Mining Tolerance Regions with Model Trees

Annalisa Appice and Michelangelo Ceci

Dipartimento di Informatica, Università degli Studi di Bari
via Orabona, 4 - 70126 Bari - Italy
{appice, ceci}@di.uniba.it

Abstract. Many problems encountered in practice involve the prediction of a continuous attribute associated with an example. This problem, known as regression, requires that samples of past experience with known continuous answers are examined and generalized in a regression model to be used in predicting future examples. Regression algorithms deeply investigated in statistics, machine learning and data mining usually lack measures to give an indication of how "good" the predictions are. Tolerance regions, i.e., a range of possible predictive values, can provide a measure of reliability for every bare prediction. In this paper, we focus on tree-based prediction models, i.e., model trees, and resort to the inductive inference to output tolerance regions in addition to bare prediction. In particular, we consider model trees mined by SMOTI (Stepwise Model Tree Induction) that is a system for data-driven stepwise construction of model trees with regression and splitting nodes and we extend the definition of trees to build tolerance regions to be associated with each leaf. Experiments evaluate validity and quality of output tolerance regions.

1 Introduction

Regression trees extend well-known decision trees to deal with a continuous goal variable Y [1]. Given a training set $D = \{(\boldsymbol{x}, y) \in \boldsymbol{X} \times Y \mid y = g(\boldsymbol{x})\}$, regression trees are built *top-down* by recursively partitioning a feature space $\boldsymbol{X}$ spanned by m independent (or predictor) variables x_i (both continuous and categorical) to mine a piecewise function f that is hopefully close to g on the domain $\boldsymbol{X}$. Regression trees differ from decision trees in associating continuous constant values rather than discrete classes with leaves.

Model trees generalize regression trees in the sense that they approximate g by means of a piecewise (linear) function instead of a constant value. They are mined in form of hierarchies of nodes starting with a root node (top-level node), where internal nodes (splitting nodes) are generally associated with a logical test on predictor variables, while leaves (i.e. bottom nodes in the hierarchy) are associated with (linear) functions [16,10,19,17,18,2,11,4]. A different tree structure with both regression and splitting nodes is mined by the system SMOTI [12]. Regression nodes perform straight-line regression, while splitting nodes partition the training space.

Model trees mined by SMOTI are proved to be characterized by understandability properties and valuable predictive capabilities [12]. Anyway, they lack

F. Esposito et al. (Eds.): ISMIS 2006, LNAI 4203, pp. 560–569, 2006.

measures (confidence values) to give an indication of how good the predictions are. This is a common limitation of the most of regression systems, included model tree induction systems, which output the bare prediction and lead data miners to rely on the previous experience or relatively loose theoretical upper bounds on the probability of error to ensure the quality of the given prediction.

Constructing a tolerance region [6] is a relevant notion in estimating the confidence value of a continuous prediction. Although, this is an established area in statistics, no method developed in traditional statistics takes into account high-dimensional problem typical of data mining. Some solutions come from Bayesian methods and PAC theory, which give confidence values to complement bare predictions. Anyway, Bayesian methods are only applicable if the stochastic mechanism generating the data is known in every detail [13]. In contrast, PAC theory assumes data is generated by some completely unknown i.i.d. distribution, but they give non trivial-results only when data are particularly clean [13].

Algorithmic theory of randomness is exploited to provide valid confidence information [9] within a transductive inference framework [7]. In this case, the main disadvantage is the relative computational inefficiency, since for every test all computations need to be started from scratch. To overcome inefficiency limitations of transductive inference without serious loss in the confidence of the quality values, we follow the main inspiration in [14] of replacing transductive inference with inductive inference. In particular, we resort to inductive inference to construct the tolerance regions to be associated with the leaves of model trees mined with SMOTI. The idea is to randomly split training data in proper data and calibration data. The model tree is top-down mined on proper training data, while the tolerance regions are constructed on calibration data. Tolerance regions are constructed by strictly including calibration values predicted from the model tree at each leaf with an error event with probability at most δ.

The paper is organized as follows. In the next Section, we briefly describe the system SMOTI, while in Section 3 we present two methods based on inductive inference to construct tolerance regions to be associate with leaves of model trees mined by SMOTI. Experimental results to evaluate validity and quality of tolerance regions constructed in both cases are commented in Section 4. Finally, conclusions and future works are drawn.

2 Mining Model Trees with SMOTI

For the sake of self-consistency of the work, we briefly recall some characteristics of SMOTI. SMOTI learning algorithm starts with a root node t_0 that is associated with the entire training data and recursively proceeds by: i) growing the tree by performing a splitting test on the current node t and introducing the nodes t_L (left child of t) and t_R (right child of t), ii) growing the tree by performing a regression step on some continuous attributes and introducing its unique child t_R,iii) stopping the tree's growth at the current node t. Hence, model trees are mined with two types of nodes: regression nodes, which perform only straight-line regressions, and splitting nodes, which partition training data.

Splitting and regression nodes pass down observations to their children in two different ways. For a splitting node t, only a subgroup of the $N(t)$ observations in t is passed to each child (left or right). No change is made on training cases. For a regression node t, all the observations are passed down to its only child, but the values of both the dependent and independent continuous variables not included in the multiple linear model associated with t are transformed to remove the linear effect of those variables already included. Thus, descendants of a regression node will operate on a modified training set. This is done in accordance to the statistical theory of linear regression, where the incremental construction of a multiple linear model is made by removing the linear effect of introduced variables each time a new independent variable is added to the model [5]. Leaves are associated with (multiple) linear functions built by combining straight-line regressions along the path from the root to the current node. In this way SMOTI potentially solves the problem of modeling phenomena, where some variables have a global effect while others have only a local effect and prevent problems related to two-staged induction algorithms, which first build the tree structure and then associate leaves with multiple models. On the other hand, the stepwise construction provides a solution to problems of efficiency and collinearity by selecting a subset of variables. A more detailed explanation of SMOTI and a comparison with other TDIMT methods are reported in [12].

Although experimental results prove that model trees mined by SMOTI are simple and easily interpretable, and their analysis reveals interesting patterns, no indication on the reliability of performed predictions is provided. Such reliability can be estimated in form of tolerance regions.

3 Inducting Tolerance Regions in SMOTI

Tolerance regions can be efficiently induced within the Inductive Confidence Machine (ICM) framework [14] by overcoming computational inefficiency of Transductive Confidence Machine (TCM) [15,9], where the confidence of prediction is estimated by resorting to transductive inference [7]. In this paper, we extend the ICM framework to the case of tolerance regions to be predicted with a model tree. In particular, we intend to ensure the confidence of predictive patterns mined with SMOTI by providing some practical information on the prediction reliability. To this aim, the training set $D = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_n}, y_n)\}$ is randomly partitioned into two sub-sets, that is, the proper set $P = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_l}, y_l)\}$ and the calibration set $C = \{(\boldsymbol{x_{l+1}}, y_{l+1}), \ldots, (\boldsymbol{x_n}, y_n)\}$ $(l \leq n)$.

SMOTI-ICM mines a model tree τ from the proper set P and builds a tolerance region for each leaf of τ by considering only the calibration examples of C reaching the leaf. Formally, let N_τ^L be the set of leaves in τ, $C(t) = \{(\boldsymbol{x_{t_1}}, y_{t_1}), \ldots, (\boldsymbol{x_{t_k}}, y_{t_k})\} \subseteq C$ is the sub-set of calibration examples reaching $t \in N_\tau^L$. Fixed the confidence level $1 - \delta$, SMOTI-ICM processes $C(t)$ to induce a tolerance region $TR_t(\boldsymbol{x})$, to be predicted for any example $\boldsymbol{x} \in \boldsymbol{X}$ reaching t.

Given an example $(\boldsymbol{x}, y)$ (with possible unknown value of y) that reaches t in τ, the predicted tolerance region is $TR_t(\boldsymbol{x})$ such that the probability that $y \in TR_t(\boldsymbol{x})$ is $1-\delta$. The tolerance region depends on the bare prediction $\widehat{y} = f_t(\boldsymbol{x}) =$

$\beta_0 + \sum_j (\beta_j \times x_j)$ that is obtained by combining all straight-line regressions along the path from the root to t with the best straight-line regression associated with t in τ [5]. The tolerance region TR_t is built according to two procedures, i.e., the *count-based procedure* and the *normal distribution based procedure*.

3.1 Count-Based Procedure

The count-based procedure is inspired by the work of [14] and operates on the list of absolute residuals. More formally, for each leaf $t \in N_\tau^L$, the bag of the absolute residuals $AR(t)$ is built as follows:

$$AR(t) = \{ar_i \quad | \quad ar_i = |y_i - \widehat{y_i}| \wedge (\boldsymbol{x}_i, y_i) \in C(t) \wedge \widehat{y_i} = f_t(\boldsymbol{x}_i)\} \tag{1}$$

The bag of absolute residuals is processed to obtain $AR'(t)$ that contains absolute residuals of $AR(t)$ without repetitions. $AR'(t)$ is used to build a tolerance region predicted at the leaf t by imposing that the probability that examples fall in this region is greater or equal to (1-δ). Operatively, by denoting with:

$$count_p(t) = \#\{ar_i \quad | \quad ar_i \in AR(t) \wedge ar_i > ar'_p\} \quad p = 1, \ldots, \#AR'(t) \tag{2}$$

where ar'_p is the p-th value in $AR'(t)$. In this definition, $count_p(t)$ represents the number of calibration examples $(\boldsymbol{x}_i, y_i) \in C(t)$ whose absolute residual $|y_i - \widehat{y_i}|$ is greater than the *p-th* absolute residual from $AR'(t)$. Hence, the tolerance region $TR_t(\boldsymbol{x})$ is $[\hat{y} - \alpha_\delta(t), \hat{y} + \alpha_\delta(t)]$, where $\alpha_\delta(t)$ is estimated as follows:

$$\alpha_\delta(t) = min\{ar'_p \quad | \quad ar'_p \in AR'(t) \wedge \delta \geq \frac{count_p(t)}{\#AR(t) + 1}\} \tag{3}$$

Intuitively, this approach allows to *locally* build the tightest tolerance region for each leaf $t \in N_\tau^L$ with respect to $C(t)$ such that the absolute frequency of calibration examples $(\boldsymbol{x}_i, y_i) \in C(t)$ having $y_i \in [\widehat{y_i} - \alpha_\delta(t), \widehat{y} + \alpha_\delta(t)]$ is (1-δ) at least. This is quite different from [14], where a single tolerance region is induced on the entire calibration set C according to the single rule returned by the Ridge Regression. Such difference depends on the tree structure we are dealing with that permits to estimate a different $\alpha_\delta(t)$ for each leaf t. However, inducing a tolerance region for each leaf t may pose problems when $C(t)$ contains few examples or $C(t)$ is empty ($C(t) = \oslash$). In this case, $\alpha_\delta(t)$ cannot be reliably estimated. To avoid this problem we *globally* estimate $\alpha_\delta(t)$ by considering the entire calibration set C when $\#C(t) \leq \varphi \times \#C$ ($\varphi \in [0, 1]$ is a user-defined parameter). In such situations global estimation is preferred to local one.

3.2 Normal Distribution Based Procedure

The normal distribution based procedure resorts to the statistical interpretation of a tolerance region in form of a confidence interval, i.e., range of values around a point estimates. In statistics, confidence intervals are the most prevalent form of interval estimation to address the issue of taking a random sample from a population and computing a statistic, such as the mean, from observed data.

A confidence interval is expressed by means of a lower and upper limit for the mean, which give an indication of how much uncertainty there is in the estimate of the true mean [8]. More in general, let W be a random variable, we denote by $\widehat{\theta}(w_1, \ldots, w_n)$ an estimator of some unknown population parameter θ such that $\widehat{\theta}(w_1, \ldots, w_n)$ is computed on a set of observations $w_1, \ldots, w_n$ collected for W. We assume that $g(w)$ is the known probability density function (p.d.f.) of W and determine the value α'_δ (α''_δ) such that there is a fixed probability $\delta/2$ that $w > \alpha'_\delta$ ($w < \alpha''_\delta$). Formally:

$$\delta/2 = P(w > \alpha'_\delta) = \int_{\alpha'_\delta}^{\infty} g(w)dw \quad \left(= P(w < \alpha''_\delta) = \int_{-\infty}^{\alpha''_\delta} g(w)dw \right) \tag{4}$$

Hence, the probability that $\widehat{\theta} \in [\alpha'_\delta, \alpha''_\delta]$ can be computed as follows:

$$P(\alpha'_\delta < \widehat{\theta} < \alpha''_\delta) = \int_{\alpha'_\delta}^{\alpha''_\delta} g(w)dw = 1 - \delta/2 - \delta/2 = 1 - \delta \tag{5}$$

This suggests the idea that the tolerance region to be predicted for any test example can be built as a confidence interval. In particular, for each leaf node $t \in N_\tau^L$, we consider the random variable $W_t = |Y - \hat{Y}| - S_{Y\hat{Y}_t}$ where $\hat{Y} = f_t(\boldsymbol{X})$ and $S_{Y\hat{Y}_t} = \frac{1}{\#C(t)} \sum_{(\boldsymbol{x}_i, y_i) \in C(t)} (y_i - \widehat{y}_i)$ and assume that values of W_t observed for calibration examples reaching t are i.i.d. distributed with respect to a Normal distribution with mean μ_{θ_t} and variance $\sigma^2_{\theta_t}$ ($W_t \sim N(\mu_{\theta_t}, \sigma^2_{\theta_t})$). Thus $g(w)$ in Equation 5 is $g(w) = \frac{1}{\sigma_{\theta_t}\sqrt{2\pi}} \exp\left(-\frac{\left(w-\mu_{\theta_t}\right)^2}{2\sigma^2_{\theta_t}}\right)$. Fixed the confidence level δ, the confidence interval for the mean of W_t is $[\mu_{\theta_t} - z^*\sigma_{\theta_t}, \mu_{\theta_t} + z^*\sigma_{\theta_t}]$, where z^* represents the point on the standard normal density curve such that the probability of observing a value greater than z^* is equal to $\delta/2$ [3].

In this setting, a test example $(\boldsymbol{x}_i, y_i)$ (with possible unknown value of y_i) that reaches t is correctly predicted by the confidence interval when $(|y_i - \widehat{y}_i| - S_{Y\widehat{Y}_t}) \in [\mu_{\theta_t} - z^*\sigma_{\theta_t}, \mu_{\theta_t} + z^*\sigma_{\theta_t}]$ that is,

$$\begin{aligned} y_i \in & [\widehat{y}_i - (\mu_{\theta_t} + S_{Y\widehat{Y}_t} + z^*\sigma_{\theta_t}), \widehat{y}_i - (\mu_{\theta_t} + S_{Y\widehat{Y}_t} - z^*\sigma_{\theta_t})] \\ & \cup [\widehat{y}_i + \mu_{\theta_t} + S_{Y\widehat{Y}_t} - z^*\sigma_{\theta_t}, \widehat{y}_i + \mu_{\theta_t} + S_{Y\widehat{Y}_t} + z^*\sigma_{\theta_t}] \end{aligned} \tag{6}$$

When $\mu_{\theta_t} + S_{Y\widehat{Y}_t} - z^*\sigma_{\theta_t} \leq 0$, that is, the mean $S_{Y\widehat{Y}_t}$ of residuals is not relatively high with respect to the variance of W_t, the tolerance region for y_i is:

$$y_i \in [\widehat{y}_i - (\mu_{\theta_t} + S_{Y\widehat{Y}_t} + z^*\sigma_{\theta_t}), \widehat{y}_i + (\mu_{\theta_t} + S_{Y\widehat{Y}_t} + z^*\sigma_{\theta_t})] \tag{7}$$

The use of the random variable W_t is motivated by the fact that this allows to identify not only upper limits, but also lower limits for the residuals when this is statistically significant. This turns to have a significant play-off when no all residuals are distributed in the neighborhood of the regression line.

Similarly to the count-based procedure, some problems may occur in constructing tolerance regions when $C(t)$ contains only few calibration examples.

This is mainly due to the assumption of normal distribution for W_t that follows from the Central Limit Theorem: the random variable W_t can be assumed normally distributed when the sum of n (in the limit of large n) independent observations for W_t has a finite variance [3]. This means that the $\alpha_\delta(t)$ must be *globally* estimated by considering the entire calibration set C when $\#C(t) \leq \varphi \times \#C$.

4 An Illustrative Example

In this section we present an illustrative example that shows how the two approaches work on a simple dataset. Let us consider the proper training set:

X	1	2	3	4	5	6	7	8	9
Y	2	4	3	4	7	7	8.5	8	10

where X is the independent variable, while Y is the response variable. We are interested in building tolerance regions to be predicted at the leaves of the model tree mined by SMOTI on the proper set above. In particular, we consider the case that the tolerance regions are built by using the calibration set C_1, that is:

X	1	2	3	4	5	6	7	8	9
Y	5	1	5	5	6	7	7	8	11

In this case, SMOTI-ICM mines a model tree with a single leaf labeled with the straight-line regression: $f(x) = 0.9667x + 1.111$. Tolerance regions predicted at this leaf are shown in Figure 1. They are built by using either the count-based procedure or the normal distributed based procedure. More precisely, the count-based procedure builds the tolerance region $[\widehat{y} - 2.92, \widehat{y} + 2.92]$, while the normal distribution based procedure returns the tolerance region $[\widehat{y} - 2.508, \widehat{y} + 2.508]$. This suggests the idea that the normal distribution based procedure is more conservative than the count-based one, since it seems to identify tighter intervals. Moreover, this case shows that tolerance regions built from the normal distribution based procedure may collapse in a single interval area when $\mu_{\theta_t} + S_{Y\widehat{Y}_t} - z^* \sigma_{\theta_t} \leq 0$. Differently, when the calibration set is C_2:

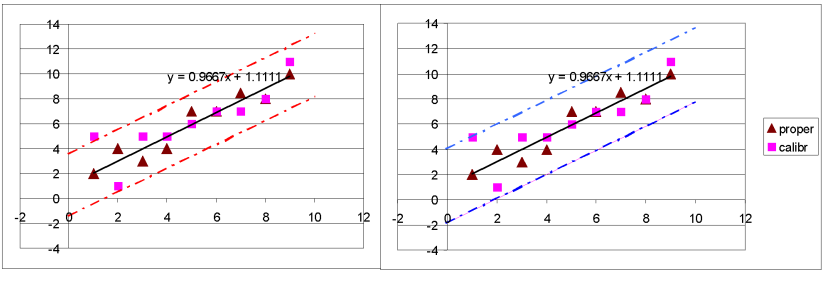

Fig. 1. Tolerance regions built on the calibration set C_1. On the left side, the tolerance region built with the normal distribution based procedure. On the right side, the tolerance region built with the count-based procedure.

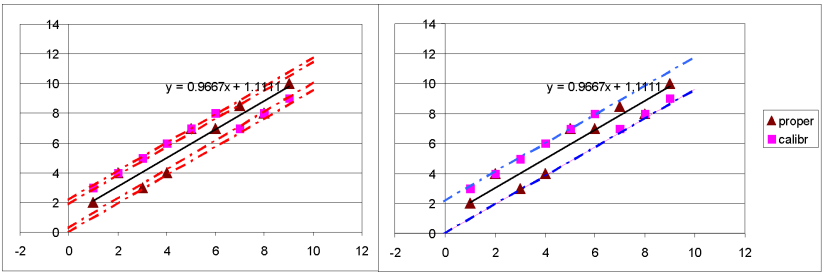

Fig. 2. Tolerance regions built on the calibration set C_2. On the left side, the tolerance region built with the normal distribution based procedure. On the right side, the tolerance region built with the count based procedure.

X	1	2	3	4	5	6	7	8	9
Y	3	4	5	6	7	8	7	8	9

the tolerance region built with the normal distribution based procedure includes two separate interval areas (see Figure 2).

5 Experiments

Tolerance regions predicted by SMOTI-ICM are empirically evaluated on eight datasets (see Table 1) taken from either the UCI ML Repository[1] or the HTL web site[2], which have a continuous variable to be predicted. Experiments aim to check reliability and tightness of tolerance regions predicted by SMOTI-ICM according to both the count-based procedure (CBP) and the normal distribution based procedure (NBP). Reliability is estimated by counting how many times SMOTI-ICM fails to predict the tolerance region that contains the real value of some test example, while tightness is estimated by computing the median value of the width of the tolerance regions predicted at each leaf of the mined model tree. Tightness is estimated in terms of median value instead of mean since the median is less sensitive to noise, overfitting and outliers than the mean [14]. CBP and NBP are compared by a 10-fold cross validation of data. Reliability is estimated on the basis of the average predictive error percentage, that is:

$$AvgPredictiveError\% = \frac{1}{k}\sum_{v_i \in V}\left(\frac{1}{\#v_i}\sum_{(\boldsymbol{x_j}, y_j)\in v_i} D_i(y_j, \hat{y_j})\right), \tag{8}$$

with $V = \{v_1, \ldots, v_k\}$ a cross-validation partition where each v_i is a set of indices of training cases, k is the number of folds (i.e., 10), $\#v$ is the number of cases in v_i and y_j is the continuous response value for the j-th testing case from v_i. $D_i(y_j, \hat{y_j}) = 0$ if y_j belongs to the tolerance region predicted by the model tree

[1] http://www.ics.uci.edu/∼mlearn/MLRepository.html
[2] http://www.niaad.liacc.up.pt/∼ltorgo/Regression/Data Sets.html

Table 1. Datasets used in evaluation of tolerance regions built by SMOTI-ICM

Dataset	#Cases	#Attributes	Goal	Range of Y
Auto-Mpg	392	8	Predicting fuel consumption	[9.0,46.6]
Auto-Price	159	27	Predicting auto price	[5118,35056]
Bank8FM	4499	9	Predicting fraction of bank customers who leave the bank due to full queues	[0.0,0.8022]
Cleveland	297	14	Predicting heart disease in a patient	[0.0, 4.0]
Housing	506	14	Housing values in areas of Boston	[5.0, 50.0]
Pyrimidines	74	28	Predicting activity (QSARs) from the descriptive structural attributes	[0.1,0.9]
Triazines	74	61	Predicting structure (QSARs) from the descriptive structural attributes	[0.1, 0.9]
Wisconsin Cancer	186	33	Predicting time to recur for a breast cancer case	[1,125]

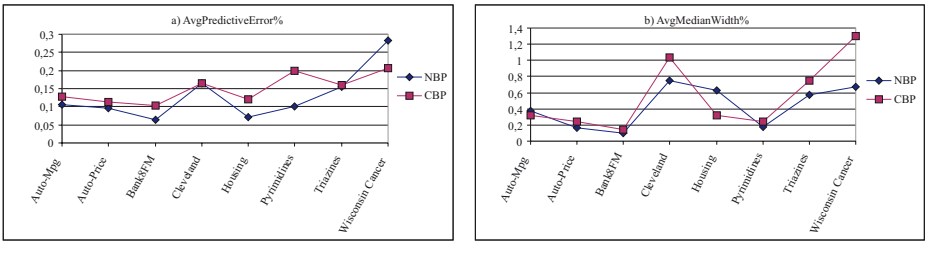

Fig. 3. Average a) predictive error and b) median width percentage for the tolerance regions predicted by SMOTI-ICM according to both CBP and NBP with $\delta = 0.1$

τ_i when applied to $\boldsymbol{x}_j$, 1 otherwise. The training set V/v_i is randomly divided into proper set P_i (70%) and calibration set C_i (30%). Each model tree τ_i is grown on P_i, while the tolerance regions at leaves of τ_i are induced on C_i.

Similarly, tightness is estimated on the basis of average percentage of the median values of the width of the tolerance regions predicted at the leaves of τ_i with respect to the range of Y in each corresponding dataset, that is:

$$AvgMedian\% = \frac{1}{k} \sum_{v_i \in V} \frac{median_{t_j \in N^L_{\tau_i}}(width(TR(t_j))}{width(range(y))}. \tag{9}$$

Reliability of tolerance regions predicted by SMOTI-ICM is reported in Figure 3.a. Tolerance regions are built according to both CBP and NBP with $\delta = 0.1$ and $\varphi = 0.05$. In both settings, results confirm on the testing data that both procedures build tolerance regions that contain the true label y with an error event with probability close to δ. Moreover, we observe that NBP outputs tolerance regions that perform, in general, a lower error than CBP. The exception is represented by Wisconsin Cancer, where both NBP and CBP perform an error

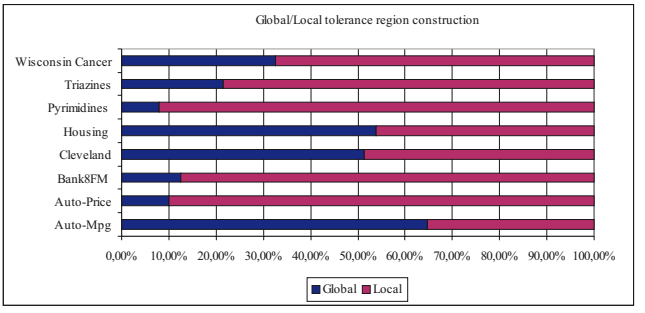

Fig. 4. Average percentage of tolerance regions globally built on the entire calibration set vs. percentage of tolerance regions locally built on sub-set of only calibration examples reaching each leaf node t

(0.28 vs. 0.20) that is significantly higher than δ. This is due to the fact that SMOTI stepwise procedure is failing in mining a piecewise function f close to the underlying function g [12]. Finally, when we compare the tightness of tolerance regions reported in Figure 3.b, we observe that low error performed by NBP is in general combined with tight tolerance regions, that is the best case.

The average percentage of tolerance regions built by using the entire calibration set (global approach) vs the average percentage of tolerance regions built by using only calibration examples reaching a leaf (local approach) are reported in Figure 4. Global construction is preferred to the local one when the percentage of calibration examples reaching the leaf node is less than the 5% (φ) of the entire calibration set. Results show that local approach is profitably preferred to the global one when enough calibration examples reach the leaf. This allows to avoid overfitting problems when building the tolerance region locally to the leaf.

6 Conclusions

In this work we resort to an inductive inference approach to construct tolerance regions as a measure of reliability of the bare prediction output by model trees mined with SMOTI. This is done by randomly splitting training data into proper data and calibration data such that the model tree structure is grown on the proper data, while the tolerance regions to be predicted at each leaf node of the tree are constructed from the calibration data. Tolerance regions are constructed according to two different procedures, named count-based procedure and normal distribution based procedure, to give an indication of how "good" the predictions are. Experimental results prove validity and quality of output tolerance regions in terms of reliability and tightness. In particular, the tolerance regions predicted by the normal distribution based procedure seem to generally correspond to the best case, that is, a lower error obtained by means of tight tolerance regions. As future work we plan to stepwise build tolerance regions to be associated with intermediate regression nodes of model trees mined by SMOTI.

Acknowledgment

This work is partial fulfillment of the research objective of COFIN-2005 project "Metodi di Data Mining Spaziali si supporto alle decisioni ambientali".

References

1. L. Breiman, J. Friedman, R. Olshen, and J. Stone. *Classification and regression tree*. Wadsworth & Brooks, 1984.
2. P. Chaudhuri, M. C. Huang, W. Y. Loh, and R. Yao. Piecewise-polynomial regression trees. *Statistica Sinica*, 4:143–167, 1994.
3. G. Cowan. *Statistical Data Analysis*. Oxford University Press, USA, 1998.
4. A. Dobra and J. E. Gehrke. SECRET: A scalable linear regression tree algorithm. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, 2002.
5. N. R. Draper and H. Smith. *Applied regression analysis*. John Wiley & Sons, 1982.
6. D. A. S. Fraser. *Non-parametric Methods in Statistics*. New York: Wiley, 1957.
7. A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *14th Conference on Uncertainty in Artificial Intelligence*, pages 148–155, 1998.
8. G. J. Hahn and W. Q. Meeker. *Statistical Intervals : A Guide for Practitioners*. John Wiley & Sons: Series in Probability and Statistics, 1991.
9. V. V. Ilia Nouretdinov and A. Gammerman. Transductive confidence machine is universal. *Lecture Notes in Artificial Intelligence*, 2842:283 – 297, 2003.
10. A. Karalic. Linear regression in regression tree leaves. In *Proceedings of Int. School for Synthesis of Expert Knowledge, ISSEK 1992*, pages 151–163, 1992.
11. W. Y. Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12:361–386, 2002.
12. D. Malerba, F. Esposito, M. Ceci, and A. Appice. Top down induction of model trees with regression and splitting nodes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):612–625, 2004.
13. T. Melluish, C. Saunders, I. Nouretdinov, and V. Vovk. Comparing the bayes and typicalness frameworks. In *12th European Conference on Machine Learning, ECML 2001*, volume 2167 of *LNCS*. Springer-V, 2001.
14. H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *13th European Conference on Machine Learning, ECML 2002*, pages 345–356, London, UK, 2002. Springer-V.
15. K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman. Transductive confidence machines for pattern recognition, 2001.
16. J. R. Quinlan. Learning with continuous classes. In Adams and Sterling, editors, *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, pages 343–348. World Scientific, 1992.
17. L. Torgo. Functional models for regression tree leaves. In D. Fisher, editor, *Proceedings of the Fourteenth International Conference on Machine Learning, ICML 1997*, pages 385–393, Nashville, Tennessee, 1997.
18. L. Torgo. *Inductive Learning of Tree-based Regression Models*. PhD thesis, Department of Computer Science, University of Porto, Porto, Portugal, 1999.
19. Y. Wang and I. Witten. Inducing model trees for continuous classes. In M. Van Someren and G. Widmer, editors, *Proceedings of European Conference of Machine Learning, ECML 1997*, pages 128–137, 1997.