Mining Model Trees with Regression and Splitting Nodes

Annalisa Appice Michelangelo Ceci Floriana Esposito Donato Malerba

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

Abstract. Model trees are tree-based models that associate leaves with multiple linear models and are used to solve prediction problems in which the response variable is numeric. In this paper a method for mining model trees is presented. Its main characteristic is the construction of trees with two types of nodes: regression nodes, which perform only straight-line regression, and splitting nodes, which partition the feature space. The multiple linear model associated to each leaf is then built stepwise by combining straight-line regressions reported along the path from the root to the leaf. In this way, internal regression nodes contribute to the definition of multiple models and capture global effects, while straight-line regressions at leaves can only capture local effects. The proposed method has been implemented in the system SMOTI and evaluated on both artificially generated datasets and benchmark datasets used for studies on both regression and model trees. The first set of results show that SMOTI outperforms the state-of-the-art model tree induction system M5', while the second set of results do not allow us to draw statistically significant conclusions. However, model trees induced by SMOTI are generally easily interpretable and their analysis may reveal interesting patterns in the data.

1. Background and motivation

In the classical regression setting, data is generated independently and identically distributed from an unknown distribution on some domain **X** and labeled according to an unknown function g with range Y. The domain **X** is spanned by m independent (or predictor) random variables x_i (both numerical and categorical), while Y is a subset of $\Re$, that is, the dependent (or response) variable y is continuous. A learning algorithm receives a training sample $S=\{(\mathbf{x}, y) \in \mathbf{X} \times Y \mid y=g(\mathbf{x})\}$ and attempts to return a function f close to g on the domain **X**, where closeness is often measured by means of the expected square error. Learning algorithms that induce a *model tree* approximate the function g by a piecewise linear function, since they associate leaves with multiple linear models. They are evolutions of algorithms that learn *regression trees* [2], which associate a constant to each leaf, and approximate the function g by means of a piecewise constant one.

Some of the model tree induction systems developed are: M5 [12], RETIS [5], M5' [15], TSIR [6], and HTL [13]. Almost all these systems perform a *top-down* induction of models trees (TDIMT) by first building the tree structure through recursive partitioning of the training set and *then* by associating leaves with models. This means that the partitioning of the feature space (*splitting stage*) does not take into account the regression models that can be associated with the leaves (*predictive stage*). Consequently, the choice of the heuristic evaluation function, used to select the best partition, is not coherent with the prediction model associated to the leaves, and the induced tree may fail to capture the underlying model.

This problem is solved in RETIS, whose heuristic evaluation function used for a binary split minimizes a function of the mean square error (MSE) computed with respect to the regression planes found for both the left and the right child. In practice, for each possible partitioning, the best regression planes at leaves are chosen, so that the selection of the optimal partitioning can be based on the result of the prediction stage.

One weakness of this solution is that the regression planes involve all continuous variables. When some of the independent variables are linearly related to each other, that is, they are (approximately) collinear, several problems may occur [3]. First, if at least one of the independent variables is a perfect linear function of one or more other independent variables in the equation, the coefficients may not be uniquely determined. Second, estimates of the regression coefficients fluctuate markedly from sample to sample. Regression coefficients cannot be used as interpretive tools to evaluate the relative importance of the independent variables. Interestingly, problems due to collinearity do not show in the model's fit. The resulting model may have very small residuals, but the regression coefficients are actually poorly estimated. A treatment suggested in this case is deleting some of the variables in the full fitted model. Therefore, *variable subset selection* is a desirable part of regression analysis that is not supported by RETIS.

An additional problem of almost all TDIMT systems is that the regression model associated with a leaf is built on the basis of those training cases falling in the corresponding partition of the feature space. Therefore, models in the leaves have only a *local* validity and do not consider the *global* effects that some variables might have in the underlying model. In model trees, global effects can be represented by variables that are introduced in the linear models at higher levels of the tree. However, this requires a different tree-structure, like that adopted in TSIR, where internal nodes can either define a partitioning of the feature space or introduce some regression variables in the models to be associated to the leaves.

In this paper, a new TDIMT system, named SMOTI, is presented. It overcomes problems encountered in existing systems by exhibiting the following characteristics:

1. Induced model trees have two types of internal nodes: regression nodes, which perform only straight-line regression, and splitting nodes, which partition the feature space. Leaves are always regression nodes.
2. A *multiple linear model* can be associated to each leaf. It involves all the numerical variables in the regression nodes along the path from the root the leaf.
3. Variables involved in regression nodes at top levels of the tree capture global effects, while those involved in regression nodes close to the leaves capture local effects.

4. The heuristic evaluation function is coherent with respect to the linear model associated to the leaves.
5. Only a subset of numerical variables may be involved in multiple linear models associated to the leaves, thus solving problems due to collinearity.
6. Induced model trees can be easily interpreted.

This paper extends and revises the work in [7,8] by removing the effect of independent variables also from the dependent variable, by improving the look-ahead strategy for regression nodes, by introducing tests on the equivalence of two straight-line regressions, by defining a better selection strategy for subsets of discrete variables, by adding two stopping criteria,by amending the computational complexity analysis and by providing the reader with new experimental results.

The paper is organized as follows. In the next Section the method implemented in SMOTI is introduced, and its computational complexity is analyzed. In Section 3 some experimental results are reported for both artificially generated data and benchmark datasets. For this second set of results the detected presence of some interesting patterns is also discussed.

2. Stepwise construction of model trees

In SMOTI, the development of a tree structure is not only determined by a recursive partitioning procedure, but also by some intermediate prediction functions (see Figure 1). This means that there are two types of nodes in the tree: regression nodes and splitting nodes. They 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, and no change is made on the variables. For a regression node *t*, all the observations are passed down to its only child, but both the values of the dependent variable and the values of the (continuous) independent variables not included in the model are transformed, to remove the linear effect of those variables already included.[1] Thus, descendants of a regression node do operate on a modified dataset. This transformation is coherent with the statistical procedure for the incremental construction of multiple linear regression models, according to which each time a new independent variable is added to the model its linear effect on remaining variables has to be removed [3].

The validity of either a regression step or a splitting test on a variable X_i is based on two distinct evaluation measures, $\rho(X_i,Y)$ and $\sigma(X_i,Y)$ respectively. The variable X_i is of a continuous type in the former case, and of any type in the latter case. Both $\rho(X_i,Y)$ and $\sigma(X_i,Y)$ are MSE[2], therefore they can be actually compared to choose between three different possibilities:

- growing the model tree by adding a regression node *t*
- growing the model tree by adding a splitting node *t*
- stopping the tree's growth at node *t*.

[1] Differently from SMOTI, only the dependent variable is transformed in TSIR. Hence, the linear model associated by TSIR to the leaves cannot be interpreted from a statistical viewpoint.

[2] This is another difference with TSIR, which, in the case of node selection, minimizes the absolute deviation between a *constant* value (the median) and the observed values *Y*. On the contrary, SMOTI coherently minimizes the square error with respect to the partially constructed regression model at each node.

```
PROCEDURE build-SMOTI-tree(X, Y, R, L, T)
Input:
   X: a set of m independent variables $X_i$,
   Y: the dependent variable
   R: a subset of numerical variables in X
   L={($\mathbf{x}_j$, $y_j$) | j=1, …, N }  a training set where $\mathbf{x}_j$=( $x_{1j}$, …, $x_{mj}$)
Output:
   T: a model tree with regression and split nodes built on (X,Y)
Best-$\rho$ = $\infty$; Best-$t_r$ = a node whose model is the estimated mean $\overline{Y}$
FOR each numeric variable $X_i \in R$ DO
  Compute the best regression node $t_r$ with variable $X_i$
  Compute the evaluation measure $\rho(X_i,Y)$ for $t_r$
  IF $\rho(X_i,Y)$ <= Best-$\rho$ THEN Best-$\rho$ = $\rho(X_i,Y)$; Best-$t_r$ = $t_r$ END IF
END FOR
IF stopping criteria THEN T = leaf Best-$t_r$;
ELSE
  Best-$\sigma$ = $\infty$; Best-$t_s$ = nil;
  FOR each variable $X_i \in X$ DO
    Compute the best split node $t_s$ with variable $X_i$
    Compute the evaluation measure $\sigma(X_i,Y)$ for $t_s$
    IF $\sigma(X_i,Y)$ <= Best-$\sigma$ THEN Best-$\sigma$ = $\sigma(X_i,Y)$; Best-$t_s$ = $t_s$ END IF
  END FOR
  IF Best-$\sigma$ > Best-$\rho$ THEN
    build-SMOTI-tree(X, Y, R, {($\mathbf{x}_j$, $y_j$)$\in$L | test in Best-$t_s$ is true}, $T_L$)
    build-SMOTI-tree(X, Y, R, {($\mathbf{x}_j$, $y_j$)$\in$L | test in Best-$t_s$ is false}, $T_R$)
    T = tree with root in Best-$t_s$, left branch $T_L$, right branch $T_R$
  ELSE
    Let $X_r$ be the variable in Best-$t_r$;
    FOR EACH case ($\mathbf{x}_j$, $y_j$)$\in$L DO
      FOR EACH numeric variable $X_i \in X-\{X_r\}$ DO
      $x'_{ij}$ = residuals of $x_{ij}$ after removing the effect of $X_r$
      END FOR
      $y'_j$ = residuals of $y_j$ when the regression in Best-$t_r$ is performed
    END FOR
    build-SMOTI-tree($\{X'_j\} \cup \{X_r\}$, $Y'$, $\{X'_j\}$, {($\mathbf{x}'_j$, $y'_j$) | ($\mathbf{x}_j$, $y_j$)$\in$L}, T')
    T = tree with root in Best-$t_r$ and child T'
  END IF
END IF
END PROCEDURE
```

Fig. 1. Main procedure of SMOTI. Revisions of the procedure with respect to the original versions described in [7,8] are reported in italics. They concern key steps of the algorithm.

The evaluation measure $\sigma(X_i,Y)$ is coherently defined on the basis of the multiple linear model to be associated with each leaf. In the case of SMOTI it is sufficient to consider the best straight-line regression associated to each leaf t_R (t_L), since regression nodes along the path from the root to t_R (t_L) already partially define a multiple linear regression model (see Figure 2).

If X_i is continuous and α is a threshold value for X_i then $\sigma(X_i,Y)$ is defined as:

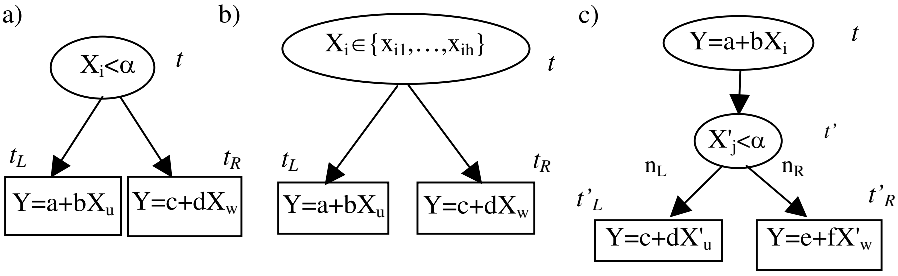

Fig. 2. (*a*) A continuous split node *t* with two straight-line regression models in the leaves. (*b*) A discrete split node discrete in the right tree. (*c*) Evaluation of a regression step at node *t*, based on the best splitting test below.

$$\sigma(X_i,Y) = \frac{N(t_L)}{N(t)} R(t_L) + \frac{N(t_R)}{N(t)} R(t_R)$$

where *N(t)* is the number of cases reaching *t*, *N*(t_L) (*N*(t_R)) is the number of cases passed down to the left (right) child, and *R*(t_L) (*R*(t_R)) is the resubstitution error of the left (right) child, computed as follows:

$$R(t_L) = \sqrt{\frac{1}{N(t_L)} \sum_{j=1}^{N(t_L)} (y_j - \hat{y}_j)^2} \quad \left(R(t_R) = \sqrt{\frac{1}{N(t_R)} \sum_{j=1}^{N(t_R)} (y_j - \hat{y}_j)^2} \right).$$

The estimate $\hat{y}_j = a_0 + \sum_{s=1}^{m} a_s x_s$ is computed by combining all univariate regression lines associated to regression nodes along the path from the root to t_L (t_R).

Possible values of α are found by sorting the distinct values of X$_i$ in the training set associated to *t*, then identifying one threshold between each pair of adjacent values. Therefore, if the cases in *t* have *k* distinct values for X$_i$, *k*-1 thresholds are considered. Obviously, the lower σ(X_i,*Y*), the better the split X$_i$≤α.

If X_i is discrete, SMOTI partitions attribute values into two sets, so that binary trees are always built. Some TDIMT systems, such as HTL and M5', use the same criterion applied in CART. More precisely, if *k* is the number of distinct values for X_i and $S_{X_i} = \{x_{i_1}, x_{i_2}, \dots, x_{i_k}\}$ is the set of distinct values of X_i, S_{X_i} is sorted according to the sample mean of *Y* over all cases in *t*. This approach is based on the assumption that the models associated to the leaves are the sample means. However, this is not the case of SMOTI, which associates a linear regression function to each partition once the split is defined. Therefore, a different heuristic is necessary. SMOTI uses a greedy strategy as suggested by [10]. The system starts with an empty set Left$_X$=∅ and a full set Right$_X$=S$_x$. It moves one element from Right$_X$ to Left$_X$ such that the move results in a better split. The evaluation measure σ(X_i,*Y*) is computed as in the case of continuous variables, therefore, a better split decreases σ(X_i,*Y*). The process is iterated until there is no improvement in the splits. For all possible splits, the evaluation measure σ(X_i,*Y*) is computed as in the continuous case.

The split selection criterion explained above can be improved to consider the special case of identical regression model associated to both children (left and right). When this occurs, the best straight-line regression associated to *t* is the same as that associated to both t$_L$ and t$_R$, up to some statistically insignificant difference. In other words, the split is useless and can be filtered out from the set of alternatives. To check this special case, SMOTI compares the two regression lines associated to the children according to a statistical test for coincident regression lines [16, pp. 162-167].

The evaluation of a regression step $Y=a+bX_i$ at node t cannot be naïvely based on the resubstitution error $R(t)$:

$$R(t) = \sqrt{\frac{1}{N(t)} \sum_{j=1}^{N(t)} (y_j - \hat{y}_j)^2}$$

where the estimator $\hat{y}_i$ is computed by combining all univariate regression lines associated to regression nodes along the path from the root to t. This would result in values of $\rho(X_i,Y)$ less than or equal to values of $\sigma(X_i,Y)$ for some splitting test involving X_i. Indeed, the splitting test "looks-ahead" to the best multiple linear regressions after the split on X_i is performed, while the regression step does not. A fairer comparison would be growing the tree at a further level in order to base the computation of $\rho(X_i,Y)$ on the best multiple linear regressions after the regression step on X_i is performed (see Figure 2).

Let t' be the child of the regression node t, and suppose that it performs a splitting test. The best splitting test in t' can be chosen on the basis of $\sigma(X_j,Y)$ for all possible variables X_j, as indicated above. Then $\rho(X_i,Y)$ can be defined as follows:

$\rho(X_i,Y) = min \{R(t), \sigma(X_j,Y)$ *for all possible variables* $X_j \}$.

The possibility of statistically identical regression models associated to the children of t' may also occur in this case. When this happens, the splitting node is replaced by another regression node t' where the straight-line regression model is the same as that in the children of the splitting node. Therefore, in this special case $\rho(X_i,Y)$ can be defined as follows:

$$\rho(X_i,Y) = min \{ R(t), R(t')\}.$$

Having defined both $\rho(X_i,Y)$ and $\sigma(X_i,Y)$, the criterion for selecting the best node is fully characterized as well. At each step of the model tree induction process, SMOTI chooses the apparently most promising node, according to a greedy strategy. A continuous variable selected for a regression step is no longer considered for regression purposes, so that it can appear only once in a regression node along a path from the root to a leaf.

In SMOTI five different stopping criteria are implemented. The first uses the partial F-test to evaluate the contribution of a new independent variable to the model [3]. The second requires the number of cases in each node to be greater than a minimum value. The third stops the induction process when all continuous variables along the path from the root to the current node are used in regression steps and there are no discrete variables in the training set. The fourth creates a leaf if the error in the current node is below a fraction of the error in the root node, as in [14, p. 60]. Finally, the fifth stops the induction process when the *coefficient of determination* is greater than a minimum value [16, pp. 18-19]. This coefficient is a scale-free one-number summary of the strength of the relationship between independent variables in the actual multiple linear model and the response variable.

2.1 Computational complexity

The computational complexity of adding a splitting node t to the tree depends on the complexity of a splitting test selection in t multiplied by the complexity of the best regression step selection in the children nodes t_R and t_L. On the contrary, the computational complexity of adding a regression node t depends on the complexity of a regression step selection in t multiplied by the complexity of the best splitting test in its child t'.

A splitting test can be either continuous or discrete. In the former case, a threshold α has to be selected for a continuous variable. Let N be the number of examples in the training set, then the number of distinct thresholds can be N-1 at worst. They can be determined after sorting the set of distinct values. If m is the number of independent variables, the determination of all possible thresholds has a complexity $O(mNlogN)$ when an optimal algorithm is used to sort the values. For each of the $m(N$-1) thresholds, SMOTI finds the best straight-line regression at both children, which has a complexity of $m(N$-1) in the worst case. Therefore, the splitting test has a complexity $O(mNlogN+m^2(N-1)^2)$, that is $O(m^2N^2)$. Similarly, for a discrete splitting test, the worst case complexity is $O(mk^2)$ where k is the maximum number of distinct values of a discrete variable. The selection of the best discrete splitting test has a complexity $O(m^2k^2N)$. Therefore, finding the best continuous or discrete splitting node has a complexity $O(m^2N^2 + m^2k^2N)$, and under the reasonable assumption that $k^2 \leq N$, that is, the number distinct values of a discrete variable is less then the square root of the number of cases, the worst case complexity is $O(m^2N^2)$.

The selection of the best regression step requires the computation of a straight-line regression, whose complexity is linear in N, for each of the m variables. Moreover, for each straight-line regression, a splitting test is required. As reported above, the splitting test has a worst case complexity of $O(m^2N^2)$. Therefore, the selection of the best regression step has a complexity $O(mN+m^3N^2)$, that is $O(m^3N^2)$.

The above results lead to an $O(m^3N^2)$ worst case complexity for the selection of any node (splitting or regression). It can be proven that RETIS has the same complexity for node selection, although RETIS does not select a subset of variables to solve collinearity problems. TSIR, which adopts a v-fold cross-validation strategy without look-ahead is more efficient, with a complexity $O(mvN)$ for regression nodes and $O(mvN^2)$ for splitting nodes. However, the model that TSIR considers at the children of a discrete splitting node during its evaluation is the sample mean and not a linear regression, which means that it suffers from the problems of adopting a heuristic evaluation function which is not coherent with the models associated to the leaves.

3. An empirical evaluation of SMOTI

SMOTI has been empirically evaluated both on artificially generated data and on datasets typically used in the evaluation of regression and model trees. Each dataset is analyzed by means of a 10-fold cross-validation. Performances are evaluated on the basis of the average MSE on the ten hold-out sets.

For pairwise comparison with M5', which is the state-of-the-art model tree induction system, the non-parametric Wilcoxon two-sample paired signed rank test is used [11], since the number of folds (or "independent" trials) is relatively low and does not justify the application of parametric tests, such as the t-test. To perform the test, we assume that the experimental results of the two compared methods are independent pairs of sample data $\{(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n)\}$. We then rank the absolute value of the differences $u_i - v_i$. The Wilcoxon test statistics W^+ and W^- are the sum of the ranks from the positive and negative differences, respectively. We test the null hypothesis H_0: "no difference in distributions" against the two-sided alternative H_a: "there is a difference in distributions". More formally, the hypotheses are: H_0: "$\mu_u=\mu_v$" against H_a: "$\mu_u\neq\mu_v$". Intuitively, when $W^+ >> W^-$ and viceversa, H_0 is rejected. Whether W^+ should be considered "much greater than" W^- depends on the significance level α. The basic assumption of the statistical test is that the two populations have the same continuous distribution (and no ties occur). Since, in our experiments, u_i and v_i are MSE, $W^+ >> W^-$ implies that the second method (V) is better than the first one (U). In all experiments reported in this empirical study, the significance level α used in the test is set to 0.05.

3.1 Experiments on artificial data sets

SMOTI was initially tested on artificial datasets randomly generated for model trees with both regression and splitting nodes. These model trees where automatically built for learning problems with nine independent variables (five continuous and four discrete), where discrete variables take values in the set {A,B,C,D,E,F,G}. The model tree building procedure is recursively defined on the maximum depth of the tree to be generated. The choice of adding a regression or a splitting node is random and depends on a parameter $\theta\in[0,100]$: the probability of selecting a splitting node is θ; conversely, the probability of selecting a regression node is $(1-\theta)$. In the experiments reported in this paper θ is fixed to 0.5, while the depth is varied from four to nine. Fifteen model trees are generated for each depth value, for a total of ninety trees. Sixty data points are randomly generated for each leaf, so that the size of the data set associated with a model tree depends on the number of leaves in the tree itself. Data points are generated by considering the various constraints associated to both splitting nodes and regression nodes. A normally distributed error is added to each model in order to introduce the noise effect. In all experiments, the thresholds for stopping criteria are fixed as follows: the significance level α used in the F-test is set to 0.075, the minimum number of cases falling in each internal node must be greater than the square root of the number of cases in the entire training set, the error in each internal node must be greater than the 0.01% of the error in the root node, the coefficient of determination in each internal node must be below 0.99.

The results of Wilcoxon signed rank test on the accuracy of the induced model tree are reported in Table 1. Three main conclusions can be drawn from experimental results. First, SMOTI performs generally better than M5′ on data generated from model trees where both local and global effects can be represented. Second, by increasing the depth of the tree SMOTI tends to be more accurate than M5′. Third, when SMOTI performs worse than M5′, this is due to relatively few hold-out sets in the cross validation, so that the difference is never statistically significant in favor of M5′. These conclusions are at variance with respect to those reported for the first version of SMOTI [7,8], where we observed that SMOTI performed better than M5 when split nodes were slightly preferred to regression nodes by means of a weighting factor not used in this experimentation. Moreover, in previous experiments we observed that the depth of the tree had no clear effect on the predictive accuracy of the induced model tree. We attribute the better performance shown by the current version of SMOTI to the extensions/revisions of the method described in Section 2.

Table 1. Results of the Wilcoxon signed rank test on the accuracy of the induced model trees. The statistically significant values (p-value≤α/2) are in boldface. All statistically significant values are favorable to SMOTI.

Tree\depth	*4*	*5*	*6*	*7*	*8*	*9*
1	**0.0019**	0.3750	0.2750	**0.0019**	**0.0039**	**0.0097**
2	**0.0019**	**0.0190**	0.0839	**0.0019**	**0.0019**	**0.0019**
3	0.0645	**0.0019**	0.0644	**0.0019**	**0.0019**	**0.0019**
4	**0.0019**	**0.0058**	0.7695	0.3750	**0.0136**	**0.0019**
5	**0.0019**	0.7695	0.4316	**0.0019**	**0.0019**	0.0644
6	**0.0019**	0.12	**0.0019**	**0.0058**	0.0839	**0.0019**
7	0.8457	0.2754	**0.0136**	0.4922	0.0839	**0.0019**
8	**0.0234**	0.375	**0.0039**	0.2324	**0.0019**	**0.0019**
9	0.0644	**0.0019**	**0.0019**	**0.0019**	**0.0019**	**0.0019**
10	0.2754	0.1934	**0.0019**	**0.0097**	**0.0019**	**0.0039**
11	**0.0136**	0.2969	**0.0097**	0.0839	0.6953	**0.0195**
12	**0.0273**	**0.0019**	**0.0019**	**0.0019**	0.4316	**0.0019**
13	**0.0019**	**0.0019**	**0.0195**	**0.0019**	0.1309	**0.0039**
14	**0.0019**	**0.0019**	**0.0039**	**0.0019**	**0.0019**	**0.0019**
15	0.3750	0.1934	**0.0039**	0.1934	**0.0058**	**0.0039**

3.2 Experiments on benchmark data sets

SMOTI was also tested on twelve data sets taken from either the UCI Machine Learning Repository or the site of WEKA (www.cs.waikato.ac.nz/ml/weka/) or the site of HTL (www.niaad.liacc.up.pt/~ltorgo/Regression/DataSets.html) They are listed in Table 2 and have been used as benchmarks in related studies on regression/model trees.

In all experimental results reported below, the thresholds for the stopping criteria are set to the same values used in the experiments on artificial datasets, except for the coefficient of determination which is set to 0.9.

Experimental results are reported in Table 3, where SMOTI is compared to M5' on the basis of the average MSE. The Wilcoxon test statistics W^+ (W^-) is the sum of the ranks from the positive (negative) differences between M5' and SMOTI. Therefore, the smaller W^+ (W^-), the better for SMOTI (M5'). As in the previous experimentation, differences are statistically significant when the p-value ≤ α/2.

Table 2. Datasets used in the empirical evaluation of SMOTI.

DATASET	No. Cases	No. Attributes	Continuous	Discrete
Abalone	4177	10	9	1
Auto-Mpg	392	8	5	3
Auto-Price	159	27	17	10
Bank8FM	4499	9	9	0
Cleveland	297	14	7	7
Delta Ailerons	7129	6	6	0
Delta Elevators	9517	7	7	0
Housing	506	14	14	0
Kinematics	8192	9	9	0
Pyrimidines	74	28	28	0
Triazines	74	61	61	0
Wisconsin Cancer	186	33	33	0

Table 3. Results of the Wilcoxon signed rank test on the accuracy of the induced models. The best *W* value is in boldface, while the statistically significant values ($p\leq\alpha/2$) are in italics.

Dataset	*SMOTI Avg. MSE*	*M5'Avg. MSE*	*W+*	*W-*	*p-value*	*Result*
Abalone	2.53637	2.77242	**14**	41	0.1934	=
Auto-Mpg	3.14938	3.20106	**21**	34	0.5566	=
Auto-Price	2246.039	2358.819	**23**	32	0.6953	=
Bank8FM	0.03833	0.04099	**9**	46	0.064	=
Cleveland	1.31603	1.24963	40	**15**	0.2324	=
Delta Ailerons	0.000232	0.0002	21.5	**14.5**	0.6404	=
Delta Elevators	0.00476	0.00163	41	**14**	0.1934	=
Housing	3.58	4.27927	**8**	47	0.048	=
Kinematics	0.1581	0.194737	**1**	54	*0.0039*	+
Pyrimidines	0.10566	0.09279	30	**25**	0.8457	=
Triazines	0.2017	0.15503	49	**6**	*0.02*	-
Wisconsin Cancer	51.41376	45.40644	33	**22**	0.625	=

Differently from artificially generated data, SMOTI does not exhibit a clear superiority with respect to M5', although results are still good and better than those reported in [7,8]. A deeper analysis of the experimental results has highlighted that for some training sets, the thresholds defined for the stopping criteria prevented SMOTI from growing model trees more accurate than those built by M5'. This problem cannot be straightforwardly solved by defining higher thresholds, since that would lead to data overfitting problems.

The interesting aspect of this experimentation is that for some datasets SMOTI detected the presence of interesting patterns that no previous study on model trees had ever revealed. In the following, we report some of them, thus proving another desirable characteristics of the system, which is easy interpretability of the induced model trees.

Abalone. Abalones are marine crustaceans, whose age can be determined by counting under the microscope rings in the cross section of the shell. Other measurements, which are easier to obtain, can be used to predict the age. For all ten cross-validated training sets, SMOTI builds a model tree with a regression node in the root. The straight-line regression selected at the root is almost invariant for all model trees and expresses a linear dependence between the number of rings (dependent variable) and the shucked weight (independent variable). This is a clear example of global effect, which cannot be grasped by examining the nearly 350 leaves of a model tree induced by M5' on the same data. Interestingly, the child of the root is always a splitting test on the whole weight, or, more precisely, on the residuals of the whole weight once the effect of the shucked weight has been removed. Unfortunately, this stability of the tree structure occurs only at the root and its child.

Auto-Mpg. The data concerns city-fuel consumption in miles per gallon. For all ten cross-validated training sets, SMOTI builds a model tree with a discrete split test in the root. The split partitions the cars in two subgroups, one whose *model year* is between 70 and 77 and the other whose *model year* is between 78 and 82. That can be easily explained with the measures for energy conservation prompted by the 1973 OPEC oil embargo. Indeed, in 1975 the U.S. Government set new standards on fuel consumption for all vehicles. These values, known as C.A.F.E. (Company Average Fuel Economy) standards, required that by 1985 automakers doubled average new car fleet fuel efficiency. These standards came into force only in 1978 and model trees induced by SMOTI capture this temporal watershed. Moreover, in the case of *model year* between 70 and 77, SMOTI performs another discrete splitting test on the number of cylinders, while in the case of *model year* between 78 and 82 SMOTI introduces a regression step generally involving the variable *weight*. Also this difference seems reasonable, since it captures the different technologies (e.g., lightweight materials) adopted by automakers before and after the introduction of C.A.F.E. standards. Differently from SMOTI, M5' performs a first continuous splitting on the variable *displacement* (≤191 vs. >191) and a second split on the variable *horsepower* for both left and right child. A test on the variable *model year* appears only at lower levels.

4. Conclusions

SMOTI is a TDIMT system characterized by the induction of model trees with two types of nodes (regression and splitting nodes). This allows SMOTI to discover both global and local effects of variables in the various regression models at leaves. SMOTI also presents other three advantages with respect to other systems reported in the literature. First, it defines the best partitioning of the feature space coherently with respect to the model tree being built. Second, it provides a solution to the problems of collinearity. Third, its results are easily interpretable. It is noteworthy that the implementation of SMOTI in the data mining system KDB2000 [1] access data stored in a relational DBMS through ODBC. Currently, this is the only TDIMT system with this characteristic.

As future work, we intend to investigate how model trees induced by SMOTI compare to other approaches, such as the hierarchical mixture-of-experts architecture [4] and the support vector machines [9]. Moreover, we intend to study the *a posteriori* simplification (pruning) of model trees with both regression nodes and splitting nodes in order to offer a solution to data overfitting problems.

Acknowledgments

This work has been supported by the annual Scientific Research Project "Scoperta di conoscenza in basi di dati: metodi e tecniche efficienti e robuste per dati complessi" Year 2002 funded by the University of Bari.

References

1. Appice, A., Ceci, M., and Malerba D.: KDB2000: Uno strumento per la scoperta della conoscenza nelle basi di dati. *Atti del Decimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati SEBD'2002*, 417-421, 2002.
2. Breiman L., Friedman J., Olshen R. and Stone J.. *Classification and regression tree*. Wadsworth & Brooks, 1984.
3. Draper N.R. and Smith H.: *Applied regression analysis*, John Wiley & Sons, 1982.
4. Jordan M.I., Jacobs R.A.: Hierarchical mixture of experts and the EM algorithms. *Neural Computation, 6*, pages 181-214, 1994.
5. Karalic A.: Linear regression in regression tree leaves. In *Proceedings of ISSEK '92 (International School for Synthesis of Expert Knowledge)*, Bled, Slovenia, 1992.
6. Lubinsky D.: Tree Structured Interpretable Regression. In Fisher D. & Lenz H.J. (Eds.), *Learning from Data*, Lecture Notes in Statistics, 112, Springer, pages 387-398, 1994.
7. Malerba, D., Appice, A., Bellino, A., Ceci, M. and Pallotta D.: Stepwise Induction of Model Trees, in F. Esposito (Ed.), *AI*IA 2001: Advances in Artificial Intelligence*, Lecture Notes in Artificial Intelligence, 2175, Springer, Berlin, Germany, 2001.
8. Malerba, D., Appice, A., Ceci, M. and Monopoli, M.: Trading-off local versus global effects of regression nodes in model trees, in H.-S. Hacid, Z.W. Ras, D.A. Zighed & Y. Kodratoff (Eds.), *Foundations of Intelligent Systems, 13th International Symposium, ISMIS'2002*, Lecture Notes in Artificial Intelligence, 2366,393-402, Springer, Berlin, Germany, 2002.
9. Mangasarian O. L. & Musicant. D. R.: Robust Linear and Support Vector Regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 22. pages 950-955, 2000.
10. Mehta M., Agrawal R. and Rissanen J.. SLIQ: A fast scalable classifier for data mining. In *Proceedings of the Fifth International Conference on Extending Database Technology*, 1996.
11. Orkin. M.. Drogin. R.. *Vital Statistics*. McGraw Hill. New York . 1990.
12. Quinlan J. R.. Learning with continuous classes. In Adams & Sterling (Eds.), *Proceedings AI'92, World Scientific*. pages 343-348, 1992.
13. Torgo L.. Functional Models for Regression Tree Leaves. In D. Fisher (Ed.), *Proceedings of the Fourteenth International Conference (ICML '97)*, Nashville, Tennessee, pages 385-393, 1997.
14. Torgo, L.: *Inductive Learning of Tree-based Regression Models*, Ph.D. Thesis, Department of Computer Science, Faculty of Sciences, University of Porto. 1999.
15. Y. Wang & I.H. Witten. Inducing Model Trees for Continuous Classes. In M. van Someren. & G. Widmer (Eds.),*Poster Papers of the 9th European Conference on Machine Learning (ECML 97)*,Prague,Czech Republic,pages 128-137, 1997.
16. Weisberg S. *Applied regression analysis*, 2nd edn. New York: Wiley, 1985.