3rd and 1st quartiles of the target variable. The upper adjacent value, adj_H, is defined as the largest observation that is less or equal to the 3rd quartile plus $1.5r$. Equivalently, the lower adjacent value, adj_L, is defined as the smallest observation that is greater or equal to the 1st quartile minus $1.5r$. Given these two limits we can define our rare extreme values as,

$$\begin{aligned} O &= \{y \in D \mid y > adj_H \vee y < adj_L\} \\ O_H &= \{y \in D \mid y > adj_H\} \\ O_L &= \{y \in D \mid y < adj_L\} \end{aligned} \tag{1}$$

Depending on the application we may have either O_L or O_H empty[3]. Figure 1 shows the box plots of the targets in two applications where we have different types of outliers. These values are drawn with circles in these graphs.

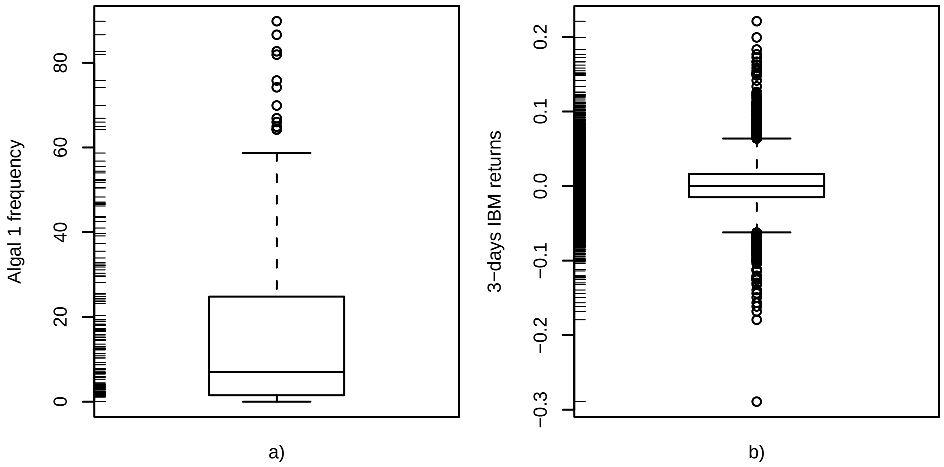

Fig. 1. Two example box plots with different types of extreme values: a) The frequency of an harmful algal; b) The 3-days returns of IBM closing prices.

Having described the main features of our target applications we need to define some evaluation criteria to guide the search for the best models. Typical performance measures used in regression settings, such as the mean squared error, are inadequate as they do not stress the fact that we are only interested in the performance in extreme values. This is the same kind of phenomenon as the one reported regarding the use of classification accuracy on problems with unbalanced class distributions [10, 13].

In the information retrieval literature (e.g. [12]) the notion of relevance seems particularly adequate to our needs. Relevance is defined as the value or utility of a system output as a result of a user search. Relevance is most of the times assessed using two measures: *precision* and *recall*. Precision is defined as the proportion of the cases predicted as target events that really are target events. Recall is defined as the proportion of existing target events that are captured by the model. Our proposal consists of adapting these two measures to our problem setup with the goal of developing a learning tool that maximizes the relevance of the induced model to our application goals.

[3] We will discard applications where both sets are empty as these are not relevant for this study.

We define recall in the context of our target applications as the proportion of extreme-valued outliers in our data that are predicted as such (i.e. covered) by our model,

$$recall = \frac{\mid \{\hat{y} \in \hat{Y}_O \mid (y \in O_H \wedge \; \hat{y} > adj_H) \vee (y \in O_L \wedge \hat{y} < adj_L)\} \mid}{\mid O \mid} \tag{2}$$

where $\hat{Y}_O$ is the set of $\hat{y}$ predictions of the model for the outlier cases (i.e. O).

With respect to precision, if we use its standard definition we have,

$$precision_{stand} = \frac{\mid \{\hat{y} \in \hat{Y}_O \mid (y \in O_H \wedge \; \hat{y} > adj_H) \vee (y \in O_L \wedge \hat{y} < adj_L)\} \mid}{\mid \{\hat{y} \in \hat{Y} \mid \hat{y} < adj_L \vee \hat{y} > adj_H\} \mid} \tag{3}$$

where $\hat{Y}$ is the set of $\hat{y}$ predictions of the model.

However, this definition is not adequate to our goals. For instance, with this formulation, assuming $adj_H = 5.6$, a predicted value of 5.8 would have the same value as a prediction of 10.1, for a test case where the true value is 10.5. In our applications this is not acceptable. Otherwise, the best solution would probably be to discretize the target variable and handle the problem as a classification task with differentiated misclassification costs. As we want to distinguish this kind of errors we need to use another definition of precision ($precision_{regr}$) that takes into account the distance between the predicted and true values. At the same time we want to maintain the scale of the measure within the 0..1 interval so that we are able to integrate recall and precision into a single measure using standard approaches. Our proposed definition of $precision_{regr}$ is the following,

$$precision_{regr} = 1 - NMSE_O \tag{4}$$

where $NMSE_O$ is the normalized squared error of the model for the outliers,

$$NMSE_O = \frac{\sum_{y_i \in O} (\hat{y}_i - y_i)^2}{\sum_{y_i \in O} (\bar{Y} - y_i)^2} \tag{5}$$

where $\bar{Y}$ is the average Y in the training data.

The value of $NMSE_O$ will usually be between 0 and 1. For the cases where this value goes above 1, which means that the model is performing worse than the naive average model, we consider that the $precision_{regr}$ of the model is 0.

Obtaining an overall evaluation measure from the values of recall and $precision_{regr}$ provides a global preference criterion that can be used to guide the search for the models. The F-measure [14] is among the most used measures and is defined as,

$$F = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall} \tag{6}$$

where β controls the relative importance of recall to precision. This is definition we use replacing precision by our proposed $precision_{regr}$.

5 An Approach Using Regression Trees

Regression trees are known for their computational efficiency, model interpretability and competitive accuracy. For these reasons we have decided to use these models as the base paradigm behind our proposal.

Standard regression trees are obtained using a procedure that minimizes the squared error. This means that the best splits for each tree node are chosen to minimize the weighed squared error between the two branches. As mentioned by Buja and Lee [2] this criterion is not adequate for several data mining applications. That is also the case of our target problems. Moreover, outliers can be a problem for standard regression trees as they may distort the selection of the best splits and may also have a large impact on the average values chosen for the leaves of the trees [17].

Just as an illustration of these issues let us consider our target application problem, regarding the predictions of frequency of harmful algal 1 (c.f. Figure 1 for a box plot of the target variable of this problem). The standard regression tree[4] obtained, using 200 observations, is shown on the left-hand side of Figure 2[5]. In spite of the existence of several extreme-valued outliers in the data (c.f. Figure 1a), we can see that the largest value that will ever be predicted by this tree is 49.8, which is not even considered an extreme value according to our definition of Equation (1). This means that this model clearly ignores these extreme values, and if these outliers are of high importance for the application (which is the case on harmful algae blooms), the model is useless. Moreover, the extreme values cause problems to the tree by leading to a distortion of the averages used in the leaves. For instance, the leaf predicting a frequency of 13 includes, among its 56 supporting cases, several extreme values. This can be confirmed by the plot of the errors on the right-hand side of Figure 2, which shows large errors on this leaf. As the value 13 is obtained by averaging the target variable of all the 56 cases in this leaf, the few extreme values are distorting this average, which means that 13 is not a good representative of the most frequent value of the cases in this leaf. In summary, this tree is not very useful for predicting algal blooms (high occurrences of an harmful algal). Moreover, the tree is not interpretable from this perspective, as it does not show when algal blooms occur. Notice, however,that the tree may still achieve good average predictive accuracy, as most of the times the algal frequencies are low.

The main idea of our proposal to avoid the problems reported above is to use the F-measure presented in Equation (6) to guide the split selection procedure used to grow the trees. As such, the key distinguishing feature of our method is the criterion used to select the best test for each tree node. In our proposal the best split s^*, is chosen using the following criterion,

$$s^*(D_t) = \max_{s \in S} \max\left(F(D_{t_L}), F(D_{t_R})\right) \tag{7}$$

where S is the set of trial splits for the node t[6]; D_{t_L} is the subset of cases in t (D_t) that satisfy the test s (i.e. the left sub-branch of t), while D_{t_R} contains the remaining cases (i.e. $D_{t_R} = D_t - D_{t_L}$); and $F(D)$ is the F-measure for a set of cases.

[4] In this paper we have used as base implementation of regression trees the package *rpart* [15] of the open source statistical software R (www.r-project.org). This package is a close re-implementation of most of CART's [1] features.

[5] The numbers between parenthesis in the tree correspond to the number of extreme-valued outliers of each node.

[6] That are the same as in a standard regression tree.

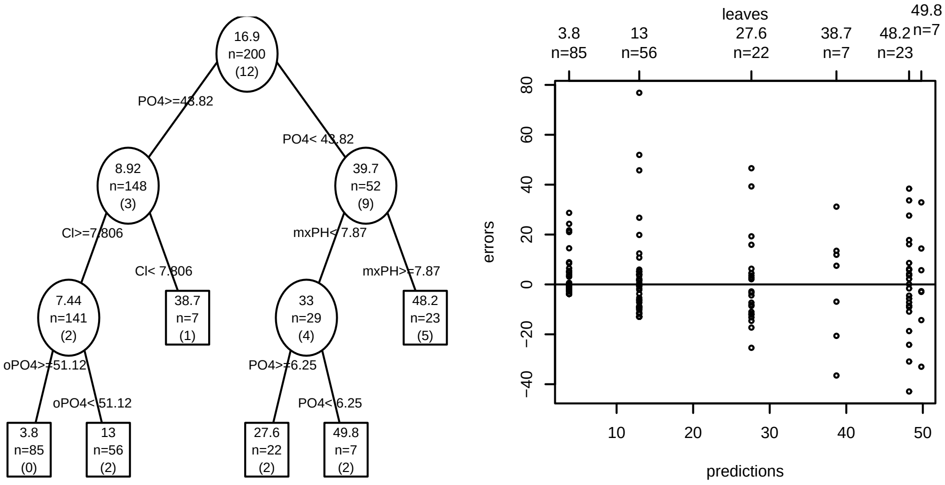

Fig. 2. A standard regression tree and its respective errors, obtained for the problem of predicting the occurrence of the harmful algal 1.

In order to obtain the F-measure for the branches of a candidate split we need to obtain the values of precision and recall. The precision of a node t is given by,

$$precision_{regr_t} = \begin{cases} 0 & \begin{array}{c} \text{if } (O_H(D_t) \cup O_L(D_t) = \phi) \vee \\ (\bar{y}_t > \tilde{Y} \wedge O_H(D_t) = \phi) \vee \\ (\bar{y}_t < \tilde{Y} \wedge O_L(D_t) = \phi) \end{array} \\ 1 - \dfrac{\sum\limits_{y_i \in O_H(D_t)} (\bar{y}_t - y_i)^2}{\sum\limits_{y_i \in O_H(D_t)} (\bar{Y} - y_i)^2} & \begin{array}{c} \text{if } \bar{y}_t > adj_H \ \wedge (\ \bar{y}_t > \tilde{Y} \vee \\ (\ \bar{y}_t = \tilde{Y} \wedge |O_H| > |O_L| \ ) \ ) \end{array} \\ 1 - \dfrac{\sum\limits_{y_i \in O_L(D_t)} (\bar{y}_t - y_i)^2}{\sum\limits_{y_i \in O_L(D_t)} (\bar{Y} - y_i)^2} & \begin{array}{c} \text{if } \bar{y}_t < adj_L \wedge \ (\ \bar{y}_t < \tilde{Y} \vee \\ (\ \bar{y}_t = \tilde{Y} \wedge |O_L| > |O_H| \ ) \ ) \end{array} \end{cases} \quad (8)$$

where $O_H(D_t)$ $(O_L(D_t))$ is the set of cases of node t that belong to $O_H(O_L)$; $\bar{y}_t$ is the average Y value in the node; and $\bar{Y}$ and $\tilde{Y}$ are the average and the median of Y in the training data, respectively.

This means that depending on the value of the node average we consider this branch as a tentative to predict high or low outliers, and calculate its precision accordingly. Even if the node average is not in the outlier range of values we still calculate the precision in the node, using the global median as a threshold for deciding whether to calculate it with respect to high or low outliers. If the node average happens to be equal to the global median, we decide the kind of outliers in an heuristic way, choosing those that are more frequent in the application domain.

Regarding the recall of a node t we use,,

$$recall_t = \begin{cases} 0 & \text{if } \bar{y}_t \geq adj_L \ \wedge \ \bar{y}_t \leq adj_H \\ \frac{\mid y \in D_t \ \wedge \ y \in O_H \mid}{\mid O_H \mid} & \text{if } \bar{y}_t > adj_H \\ \frac{\mid y \in D_t \ \wedge \ y \in O_L \mid}{\mid O_L \mid} & \text{if } \bar{y}_t < adj_L \end{cases} \tag{9}$$

When a trial split leads to a branch having an average target value that is not an outlier, the respective recall is zero. This would lead to an F value of zero according to Equation (6). This is a common situation particularly in top level nodes, where the partitions are still too big, and thus the average Y is seldom an outlier. Moreover, sometimes all trial splits for a node are in these circumstances. This means that we are not able to select the best split for these nodes as all splits have the same score, and thus the tree growth procedure would stop prematurely. These situations occur because in complex applications we seldom find a single split that is able to isolate extreme values in one of the branches so that the branch has an average target that is an outlier. This problem decreases as the tree grows because the number of cases in the nodes gets smaller and thus finding such splits is easier. Although these top level splits have zero recall we should still be able to establish a preference criterion to select one, because we can calculate their precision. In order to overcome this difficulty we have added a small threshold[7] to the value of recall in Equation (6) so that the value of F is not zero even when the recall is null.

Summarizing, our proposal consists of selecting the splits that are able to generate a branch (a subset of cases) with a high value of the F-measure. Notice, that we do not search for a weighed solution between the two branches. Even if one of the branches has a poor F score, as long as the other achieves a high F-measure we have a good candidate split. This strategy is similar to the one followed by Buja and Lee [2], which also do not search for splits with a good compromise between the left and right branch. These strategies lead to unbalanced trees. Still, we share the opinion of Buja and Lee that consider these trees more interpretable.

Another important question that needs to be addressed when developing a tree-based system, is the tree growth stopping criteria. This is a statistical estimation problem and most systems use a two-stages procedure consisting of growing an overly large tree (possibly overfitting the training data), and then use some statistical estimation procedure (e.g. cross validation) for post-pruning this tree[8]. Given that outliers are insignificant from a statistical perspective, these strategies are difficult to implement in our system because they are based on statistical significance. This is consistent with what is mentioned by Weiss and Hirsh [22] in the context of learning from small disjuncts. These authors mention that pruning is considered questionable when the learning objectives are small subsets of cases.

Currently, our method obtains a tree model in a single stage, stopping the tree growth when one of the following conditions arise:

- The F-measure of the node is above a certain user-definable threshold,
- Or the node does not contain any extreme value (i.e. $D_t \cap O = \phi$).

[7] We have used the value of 0.001 in our experiments.

[8] See [16] for an overview of pruning methods for regression trees.

In order to illustrate the effects of using the proposed splitting criteria as opposed to standard least squares methods, we show the obtained regression tree by our method for the same harmful algal of Figure 2 in Figure 3.

As we can see, clearly the first split is a very important one from our perspective. The right subtree, isolates four of the twelve extreme-valued outliers in a "pure" partition. According to this model low values of PO4 lead to high values of occurrence of this algal. This is knowledge that the standard regression tree (*c.f* Figure 2) did not induce. As we have mentioned before, one of our main objectives is to discover the conditions under which algae blooms occur. This objective is clearly met by this kind of splits. Nevertheless, the partition created by this initial split, containing only extreme-valued outliers, is further splitted in order to achieve a better precision. One may question whether this further spliting is useful or it is just overfitting the training data. Ideally, we would like to stop the tree growth, whenever the extreme-valued outliers in a given node are similar. The left subtree does not provide such clear-cut partitions containing only outliers, which possibly means that the remaining outliers do not share any particular feature. Nevertheless, it certainly distinguishes the extreme values of occurrence from the other occurrence values. It creates partitions where there is no extreme-valued outlier. An exception appears at the leaf with mean 7.67 of 140 values which has among them an extreme-valued that it could not isolate, at least with the used F-measure threshold[9]. Still, that extreme value can simply be noisy observation, caused for instance by wrong measurements. In fact, this is another major problem when dealing with this kind of applications. It is difficult to distinguish outliers from noisy observations.

In summary, in comparison to the model in Figure 2, we get a more explicit understanding of the conditions under which the extreme values of this algal appear. From this perspective, we claim that our tree is more informative than a standard regression tree. Although one may argue that this tree is simply overfitting the data, the fact is that as we will see on the results of our experiments for the prediction of this particular algal, our models achieve a better $precision_{regr}$, recall and F measure.

6 Experimental Results

In this section we perform an experimental analysis of the performance in the algae application domain of the trees obtained with our method. Our analysis compares our proposal to its base paradigm, standard regression trees.

Given that we have to predict the frequency of seven different algae, we have divided this taks in seven different multivariate regression problems. For each problem we have a dataset with the same eleven variables (eight continuous attributes - chemical parameters; and 3 nominal attributes - river characteristics), and a different target continuous variable. We have obtained the models for predicting each algal using 200 water samples. For the test phase we had available 140 water samples. Table 1 gives the number of extreme-valued outliers contained in train and test datasets for each algal. There are only high extreme-valued outliers, which correpond to the blooms.

[9] We set the F-measure threshold to 0.7

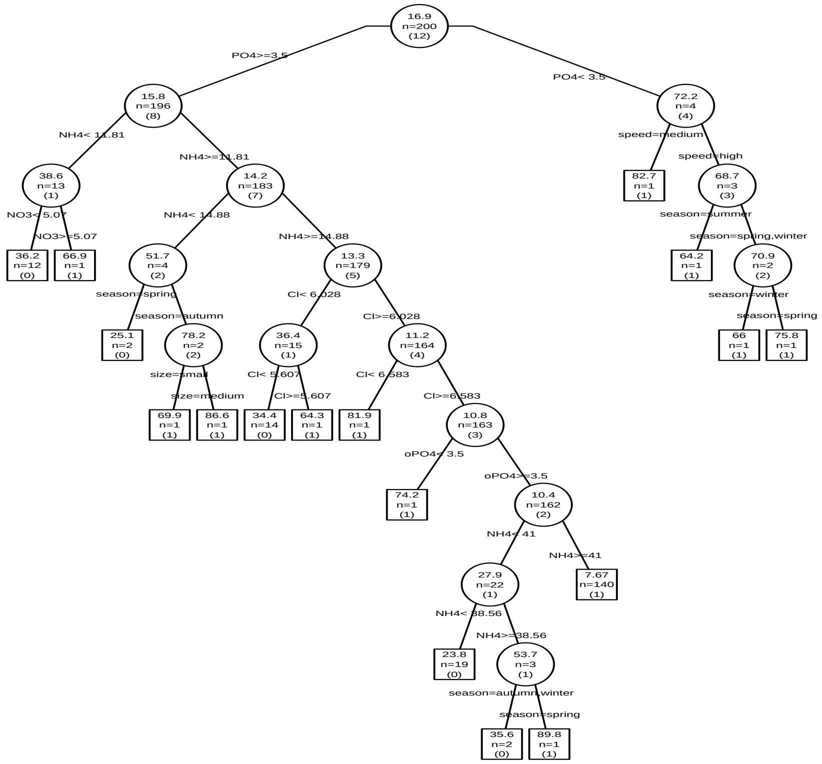

Fig. 3. Regression tree obtained by our method for the problem of predicting the occurrence of the harmful algal 1.

Datasets	**extreme outliers**	
	train	**test**
algal1	12	9
algal2	10	8
algal3	22	16
algal4	16	9
algal5	13	12
algal6	19	20
algal7	21	6

Table 1. Extreme-valued outliers contained in each dataset

We have carried out 5 repetitions of a 10-fold cross validation experiment using the 200 water samples. These experiments were designed with the goal of estimating the average difference in $precision_{Reg}$, recall, and F-measure, between a standard regression tree and our proposed method. For the standard method we have used the package *rpart* of R, using the best tree obtained based on cross validation error-complexity, according to the 0-SE rule method described in [1]. We have decided to use these trees for the same reasons we mentioned before: as we intend to predict rare values it probably will not be a good idea to prune too much the trees. The statistical significance of the observed differences was asserted through paired t-tests. Differences that are significant at the 95% level were marked with one sign, while differences significant at 99% have two signs. Plus (+) signs are used to mark differences favorable to standard regression trees, while minus (−) signs are used to indicate the significant wins of our method. Differences that are not significant at these confidence levels have no sign. The F-measure was calculated with $\beta = 1$, meaning that the same weight was given to $precision_{regr}$ and recall (c.f. Equation (6)).

The results of our experiments are shown on Table 2. This table shows an overall advantage of our method, though not too significant. In terms of the F-measure we generally achieve better results. The exception is algal 7, where a poor $precision_{regr}$ penalizes the F measure. The value of zero recall, obtained by CART for algal 1, is a consequence of models that do not predict any of the outliers as such, which occurs when a tree does not have any leaf with an average value that is an outlier (*c.f* Figure 2).

We have also carried out a different experiment, this time comparing the models over the 140 test samples. The idea here was to carry out an experiment similar to the one done during the data analysis competition that originated these data. In this comparison we have also included the predictions submitted by the winner of this competition, apart from the predictions of CART and our method. The results are shown in Table 3.

CART obtained the worst score for several algae in terms of the F-measure. It got several zero values for recall and also some zero values for $precision_{regr}$. Given the definition of $precision_{regr}$ we use (Equations (4) and (5)) it may seem strange to see some zero values in $precision_{regr}$. These occur because some models have a NMSE at predicting the outliers equal or above one.

Datasets	$precision_{regr}$			$recall$			$Fmeasure$		
	CART	Our method	Signif	CART	Our method	Signif	CART	Our method	Signif
algal1	0.2787366	0.4261633		0.0000000	0.3500000	−−	0.0000000	0.3479222	−−
algal2	0.1258636	0.1455305		0.0066667	0.0700000		0.0058272	0.0667674	
algal3	0.0654252	0.1559563	−	0.0933333	0.1980000	−	0.0714700	0.1656287	−
algal4	0.1049812	0.1358873		0.1416667	0.1973333		0.0954843	0.0985749	
algal5	0.0884105	0.1438035		0.0400000	0.0966667		0.0399061	0.0596241	
algal6	0.0934455	0.1100473		0.1276667	0.1266667		0.0848333	0.0935813	
algal7	0.1037794	0.0905430		0.0950000	0.1083333		0.0960736	0.0863567	

Table 2. Standard regression trees vs our method in terms of $precision_{Regr}$, *recall* and *Fmeasure*.

From the analysis of the results we cannot state there is a evident advantage of our method predictions over the winner predictions, at least from the perspective of the F-measure, which was the criterion used to grow our trees. The winner of the competition

Datasets	$Precision_{regr}$			$Recall$			$Fmeasure$		
	CART	Competition winner	Our method	CART	Competition winner	Our method	CART	Competition winner	Our method
algal1	0.5178113	0.6069383	0.5381055	0.0000000	0.1111111	0.2222222	0.0000000	0.1878355	0.3145459
algal2	0.2949493	0.0830053	0.0649749	0.0000000	0.0000000	0.1250000	0.0000000	0.0000000	0.0855046
algal3	0.0000000	0.2233995	0.0000000	0.0000000	0.1250000	0.1875000	0.0000000	0.1603041	0.0000000
algal4	0.0000000	0.4722144	0.0000000	0.0000000	0.2222222	0.3333333	0.0000000	0.3022206	0.0000000
algal5	0.1480915	0.1943417	0.2738103	0.0000000	0.0000000	0.3333333	0.0000000	0.0000000	0.3006541
algal6	0.1425729	0.2425430	0.0398625	0.3000000	0.1000000	0.1500000	0.1932874	0.1416132	0.0629863
algal7	0.1600282	0.1175599	0.0000000	0.2500000	0.3750000	0.1250000	0.1951429	0.1790035	0.0000000

Table 3. Evaluation of $precision_{regr}$, $Recall$ and $Fmeasure$ with different sets of predictions for the test dataset.

achieves better results in terms of $precision_{regr}$. Our advantage lies in terms of the proportion of extreme-valued outliers in the domain that are captured by the model (i.e. the recall). However, the results in terms of $precision_{regr}$ are not so interesting, we inclusively get some zero values for this statistic. During the training phase, we have tried to understand the reasons for this lack of $precision_{regr}$ in our method. We have varied the F threshold that guides the criterion for stopping tree growth and have observed some variations on these results that seem to indicate that there is some space for improvement of our method by tunning this parameter. Apparently, tree growth may be stopping too soon for these datasets, where our performance seems to be degrading. Still, if $precision_{regr}$ was the key objective we could also tune the β parameter of the F-measure that weights the preference between recall and $precision_{regr}$ when selecting the best splits of the trees[10].

From the observation of Figures 4, we can state that our method predicts more extreme-valued outliers than any other method. But this has its drawbacks in terms of $precision_{regr}$. Although there are some extreme value predictions that correspond to a real extreme value and are reasonably precise (*c.f.* Figure 4), there are other extreme-valued predictions that do not correspond to an extreme-valued outlier at all, or that totally miss on its true value (*c.f.* Figure 5), which penalizes the $precision_{regr}$.

A common problem to all the systems, is that there are several extreme-valued outliers in the test set that were not detected. This is, in fact, a more important matter. Not predicting an algal bloom that actually happened is worse then predicting a algae bloom, that did not happen after all. This type of problems should be taken care in future work.

7 Conclusions

We have described a new splitting criteria for regression trees with the goal of addressing the problem of predicting harmful algae blooms. This problem belongs to a specific class of data mining applications. In these domains the main goal of modeling is to predict accurately outlier values in the target variable and also to understand under which conditions these values occur. Our proposal addresses these application goals by leading to regression trees designed to maximize both the number of extreme-valued outliers that are captured by the model and the precision at predicting their values.

[10] In our experiments we used equal weight.

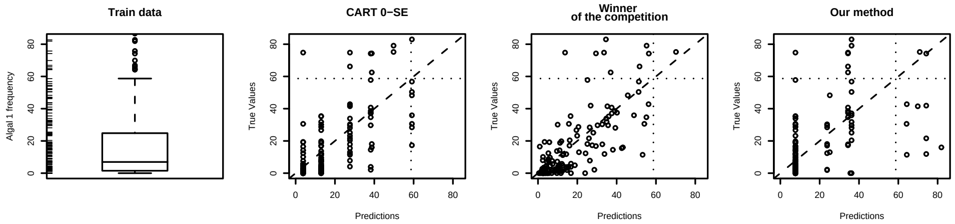

Fig. 4. Predictions obtained for the harmful algal 1, by different systems (outliers range marked by dotted lines).

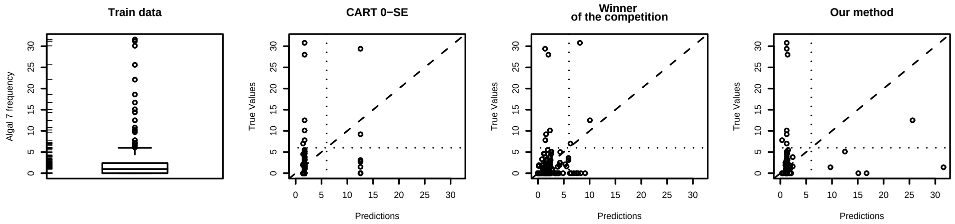

Fig. 5. Predictions obtained for the harmful algal 7, by different systems

At the same time our models are able to identify some of the characteristics shared by these rare events, which can be of key importance for taking preventive actions.

Regarding future work we plan to investigate further this application problem, trying to overcome the failure of our models in terms of $precision_{regr}$ in some situations. Our current explanation lies on the tree growth stopping criteria and we intend to explore other alternatives to the current user settable threshold on the F-measure value. We also intend to improve recall in order to minimize the algae blooms missed by our models.

References

1. L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, 1984.
2. A. Buja and Y.-S. Lee. Data mining criteria for tree-based regression and classification. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 27–36, 2001.
3. W. Cleveland. *Visualizing data*. Hobart Press, 1993.
4. T. Fawcett and F. Provost. Activity monitoring: Noticing interesting changes in behavior. In S. Chaudhuri and D. Madigan, editors, *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 53–62. ACM, 1999.

5. G.Weiss and F. Provost. The effect of class distribution on classifier learning: An empirical study. Technical Report Technical Report ML-TR-44, Department of Computer Science, Rutgers University, 2003.
6. D. M. Hawkins. *Identification of Outliers.* Chapman and Hall, 11 New Fetter Lane, London EC4P 4EE, 1980.
7. R. Holte, L. Acker, and B. Porter. Concept learning and the problem of small disjuncts. In N. Sridharan, editor, *Proceedings of the 11th International Conference on Artificial Intelligence.* Morgan Kaufmann, 1989.
8. E. Keogh, S. Lonardi, and W. Chiu. Finding surprising patterns in a time series database in linear time and space. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 550–556, 2002.
9. Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance based outliers in large datasets. In *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases*, pages 392–403. Morgan Kaufmann, San Francisco, CA, 1998.
10. I. Kononenko and I. Bratko. Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6(1):67–80, 1991.
11. R. Ng M. M. Breunig, H. P. Kriege and J. Sander. Optics of: Identifying local outliers. *Lecture Notes in Computer Science*, 1704:262–270, 1999.
12. C. Meadow, B. Boyce, and D. Kraft. *Text Information Retrieval Systems.* Academic Press, 2nd edition, 2000.
13. F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proc. 15th International Conf. on Machine Learning*, pages 445–453. Morgan Kaufmann, San Francisco, CA, 1998.
14. C. Van Rijsbergen. *Information Retrieval.* Dept. of Computer Science, University of Glasgow, 2nd edition, 1979.
15. T. Therneau and E. Atkinson. An introduction to recursive partitioning using rpart routines. Technical report, Mayo Foundation, 1997.
16. L. Torgo. A comparative study of reliable error estimators for pruning regression trees. In H. Coelho, editor, *Proceedings of the Iberoamericam Conference on AI (IBERAMIA-98)*, 1998.
17. L. Torgo. A study on end-cut preference in least squares regression trees. In P. Brazdil and A. Jorge, editors, *Proceedings of the Portuguese AI Conference (EPIA 2001)*, number 2258 in LNAI, pages 104–115. Springer, 2001.
18. L. Torgo and J. Gama. Regression using classification algorithms. *Intelligent Data Analysis*, 1(4), 1997.
19. P. Turney. Types of cost in inductive learning. In *Proceedings of the Workshop on cost-sensitive learning at the 17th ICML*, pages 15–21, 2000.
20. G. Weiss and H. Hirsh. Learning to predict rare events in event sequences. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, editors, *Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 359–363, New York, NY, 1998. AAAI Press, Menlo Park, CA.
21. G. Weiss and H. Hirsh. Learning to predict extremely rare events. In *AAAI Workshop on Learning from Imbalanced Data Sets*, pages 64–68. Technical Report WS-00-05, AAAI Press, 2000.
22. G. Weiss and H. Hirsh. A quantitative study of small disjuncts. In *Proceedings of AAAI/IAAI*, pages 665–670, 2000.