Rule-based Prediction of Rare Extreme Values

Rita Ribeiro[1] and Luís Torgo[2]

[1] LIACC - University of Porto, R. Ceuta, 118, 6o, 4050-190 Porto, Portugal,
rita@liacc.up.pt
[2] FEP/LIACC - University of Porto, R. Ceuta, 118, 6o, 4050-190 Porto, Portugal,
ltorgo@liacc.up.pt

Abstract. This paper describes a rule learning method that obtains models biased towards a particular class of regression tasks. These tasks have as main distinguishing feature the fact that the main goal is to be accurate at predicting rare extreme values of the continuous target variable. Many real-world applications from scientific areas like ecology, meteorology, finance,etc., share this objective. Most existing approaches to regression problems search for the model parameters that optimize a given average error estimator (e.g. mean squared error). This means that they are biased towards achieving a good performance on the most common cases. The motivation for our work is the claim that being accurate at a small set of rare cases requires different error metrics. Moreover, given the nature and relevance of this type of applications an interpretable model is usually of key importance to domain experts, as predicting these rare events is normally associated with costly decisions. Our proposed system (R-PREV) obtains a set of interpretable regression rules derived from a set of bagged regression trees using evaluation metrics that bias the resulting models to predict accurately rare extreme values. We provide an experimental evaluation of our method confirming the advantages of our proposal in terms of accuracy in predicting rare extreme values.

1 Introduction

In data mining there are several prediction problems for which the rare instances of the concept to be learned are the most important ones. Forecasting large changes on stock prices, ecological or meteorological catastrophes, are a few examples of applications where we are faced with this kind of problems. In all these applications, domain experts are specially interested in having accurate and interpretable predictions of such rare events as these are usually associated with costly actions/decisions. The work presented in this paper addresses this kind of applications in a regression context. These problems present difficult challenges to learning methods as we are trying to model a concept that is rare and less represented than other common concepts in the used data sets.

Predictive data mining tasks fall in two categories: classification, where the target variable is discrete; and regression, where the target variable is continuous. Within classification, this kind of problems is a well-known subject of research

and is related to the problem of unbalanced class distributions [14]. According to some studies (e.g. [5, 12, 14]), the existence of a minority class brings an additional difficulty to the traditional classification methods which are biased to the prediction of the most common values by evaluation criteria such as accuracy.

Most existing work on prediction of rare events within data mining is related to classification tasks. Still, the same type of problems appear in the context of regression. As in classification, one should change the evaluation criteria used by the traditional regression methods that are biased to the prediction of the most common values. In a previous work [9], we have handled this type of problems by proposing a new splitting criterion for CART-like regression trees [1]. Based on this previous work, we now present a rule-based regression system for the prediction of rare and extreme values of a continuous target variable. Compared to the former, this new system improves on both accuracy and interpretability due to the modular characteristics of rule-based systems.

2 Background

Predicting rare events has been receiving an increasing attention from the data mining community. This interest stems from both the important associated applications, and from the fact that learning a concept based on cases that are rare is a non-trivial task for traditional learning methods. Standard machine learning methods are biased to the prediction of the most common values and usually assume that all the prediction errors have the same "cost".

Within classification tasks, one of the proposed approaches to handle rare cases is to use misclassification costs (e.g. [10]). This allows the errors committed at some subset of cases belonging to a rare class to be more penalized and thus models will be biased towards avoiding these errors. Moreover, on these problems with an unbalanced class distribution, some authors [5, 13, 14] have shown that evaluating models by classification accuracy is not adequate. In this context, they proposed different performance metrics based on ROC curves or in measures like precision and recall. The Two-Phase Rule Induction method proposed by Joshi et al. [5] is an example of a rule induction system biased towards the minority class which induces the rule set in two steps considering recall and then precision. Given the clear tradeoff between precision and recall, some works propose the use of *F-measure* [7], one of the measures which combines those two, as shown in Equation 1.

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

where $0 \leq \beta \leq 1$, controls the relative importance of recall to precision.

Nevertheless, all these classification approaches are not directly applicable to the type of problems we wish to address here because our target variables are continuous. Although there are several works that handle regression problems through a classification approach (e.g. [4, 8, 15]), these approaches do not fully meet our target applications requirements. Our goal is not only to capture the

rare and extreme values, but also to be able to predict them as accurately as possible in a numeric perspective. This means that the degree of extremeness of the target variable is also relevant for domain experts, as different actions can be taken according to that degree. One could argue that by having several classes associated to these different degrees of extremeness would overcome this difficulty and allow classification methods to be applied. Still, we argue that this would split an already low populated class associated to rare events into even less frequent classes, thus making the problem even harder. Moreover, this would always be a coarse approximation of an inherently continuous prediction problem.

3 Our System: R-PREV

Given the requirements of our target applications, our goal was to develop a system capable of accurately predict rare extreme values of a continuous target variable in an interpretable way.

In regression methods, as in most learning tasks, the model parameters estimation process is guided by some preference criterion. The most common choices are estimators of the true average prediction error (e.g. mean squared error) of the models. These performance metrics are calculated over all the range of values of the target variable and thus will tend to bias the models to maximize performance over the most common values, as these will have a stronger impact on the overall mean error. This type of preference criteria is not suitable when the interest resides on the performance on a special subset of values that are not very frequent. This is the case of our target applications: the obtained model should perform specially well over the rare and extreme values of the continuous target variable.

The Rule-based Prediction of Rare Extreme Values system (R-PREV) we propose in this paper is based on the trees obtained by a system we have described in a previous work [9]. As such, we will now provide a detailed description of main features of this later system that we will refer as "Base Tree". This consists of a regression tree induction system based on CART [1] but with a different splitting criterion that enables the induction of trees biased towards the prediction of rare and extreme values. The main idea is to use the *F-measure* presented in Equation 1 as splitting criterion for the tree growth. The first step to allow the use of this metric is to provide a formal definition of what is a rare extreme value. In cases where no domain knowledge is available to define this notion, we have used the statistical notion of outlier, given by the box-plot [2], to establish the two thresholds that define the rare extreme high and low values of the target variable. The default thresholds are the so-called adjacent values of the box-plot of a continuous variable. The upper-adjacent value, thr_H, is defined as the largest observation that is less or equal to the 3rd quartile plus $1.5r$, where r is the interquartile range, i.e. the difference between the 3rd and 1st quartiles of the target variable. In an equivalent way, the lower adjacent value, thr_L, is the smallest observation that is greater or equal to the 1st quartile minus $1.5r$.

Once we have these two thresholds, obtained either by existing domain knowledge or by using the information of box-plots, we can define our rare extreme values as,

$$\begin{aligned} RE &= \{y \in Y \mid y > thr_H \vee y < thr_L\} \\ RE_H &= \{y \in Y \mid y > thr_H\} \\ RE_L &= \{y \in Y \mid y < thr_L\} \end{aligned} \tag{2}$$

Depending on the application, we may have either RE_H or RE_L empty. In Figure 1, there is an example of a box-plot obtained for a continuous variable representing the median values of houses in Boston residential areas, using the well-known Boston Housing data set [11]. The circles are the rare extreme values determined by the upper adjacent and lower adjacent values represented by the two horizontal lines outside the box. In this particular case, we only have high rare extreme values, that is, residential areas with extremely expensive houses which distinguish themselves from the rest. According to these thresholds, in this dataset there are no low extremes, i.e. extremely cheap houses.

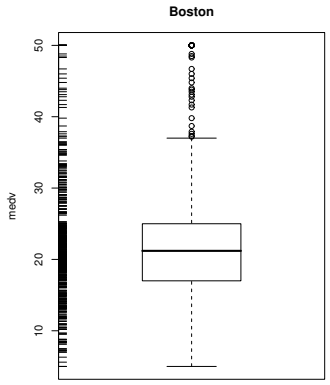

Fig. 1. Example of a box-plot for the 'medv' attribute of Boston dataset.

Once we have the concept of rare extreme value defined, it is necessary to specify how to calculate recall and precision in a regression context[3], as they are required for obtaining the *F-measure* (c.f. Equation 1).

Let $\hat{y}_i$ be the prediction obtained for the case $\langle \mathsf{x}_i \rangle$ whose target true value is y_i. We can define the following two sets:

- $\widehat{Y}_{RE} = \{\hat{y}_i \in \widehat{Y} \mid y_i < thr_L \vee y_i > thr_H\}$, i.e., $\widehat{Y}_{RE}$ is the set of $\hat{y}$ predictions of the model for the rare extreme value cases;

[3] These measures are originally defined in a classification context.

– $\widehat{Y}_{\widehat{RE}} = \{\hat{y}_i \in \widehat{Y} \mid \hat{y}_i < thr_L \vee \hat{y}_i > thr_H\}$, i.e.,$\widehat{Y}_{\widehat{RE}}$ is the set of $\hat{y}$ predictions of the model that are rare extreme values.

Given these sets, we define recall as the proportion of rare extreme values in our data that are predicted as such (i.e. covered) by our model, by the following equation,

$$recall = \frac{\mid \{\hat{y}_i \in \widehat{Y}_{RE} \mid \hat{y}_i < thr_L \vee \hat{y}_i > thr_H\} \mid}{\mid \widehat{Y}_{RE} \mid} \tag{3}$$

Regarding precision, if we use its standard definition it would be defined as the proportion of predicted rare extreme values that are really extremes (c.f. Equation 4).

$$precision_{stand} = \frac{\mid \{\hat{y}_i \in \widehat{Y}_{\widehat{RE}} \mid y_i < thr_L \vee y_i > thr_H\} \mid}{\mid \widehat{Y}_{\widehat{RE}} \mid} \tag{4}$$

However, because we are in a regression context, we adapted the concept of precision so that the amplitude of the differences between predictions and true values is taken into account (c.f. [9]). In this context, we have proposed and used the following definition of precision,

$$precision = 1 - NMSE_{\widehat{RE}} \tag{5}$$

where $NMSE_{\widehat{RE}}$ is the normalized squared error of the model for the cases predicted as rare extreme values,

$$NMSE_{\widehat{RE}} = \frac{\sum\limits_{\hat{y}_i \in \widehat{Y}_{\widehat{RE}}} (\hat{y}_i - y_i)^2}{\sum\limits_{\hat{y}_i \in \widehat{Y}_{\widehat{RE}}} \left(\overline{Y} - y_i\right)^2} \tag{6}$$

Suppose that in some application $thr_H = 10$. If we have a test case with a true value of 12, the proposed definition allows us to signal a prediction of 11 as much better (more precise) than a prediction of 30 for the same test case. Thus, with this proposed definition of precision, we are able to bias the models to be accurate in the degree of extremeness. The use of a normalized metric like $NMSE$ ensures that precision varies between 0 and 1 like recall. For rare situations where $NMSE$ goes above 1, which means that the model is performing worse than the naive average model, we consider that the precision of the model is 0.

As mentioned before the tree growth procedure is guided by a split criterion based on precision and recall, namely the *F-measure* (c.f. Equation 1). The best split will be the one that maximizes the *F-measure* in one of the partitions generated by the split. The tree continues its growing process until the *F-measure*

value of a split goes above some pre-specified threshold f[4], or until there are no more rare extreme values in the current node partition. Full details on the growth of these trees can be obtained in [9].

Interpretability is of key importance to our target applications as the predicted rare events are usually associated with costly decisions. In this context, we have decided to select a rule-based formalism to represent our models. Rules are usually considered to have greater explanatory power then trees, mainly due to their modular characteristics.

We have obtained a set of rules based on trees generated by the "Base Tree" system we have just described. A set of rules can be easily obtained from a regression tree. Each path from the root of the tree to a leaf is transformed into a rule of the form:

$$\textbf{if } cond_1 \wedge cond_2 \wedge ... \wedge cond_n \textbf{ then } v_i$$

where each $cond_k$ is a test over some predictor variable in the considered path i, and v_i is the value of the leaf at the end of that path.

Trees generate a mutually exclusive partition of the input space of a problem. This means that given any test case, only one rule will cover it. We have decided to obtain our set of rules from a set of trees obtained through a bagging process in order to both decrease the variance component of the error of the resulting model and also to eliminate this mutually exclusivity property, thus ensuring a higher modularity of each rule in the final model. We start by obtaining a pre-specified number of stratified bootstrap samples, so that for each sample we can have a similar distribution function for the target variable. For each sample we run the regression tree method referred above and then transform it into a set of rules. This process is repeated for all trees obtained from the bootstrap samples.

Once we get this large set of rules, R, originated from different trees we try to simplify it in two forms: individually, using some simple logical simplifications of the conditions on each rule; and globally by eliminating some rules from this set using the information regarding their specificity and *F-measure*. As a result of this process we obtain an ordered rule set, also known as a decision list.

We measure the specificity of a rule by the number of cases that are uniquely covered by that rule,

$$spec(r) = \mid \{\langle \mathsf{x}_i \rangle \mid cover(R, \langle \mathsf{x}_i \rangle) = \{r\}\} \mid \tag{7}$$

where R is the entire rule set and $cover(R, \langle \mathsf{x}_i \rangle)$ is the set of rules that cover the case $\langle \mathsf{x}_i \rangle$.

We want to retain rules with high specificity because they represent knowledge that is not captured by any other rule. Regarding the remaining rules

[4] Experiments carried out in a previous work [6], have shown that the "best" setting for the threshold f is domain dependent and thus for achieving top performance for a particular problem, some tunning process is recommended. In the context of the experiments of this paper we have used a the default value of 0.7, which these experiments have shown as a generally reasonable setup in many domains.

(whose specificity is zero), we order them by their evaluation criterion (*F-measure* score) and then select the top k rules according to a user-specified margin parameter, m.

This means that the final theory, T, is given by the rules belonging to the initial set, R, ordered by their *F-measure*, such that,

$$T = \{\ r \in R \mid spec(r) > 0 \vee F(r) > (1 - m) \cdot F(r_{top})\ \} \tag{8}$$

where $F(r)$ is the *F-measure* of the rule r, r_{top} is the rule with the best *F-measure* and m is the margin parameter.

Notice that if we want a theory formed only by rules that have some specificity then we can set the margin parameter (m) to 0. This is also the setting that leads to a smaller theory, as larger values will increase the number of rules.

This rule selection process removes the complete coverage property that results from the mutual exclusivity of a tree, which means that there may exist a test case that is not covered by any of the rules in the final theory, T. For these situations we have added a default rule at the end of our ordered set of rules, T. This default rule basically predicts the mean value of the target variable (c.f. Equation 9).

$$T_f = T \cup \{\ \textbf{if}\ null\ \textbf{then}\ \overline{Y}\ \} \tag{9}$$

Obtaining a forecast for a test case $\langle \mathsf{x}_i \rangle$ involves averaging all the predictions of the rules satisfied by $\langle \mathsf{x}_i \rangle$, weighted by their respective *F-measure*,

$$\hat{y}_i = \frac{\sum_{k=1}^{|C|} F(C[k]) \cdot predict(C[k], \langle \mathsf{x}_i \rangle)}{\sum_{k=1}^{|C|} F(C[k])} \tag{10}$$

where $C = cover(T_f, \langle \mathsf{x}_i \rangle)$.

The algorithm of R-PREV can be summarized by the steps given in Figure 2.

1. establish thr_L and thr_H for the target variable Y;
2. specify β and f parameters;
3. specify a margin parameter m;
4. generate n bootstrap stratified samples;
5. $R = null$;
6. for each i from 1 to n
   (a) t_i = BaseTree($sample_i$,thr_L,thr_H,β,f);
   (b) obtain the rule set R_i from the tree t_i;
   (c) perform logic simplifications over R_i;
   (d) $R = R \cup R_i$;
7. sort the rules in set R by their *F-measure* values;
8. obtain the final theory:
   $T_f = \{\ r \in R \mid spec(r) > 0 \vee F(r) > (1 - m) \cdot F(r_{top})\ \} \cup \{\ \textbf{if}\ null\ \textbf{then}\ \overline{Y}\ \}$

Fig. 2. R-PREV main algorithm.

4 An Analysis of System R-PREV

In this section we analyze the performance of our proposal with respect to the two main features that distinguish it from our previous work [9]: the improved accuracy at forecasting rare extreme values; and the interpretability advantages of its rule-based formalism.

4.1 Predictive Accuracy

We have carried out a set of experiments with the goal of estimating the performance of our proposed system in the task of predicting rare extreme values, when compared to related regression methods. Namely, we have compared several variants of our system (R-PREV) with different settings in terms of the margin parameter ($m = 0\%, 25\%, 50\%, 75\%, 100\%$), with: the base system used to obtain the trees ("BaseTree" [9]); a CART-like regression tree; and a bagged CART-like regression tree (BaggCART). The selection of competitors was carried out with the goal of having a better understanding of the gains caused by each of the added features of our system when compared to a simple CART-like tree.

The experiments were carried out on a set of real-world problems, some of which are commercial applications. The methods were tested over 24 data sets using 10 repetitions of a 10-fold cross-validation procedure. Regarding the methods that use bagging we have used 50 bootstrap stratified samples.

Taking into account recent results reported in [3] regarding the comparison of multiple models over several data sets, we have used the Friedman test and the post-hoc Nemenyi test for asserting the statistical significance of the observed differences in performance.

We have estimated the performance of the different methods by means of the *F-measure* values as this statistic is better at characterizing the performance in rare extreme values. We have used a β value of 0.5 for the *F-measure* calculation. This choice is justified by the fact that, given that we are addressing a numeric prediction task, precision is always the most important factor. With $\beta = 0.5$, we are giving it doubled importance relatively to recall. For each dataset, we calculated the mean value of F obtained by each method over all repetitions. In order to check whether the systems can be considered equivalent, we applied the Friedman test. This test ranks internally the obtained results for each dataset over all the compared methods and obtains rank data like the one shown in Table 1.

In these experiments we have used the default parameters of all systems as our goal was not to optimize their performance on each individual problem. Therefore, some individual results may not be as good as possible. This is particularly noticeable in our R-PREV system, as previous experiments [6] have revealed a certain sensitivity to the setting of parameter f.

Regardless of this, the Friedman test applied over the rank data, reported a significant difference between the 8 compared methods at a significance level of 5%. Given this result, we proceed by applying the post-hoc Nemenyi test to

datasets	R-PREV $m=0$	R-PREV $m=25$	R-PREV $m=50$	R-PREV $m=75$	R-PREV $m=100$	Base Tree	CART	BaggCART
servo	8.0	7.0	4.0	3.0	5.0	6.0	2.0	1.0
triazines	7.5	4.0	3.0	2.0	1.0	5.0	6.0	7.5
algae1	7.5	6.0	3.0	2.0	1.0	4.0	5.0	7.5
algae2	6.5	6.5	3.0	2.0	1.0	4.0	6.5	6.5
algae3	7.5	7.5	3.0	2.0	1.0	5.0	4.0	6.0
algae4	7.5	7.5	4.0	2.0	1.0	6.0	3.0	5.0
algae5	7.0	7.0	3.0	2.0	1.0	4.0	5.0	7.0
algae6	7.5	7.5	5.0	2.0	1.0	6.0	3.0	4.0
algae7	8.0	7.0	3.0	2.0	1.0	6.0	4.0	5.0
machine-cpu	8.0	6.0	7.0	1.0	2.0	4.0	5.0	3.0
china	7.0	7.0	3.0	2.0	1.0	5.0	4.0	7.0
sard0	8.0	7.0	6.0	2.0	1.0	5.0	4.0	3.0
sard2	8.0	4.0	3.0	2.0	1.0	6.0	7.0	5.0
sard3	8.0	4.0	3.0	1.0	2.0	6.0	5.0	7.0
sard4	7.5	3.0	4.0	2.0	1.0	7.5	6.0	5.0
sard5	7.0	3.0	2.0	5.0	1.0	7.0	7.0	4.0
sard0-new	8.0	7.0	5.0	2.0	1.0	6.0	4.0	3.0
sard1-new	8.0	4.0	3.0	2.0	1.0	6.0	7.0	5.0
Boston	8.0	6.0	7.0	4.0	2.0	5.0	3.0	1.0
onekm	7.0	5.0	1.0	2.0	3.0	4.0	7.0	7.0
cw-drag	8.0	7.0	5.0	4.0	3.0	6.0	1.0	2.0
co2-emission	8.0	5.0	4.0	2.0	1.0	3.0	6.0	7.0
acceleration	8.0	6.0	5.0	2.0	1.0	3.0	7.0	4.0
available-power	8.0	7.0	6.0	4.0	5.0	3.0	2.0	1.0
avg.ranks	7.65	5.88	3.96	2.33	1.62	5.1	4.73	4.73

Table 1. Ranking of the different regression methods over the set of datasets.

the rank data in order to compare all methods to each other. The results of this test are better visualized by the CD (critical difference) Diagram proposed by [3] and presented in Figure 3. This diagram represents the information on the statistical significance regarding every pairwise comparison between methods, which means that each method is represented by 7 symbols as we have 8 methods being compared. The methods are plotted at their respective average ranking value in terms of the X axis (notice that CART and BaggCART have the same average ranking). The vertical axis position has no meaning. A dotted line connecting two symbols has the meaning that according to the Nemenyi test, the methods are significantly different at a 5% level. Bold lines indicate that the difference in average ranking is not statistically significant at the same confidence level. An ideal performance would be a method whose symbols are at the right most position of the graph (lowest average ranking), and are connected only by dotted lines to every other method (all pairwise comparisons are statistically significant). Thus, we can observe that R-PREV with $m = 100$ and with $m = 75$ are clearly the two best methods with a very high statistical significance in almost all pairwise comparisons. In particular, they are significantly better than CART, BaggCART and "BaseTree". Still, we should also remark that R-PREV with $m = 0$ is significantly worse than all standard CART related methods. This clearly indicates the importance of having more rules contributing to the predictions, but unfortunately also means that our model needs theories with more rules (thus less interpretable) to achieve top performance.

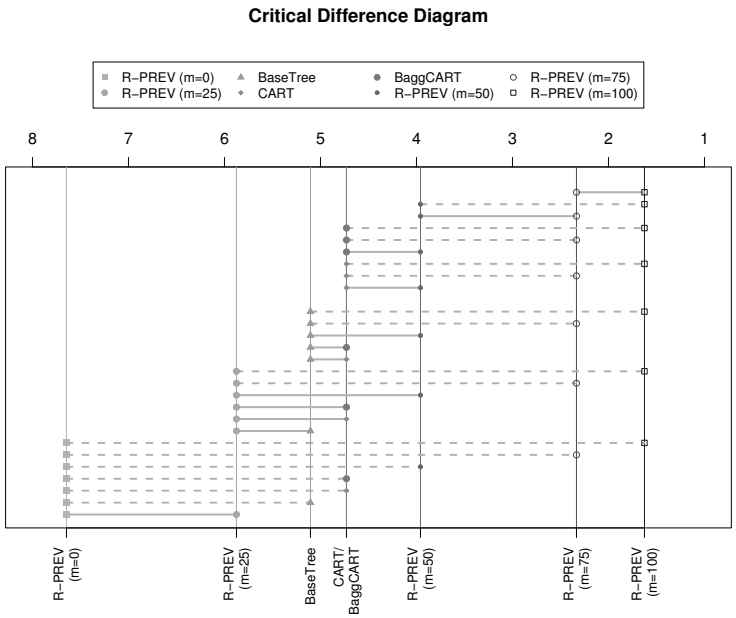

Fig. 3. The Critical Difference Diagram obtained from our experiments.

In summary, these results provide clear evidence that R-PREV can achieve very competitive predictive performance in terms of accuracy at predicting rare extreme values measured by the F statistic.

4.2 Interpretability

As we have seen in the previous section, in order to achieve top performance we need a larger number of rules. Still, for each test case only a reduced number of rules is used to obtain the prediction and domain experts can analyze these before taking any action associated with rare events. This was a property that we were seeking when developing our system: provide domain experts with comprehensible explanations of the system predictions.

In order to illustrate this comprehensibility issue we have selected the Boston Housing data set. This selection was guided by the fact that this domain concerns a topic (housing prices as a function of socio-economical factors) that is easily understandable by non-expert readers. Other data sets would require domain knowledge in order to comment the interpretability and/or reasonability of the rules obtained by R-PREV.

Figure 4 shows two of the top most valuable rules obtained by R-PREV for the Boston Housing data set. These two rules give some interesting insights regarding the more expensive areas in Boston. One of the rules tells us that an area where

houses have more than 7 rooms tends to have a very high median price of houses. The second rule says that areas with a low crime rate, near the city center, with a small percentage of lower status population and with small houses, are also quite expensive. Both seem to capture quite common sense knowledge and would probably be regarded as correct by an expert of this domain, which would then mean that this expert would easily "accept" the model predictions. This kind of knowledge cannot be captured by standard methods like CART, because the models they obtain are focused on being accurate at the more frequent cases and not the rare extreme values like ours.

```
*******************************************************************************
                DOMAIN:  Boston
                Rare Extreme Thresholds: < 5 and > 37
                Number of outliers:  37 of 506  examples.
*******************************************************************************
=======================================     =======================================
If                                          If
      rm >= 7.43700                               crim < 8.81054   and
                                                  dis in [1.13330,1.38485[   and
                                                  lstat < 12.30000 and
                                                  rm < 7.43700
Then                                        Then
      medv = 45.2129                              medv = 50
---------------------------------------     ---------------------------------------
coverage :                                  coverage :
      nr exs  =  30                               nr exs  =  3
      nr rare extreme values  =  28               nr rare extreme values  =  3
      specificity  =  0                           specificity  =  0
stats :                                     stats :
      F  =  0.8913744                             F  =  0.3061224
=======================================     =======================================
```

Fig. 4. Example of two of top best rules obtained by R-PREV for Boston domain.

5 Conclusions

In this paper we have described a rule-based regression system, called R-PREV, conceived to address a particular class of problems that occur in several real-world applications. The applications we envisage have as main objective to produce accurate and interpretable predictions of rare extreme values of a continuous target variable. We claim that this particularity makes these problems hard to solve by the standard regression methods as they are biased to achieve a good performance on the most common values of a target variable. A different evaluation criterion is needed to overcome this limitation.

In this paper we present an extension of our previous approach to this class of problems. The extension was developed with the goal of improving both the accuracy and the interpretability of the models. The experimental evaluation we have carried out provides clear evidence that R-PREV outperforms a set of other systems with a high degree of statistical confidence. Regarding interpretability,

which is crucial in most applications we are addressing, the use of a rule-based formalism leads to highly interpretable models, as we have shown by some examples.

Acknowledgements

This work was partially supported by FCT project MODAL (POSI/4049/2001) co-financed by POSI and by the European fund FEDER, and by a PhD scholarship given by FCT (SFRH/BD/1711/2004) to Rita Ribeiro.

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. W. Cleveland. *Visualizing data*. Hobart Press, 1993.
3. Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, January 2006.
4. Nitin Indurkhya and Sholom M. Weiss. Solving regression problems with rule-based ensemble classifiers. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–292, New York, NY, USA, 2001. ACM Press.
5. Mahesh V. Joshi, Ramesh C. Agarwal, and Vipin Kumar. Predicting rare classes: Comparing two-phase rule induction to cost-sensitive boosting. In *Proceedings of the Sixth European Conference, PKDD 2002*, pages 237–249, 2002.
6. R. Ribeiro. Prediction models for rare phenomena. Master's thesis, Faculty of Economics, University of Porto, Portugal, February 2004.
7. C. Van Rijsbergen. *Information Retrieval*. Dept. of Computer Science, University of Glasgow, 2nd edition, 1979.
8. L. Torgo and J. Gama. Regression using classification algorithms. *Intelligent Data Analysis*, 1(4), 1997.
9. L Torgo and R. Ribeiro. Predicting outliers. In N. et al. Lavrac, editor, *Proceedings of Principles of Data Mining and Knowledge Discovery (PKDD-03)*, volume 2838 of *LNAI*, pages 447–458. Springer-Verlag, 2003.
10. 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.
11. UCI Machine Learning Repository - http://www.ics.uci.edu/ mlearn/MLSummary.html.
12. Gary Weiss and Haym Hirsh. Learning to predict rare events in categorical time-series data. In *AAAI Workshop on Predicting the Future: AI Approaches to Time-Series Problems*, volume WS-98-07, pages 83–90. AAAI Press, 1998.
13. Gary Weiss and Haym Hirsh. Learning to predict extremely rare events. In *AAAI Workshop on Learning from Imbalanced Data Sets*, volume WS-00-05, pages 64–68. AAAI Press, 2000.
14. Gary Weiss and Foster 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, 2001.
15. Sholom M. Weiss and Nitin Indurkhya. Rule-based machine learning methods for functional prediction. *Journal of Artificial Intelligence Research*, 3:383–403, 1995.