

Non-parametric methods for data mining applications

Metodi non-parametrici per il Data Mining

Agostino Di Ciaccio

Università degli Studi di Roma “La Sapienza”

agostino.diciaccio@uniroma1.it

Giorgio Eduardo Montanari

Università degli Studi di Perugia

giorgio@stat.unipg.it

Riassunto: Nel presente lavoro vengono considerati alcuni tipici problemi inferenziali che si incontrano nelle applicazioni di Data Mining. In tale contesto i dati spesso non soddisfano le usuali ipotesi statistiche che giustificano l'utilizzo di tecniche standard basate su modelli parametrici e test di significatività. Attraverso un ampio studio empirico si esplorano invece le potenzialità dei metodi non parametrici, ed in particolare del DART, ai fini della previsione di singoli valori di una variabile oggetto di indagine e della stima della media. I risultati ottenuti documentano una notevole efficienza del DART in termini di errore quadratico medio, anche in presenza di campioni autoselezionati, relazioni non lineari e asimmetrie nelle variabili utilizzate.

Keywords: Finite population, DART, Prediction

1. Introduction

When it is necessary to analyse large administrative archives for statistical purposes, the classical inference procedures show several limits because of the particular characteristics of the analysed data. The following is a list of some of them:

- 1) Data are collected for aims which are not related to the survey and therefore often they do not constitute a census or a random sample.
- 2) There are many missing or incorrect data.
- 3) The variables that are included in the archive are not necessarily the most useful for the aims of the survey, but they are however the only variables available.
- 4) Very often, available data are both qualitative and quantitative.
- 5) The observed variables can be numerous, i.e. some hundreds, without a sufficient knowledge of the phenomena allowing the selection of a sub-set of variables or the choice of a particular model for the analysis.
- 6) Usually, a priori information is not available on the phenomena to be analysed.
- 7) The available records can be several millions or more.

These characteristics are the most frequent ones in current data-bases of a company. A common request that we have to face in this respect is to “mine the data”, in order to search for unknown patterns which might be of some interest for the company that manages the archives. The new problem the statistician has to deal with is that the basic hypotheses of the classical inferential procedures do not hold (see Coppi 2002).

In this paper we concentrate our attention on estimation problems for a quantitative variable Y taking into account a set of explanatory quantitative variables X_j ($j=1,\dots,k$) from a data-set that shows some of the characteristics quoted above. To this purpose, we do not try to look for “the Model” that describes the relation between the phenomena under consideration, since this would not be theoretically justified by the characteristics of these data. In fact, for very large archives, classical procedures of selection of the model turn out to be somewhat ineffectual. On the other hand, the practice that consists in extracting a sample from the archive and performing the analysis on such a single sample constitutes, from a statistical point of view, an absurdity and it cannot be considered as a valid solution. It is worth also to note that the common practice of carrying out reiterated tests on the same data leads to conclusions for which the true degree of reliability is not known. So, such a practice is not justified in the presence of a large amount of data. Furthermore, when rigid functional shapes (e.g. linear) are used, it is easy to object that such shapes turn out to be arbitrary and seldom suitable to describe complex relations between phenomena that are, usually, unknown.

In this paper, through a simulation approach we analyse data in which standard distributional hypotheses (e.g. normality) are not met. Moreover, we consider the effects of selection biases and non-linearity of the relations on the performances of the analysed methods.

2. Data Mining applications

We consider two estimation problems which are strongly linked: the estimation of the average of a variable Y and the estimation of the unobserved value of Y in a given unit, using the values of a set of auxiliary variables known for all population units. To that purpose, we assume that a sample of units from the population provides values for both the dependent variable Y and the explanatory or predictor variables.

There are many situations in which problems of this kind are common, as in the following cases:

- We want to analyse some characteristics of the visitors of a WEB site. For the registered visitors it is possible to have detailed information (they constitute the sample), but for the great majority of the visitors we only know the visited pages, the origin IP address, the frequency of access to the site (through the cookies). Obviously, the sample is self-selected.
- We want to analyse the behaviour of the customers of a supermarket. A set of consumers registered themselves for being able to use the fidelity card (they are the sample), and therefore several informative traits are available for them. Indeed, information is available on each purchase but for the registered consumers information is also available on their purchases carried out in a month or a year. Also in this case the sample is self-selected.
- We want to analyse some “sensible” information (income, political opinions, sexual habits...) on a population for which only administrative data are available. A sampling is carried out in order to obtain all the information on a small number of individuals. Even in this case, the sample turns out, a posteriori, almost certainly incomplete and biased because of the nature of the requested information.

In section 4 we examine the case in which we want to estimate the average (or the amount) of a variable Y , while in section 5 we consider the problem of the estimation of unobserved values of Y . In both cases we compare an approach based on the non-parametric method DART (Friedman, 1996) with more traditional statistical methods as suggested in Di Ciaccio & Montanari (2001).

DART is the name of a local learning method based on recursive covering. This approach combines the strategies of two different local learning procedures, CART and K-nearest neighbour (K-NN). The first method identifies a partition of the predictors space into a small number of regions while with the second method a large number of overlapping regions are induced.

Combining these two procedures, DART first determines a set of overlapping local regions that cover the predictor-space; then, within each region, it models simple fitting functions of Y . The parameter *dof* defines the degree of overlapping of the regions. The parameter *lin* defines the fitting function (*lin*=3 implies local linear model with linear combination splits), while *kd* is the minimum number of sample observations in each terminal region.

With eager training there are two prediction strategies named *ANS* and *AVE*. The first uses the local fitting function of the region in which the prediction point $\mathbf{x}$ (the vector with the values of the k explanatory variables) is most centred. The second prediction strategy (computational-time expensive but also much more effective) uses a weighted average of the local fitting functions (at the point $\mathbf{x}$) of all regions that contain $\mathbf{x}$. The weight given to the predicting function of each region is proportional to the distance between its centre and the prediction point $\mathbf{x}$.

3. The simulation framework

Consider a finite population of N units and assume that the values $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ of an explanatory vector $\mathbf{X}$ are known for all elements of the population but the values of a survey variable Y are observed on a sample drawn from it.

Two types of problems are considered in this simulation. The first one is the estimation of the mean of Y in the whole population and in subpopulations of interest. The second problem is the prediction of the value of Y in population units of interest. To those purposes, let $Y = f(\mathbf{x}) + \varepsilon$ be the working model for the conditional distribution of Y given $\mathbf{x}$, where ε is a random variable whose expected value is equal to zero. Once f has been estimated from sample data, the predicted value of Y for the i -th unit of the population is given by $\hat{f}(\mathbf{x}_i)$, where $\hat{f}$ is the estimate of f . Then, a naïve estimator of the population mean, $\bar{Y} = N^{-1} \sum_{i=1}^N Y_i$, is given by

$$(1) \quad \hat{\bar{Y}} = N^{-1} \sum_{i=1}^N \hat{f}(\mathbf{x}_i).$$

Several non-parametric methods have been proposed for estimating f such as kernel regression methods, K-NN and, recently, local polynomial regression methods (Breidt and Opsomer, 2000). These methods run into difficulties when $\mathbf{X}$ is multidimensional because of the curse of dimensionality, i. e. the sparseness of the sample points $\mathbf{x}$ in the predictor space. To avoid such a problem, a possible alternative is the use of DART to estimate f . Considering this approach and using (1), the performance of the resulting

DART estimator, $\hat{Y}_{DT}$, has been explored through a simulation study and was compared to the performances of the sample mean $\bar{y}$, the post-stratified estimator $\hat{Y}_{PS} = \sum_h W_h \bar{y}_h$, where W_h is the post-stratum weight and $\bar{y}_h$ is the post-stratum sample mean, and the linear regression estimator, $\hat{Y}_{RG}$, where f is estimated through a linear regression parametric model using $\mathbf{X}$ as the set of explanatory variables, i.e. $f(\mathbf{x}) = \beta' \mathbf{x}$.

We considered the following generator model of finite populations

$$Y = a(c_0 + \sum_{j=1}^k c_j X_j) + \sum_{j=1}^k \beta_j (X_j - \bar{X}_j)^2 + \varepsilon,$$

where

- k is the dimension of $\mathbf{X}$, i.e. the number of auxiliary variables;
- X_j ($j=1, \dots, k$) is the j -th auxiliary variable generated from a Beta distribution with parameters (g_j, h_j) , where the latter are drawn from a uniform distribution in the interval $[2, 10]$;
- a , c_0 , c_j and β_j ($j=1, \dots, k$) are randomly drawn from uniform distributions in the intervals $[0, 4]$, $[-5, 5]$, $[-5, 5]$ and $[-50, 50]$, respectively;
- ε is the noise generated from a gaussian distribution $N(0, \sigma_\varepsilon^2)$.

The above definitions allow us to generate additive models with different levels of non-linearity in the relation between Y and $\mathbf{X}$ and different levels of asymmetry of the predictor variables. For an easier comparability, the systematic component of the model (the signal) has been scaled so that its variance, σ_s^2 , is equal to 9.0. The variance of the noise, σ_ε^2 , was set equal to 2.25, in order to obtain a signal to noise (s/n) ratio 4:1; thus, the signal explains 80% of the total variability of Y . We chose two sample sizes, i.e. $n=200$ and $n=2000$, and the number of auxiliary variables k was set equal to 3.

Several simulations were carried out as follows. For $n=200$, as many as 500 different finite populations of size 5000 were generated first drawing the values of a , c_0 , c_j , β_j , g_j and h_j ; then, conditionally on the latter, drawing the 5000 values of each X_j ($j=1, \dots, k$) and of ε ; after all, computing Y . For $n=2000$, a number of 250 different finite populations of size 20,000 were generated as above.

For each population, 500 samples of size n were extracted and for each of them the predicted value of Y was computed for all the units in the population through the DART procedure. The latter has been implemented using the AVE option and keeping fixed the values of the tuning parameters as follows: $kd=15$, $dtf=0.30$, $lin=3$, $rdg=0.01$ (see Friedman, 1996b). Then, the estimators of the population mean, $\hat{Y}_{DT}$, $\hat{Y}_{PS}$, $\hat{Y}_{RG}$ and $\bar{y}$ have been computed. In particular, to compute the post-stratified estimator, we defined 8 post-strata partitioning the range of each predictor variable in “below its (realised) mean” and “above its mean”. Post-strata were used also as subpopulations whose means are to be estimated as well. In the latter case, the sample mean is compared with the regression and DART estimators, both computed by means of (1) with the sums running only over the subpopulation units and replacing N with the size of the subpopulation. Thus, to complete one simulation, each estimator was computed 250,000 times. At last, for each population, we computed the mean of the 500 computed values of each estimator and their Mean Square Error (MSE) and the percent contribution of the bias to the MSE.

Three sampling procedures were considered: *a*) simple random sampling, *b*) biased sampling with respect to X_1 , *c*) biased sampling with respect to Y .

Simple random sampling is equivalent to a missing at random non response mechanism and for such a sampling scheme $\bar{y}$ is design unbiased, $\hat{Y}_{PS}$ and $\hat{Y}_{RG}$ are approximately design unbiased, i.e. the bias is of order n^{-1} (Särndal *et al.*, 1992), while that is not assured for $\hat{Y}_{DT}$, although it is expected to be consistent.

Biased sampling with respect to X_1 was performed assigning a chance of selection five times greater to the population units with a value of X_1 over its mean. This kind of selection is equivalent to a missing at random non response mechanism conditional on the value of X_1 . Thus, it is an ignorable selection scheme provided that the values of X_1 are known for all population units (Sudgen & Smith, 1984). For such a sampling scheme, the post-stratified estimator is unbiased for the population mean.

Biased selection with respect to Y was performed assigning a chance of selection five times greater to the population units with a value of Y over its mean. In this case the sampling scheme is non ignorable but informative (Sudgen & Smith, 1984) and no estimator is unbiased.

Biased selections were considered to simulate self selected samples or self selected non response as it is often the case in data mining problems.

The generator model produces data with different levels of non-linearity in the relation between Y and the predictor variables. To measure the degree of linearity in a data-set we used the determination index R^2 , computed on the whole data-set; a true linear relation corresponds to a value of R^2 near its maximum, i.e. 0.80 for a s/n ratio 4:1.

4. Simulation results for estimating the population mean

Simulation results for the estimation of the population mean are summarized in Table 1 and Table 2. For each population, we computed the empirical Mean Squared Error (MSE) and the percent contribution of the bias to the MSE of each estimator. For the three sampling schemes, Table 1 compares the efficiencies of the estimators by means of the scaled MSE, having set equal to 100 the MSE of the sample mean. The values are averaged over populations with a value of R^2 belonging to selected intervals.

As expected, the efficiency of the regression estimator is highly related to the value of R^2 in the population, and it is more efficient than the non-parametric estimator $\hat{Y}_{DT}$ only when R^2 approaches the maximum value, i.e. the relation between Y and X is almost linear. This is a common result in comparing parametric and non-parametric methods. The efficiency of the post-stratified estimator, which can be seen as a naïve non-parametric estimator, is related to R^2 too, but it depends heavily on the fixed in advance number of post-strata, and on the sparseness of sample data. On the contrary, the efficiency of the DART estimator is almost independent of the value of R^2 within the class of models considered in this empirical study. A picture of the above outcomes is in Figure 1 that, for the case $n=200$, shows the scatter plot of the 500 MSE (absolute) values of each estimator against the value of R^2 of the corresponding populations. Very similar plots were obtained for the other simulations.

Under simple random sampling the contribution of the bias to the MSEs is negligible for all estimators as expected, apart the DART one that shows some small bias. Under biased sampling with respect to X_1 , the sample mean is biased, in particular for linear populations; the regression estimator is biased as well, in particular for non linear populations.

Table 1: Means of scaled MSEs and percent contribution of the bias to the MSE of post-stratified (PS), regression (RG) and DART (DT) estimators for intervals of R^2 values under simple random sampling and biased sampling with respect to X_1 and Y .

Estimator	$n = 200$							$n = 2000$						
	Scaled MSE			Contribution of Bias (% of MSE)				Scaled MSE			Contribution of Bias (% of MSE)			
	$\hat{Y}_{PS}$	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$	$\bar{y}$	$\hat{Y}_{PS}$	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$	$\hat{Y}_{PS}$	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$	$\bar{y}$	$\hat{Y}_{PS}$	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$
R^2 range	Simple random sampling													
0-10	102	99.0	34.1	0.1	0.1	0.7	6.6	98.2	95.2	37.2	0.6	0.5	0.7	27.3
10-20	96.7	88.2	33.2	0.1	0.1	0.8	6.0	95.8	87.9	31.8	0.7	0.7	1.1	21.8
20-30	87.4	75.9	32.5	0.2	0.1	0.6	5.2	86.8	76.1	31.5	0.7	0.8	0.7	21.8
30-40	82.1	67.3	31.1	0.2	0.1	0.6	4.0	81.0	66.7	32.5	0.7	0.9	0.6	14.9
40-50	74.3	56.4	29.5	0.1	0.2	0.7	2.4	72.2	55.9	29.3	0.4	0.3	0.4	12.7
50-60	66.6	45.9	29.5	0.2	0.1	0.5	1.6	65.7	45.7	27.2	0.7	0.9	0.9	12.9
60-70	59.7	35.9	28.6	0.1	0.1	0.4	1.0	59.3	36.5	25.0	0.4	0.6	0.9	8.7
70-80	54.3	26.1	27.5	0.2	0.1	0.3	0.4	52.6	24.7	24.1	0.7	1.4	0.6	6.0
R^2 range	Biased sampling with respect to X_1													
0-10	169	448	64.4	20.0	2.6	53.9	8.6	90.0	1392	37.8	53.8	0.6	85.4	24.2
10-20	104	284	42.1	49.3	1.9	56.3	7.9	64.4	475	21.1	67.6	1.0	78.9	18.5
20-30	56.0	158	21.3	70.6	1.8	57.4	7.0	24.1	908	8.6	83.9	4.6	87.6	22.2
30-40	50.2	104	20.2	70.0	1.8	46.5	7.2	3.0	96.7	1.3	97.6	1.1	71.4	13.3
40-50	35.3	70.1	15.0	77.7	2.0	47.3	5.8	2.3	30.5	0.8	98.3	0.5	83.6	12.3
50-60	25.8	49.0	10.8	82.0	1.6	49.9	10.8	3.0	40.4	1.4	97.5	0.3	86.3	11.6
60-70	25.4	30.6	11.4	78.2	1.6	47.0	2.9	11.9	61.3	5.0	89.0	0.8	85.3	6.8
70-80	13.8	8.3	6.2	87.2	2.3	37.6	2.7	10.1	29.7	4.7	89.6	0.5	82.8	5.3
R^2 range	Biased sampling with respect to Y													
0-10	97.9	96.1	15.3	98.6	98.5	98.5	92.9	97.4	95.0	9.1	99.9	99.9	99.9	99.2
10-20	89.2	83.2	15.2	98.5	98.2	98.2	93.4	92.7	84.5	10.8	99.8	99.8	99.8	99.4
20-30	75.1	66.2	13.2	98.6	98.1	97.8	92.9	76.7	65.9	8.7	99.9	99.8	99.8	99.1
30-40	63.7	51.6	12.4	98.6	97.8	97.3	92.7	61.7	47.1	7.8	99.9	99.8	99.7	99.2
40-50	54.1	39.1	11.0	98.6	97.4	96.6	91.5	52.9	37.9	7.2	99.9	99.7	99.7	99.1
50-60	44.0	27.1	9.6	98.7	96.9	95.8	91.0	42.0	25.2	6.7	99.9	99.7	99.5	99.0
60-70	36.0	18.5	8.1	98.7	96.4	94.9	89.5	35.3	17.4	6.9	99.9	99.6	99.5	99.0
70-80	26.9	9.1	6.8	98.7	95.8	93.1	89.1	25.3	7.2	6.0	99.9	99.6	99.2	99.0

Table 2: Average of mean scaled MSEs over post-strata of regression (RG) and DART (DT) estimators for intervals of R^2 values.

Sampling	$n = 200$						$n = 2000$					
	Simple random		Biased w.r. to X_1		Biased w.r. to Y		Simple sampling		Biased w.r. to X_1		Biased w.r. to Y	
	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$	$\hat{Y}_{RG}$	$\hat{Y}_{DT}$
R^2 range												
0-10	69,7	24,5	118	24,7	96,0	18,4	236	22,7	664	23,3	662	23,5
10-20	87,1	27,5	136	27,2	93,4	21,2	517	25,5	1274	25,3	1352	24,4
20-30	75,1	30,0	142	28,9	88,2	22,2	340	28,0	1112	28,1	1093	27,8
30-40	66,3	32,6	103	29,9	81,2	23,5	177	29,2	579	30,0	646	29,5
40-50	63,2	36,3	97,9	32,9	75,5	29,2	185	35,3	717	35,5	672	34,8
50-60	51,9	39,4	82,9	34,9	64,0	31,9	166	38,1	616	37,7	607	37,3
60-70	42,5	42,9	62,0	37,6	53,6	33,9	121	41,5	457	41,5	460	41,2
70-80	26,9	42,7	35,7	36,3	34,4	32,6	48,6	41,2	158	41,0	163	40,4

Figure 1: *MSEs of sample mean, post-stratified, regression and DART estimators with respect to the R^2 values (200 sample cases).*

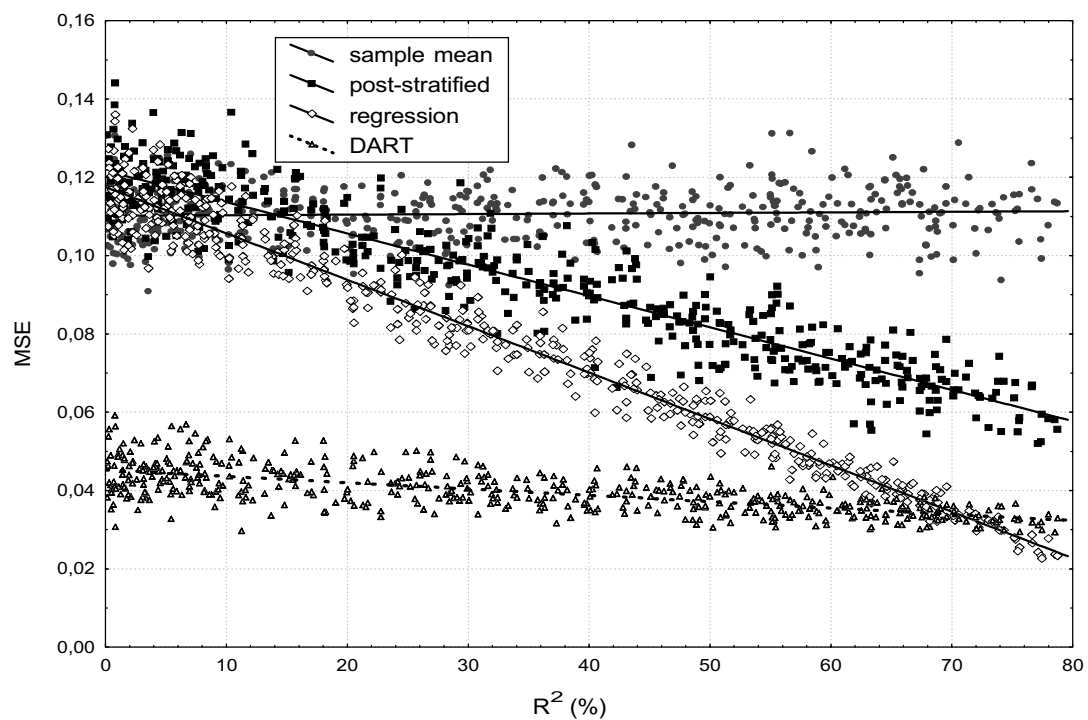
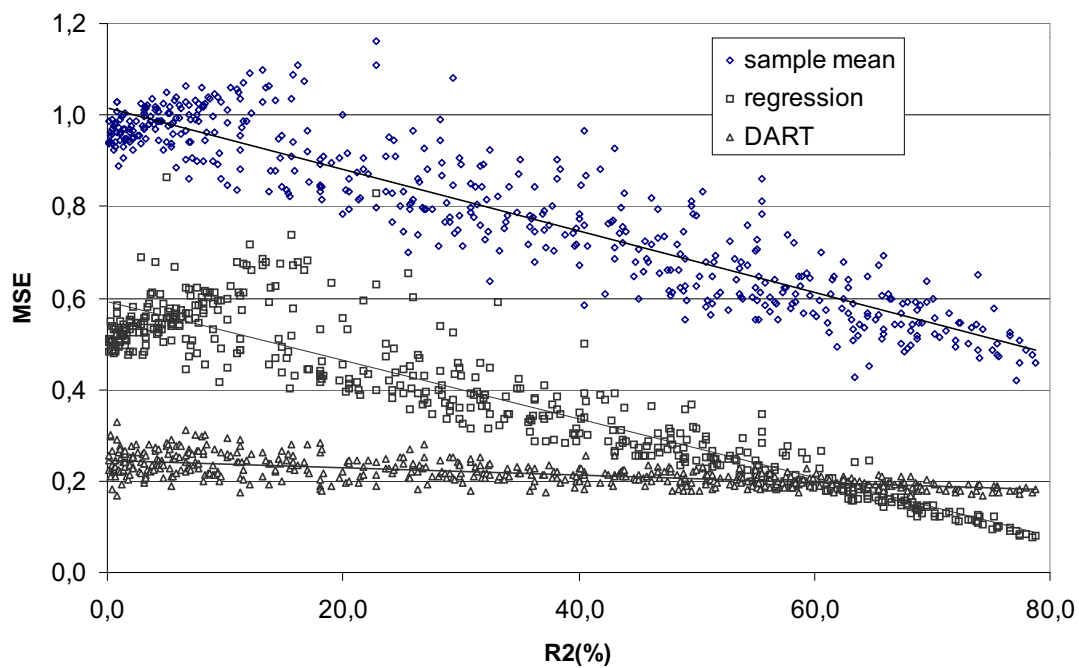


Figure 2: *Mean MSEs of sample mean, post-stratified, regression and DART estimators with respect to the R^2 values (200 sample cases).*



The post stratified estimator is almost unbiased as expected, since within the post-strata the chance of selection or the response probability is constant, but the efficiency may be low. The DART estimator shows fairly small biases, especially with the smallest sample size, for any level of non-linearity and it is the most efficient in all cases. Thus, the DART estimator seems capable of adjusting for biased selection (when the data are missing at random conditionally on the predictor variables) and has a greater efficiency than the parametric alternatives. For the informative sampling scheme, most of the MSEs is due to the biases, but that of the DART estimator is quite smaller.

Table 2 reports the results for estimators of the post-stratum means. For each post-stratum we computed the scaled MSE of the regression and DART estimators, having set equal to 100 that of the corresponding sample mean; then the means over the post-strata of the scaled MSEs were computed. In the Table, the mean scaled MSE is averaged over populations with R^2 values belonging to selected intervals for all three sampling methods. Again, the performance of the regression estimator is heavily dependent on its bias in estimating post-stratum means. On the contrary, the DART estimator is always more efficient than the sample mean for any level of non-linearity. For the case $n=200$, Figure 2 shows the scatter plot of the 500 MSE mean (absolute) values over post-strata of each estimator against the value of R^2 of the population. Note that the efficiency of the DART estimator, as measured by the mean MSE over post-strata, is much more stable across populations. For biased sampling with respect to X_1 the sample mean is unbiased for the post-stratum mean. The DART estimator is more efficient on the average and its efficiency is more stable while the regression estimator suffers from its bias when the model is misspecified. Similar results were obtained for the informative sampling scheme.

5. Simulation results for the prediction of Y .

In a recent paper, Borra & Di Ciaccio (2002) compared several non-parametric methods to analyse their predictive performances, considering three well-known artificial data-sets and some real data-sets (see also Borra & Di Ciaccio, 2001). In particular, they compared DART, PPR and MARS using bagging and boosting. A remarkable predictive performance of these non-parametric methods appeared in that paper.

Using the simulation framework described in section 3, we evaluated the predictive performance of DART with respect to non-linearity and biased sampling. For each sample, drawn from a population, we computed first the DART predictions and the linear regression predictions of Y , both for the sample units (training data) and the population units (test data), and then the corresponding Mean Squared prediction Errors, i.e. the mean of the squared differences between the prediction and the value of Y . Finally, for each population we computed the average of the MSEs obtained from the 500 realised samples.

In Figure 3-4 we show the results obtained plotting the average MSE against the value of R^2 of the population. In the plots we also included as benchmark the performance of the linear regression model. Results refer to the case of biased sampling with respect to Y (informative sampling scheme). In Figure 3, which refers to the bigger sample size ($n=2000$) and 250 populations of size 20,000, we can note the predictive power of DART (lower MSE), independently of the degree of linearity in the data-set.

Figure 3: *MSE for linear regression and DART for training (2000 cases) and test units (20,000 cases), with respect to the R^2 values. Biased sampling with respect to Y .*

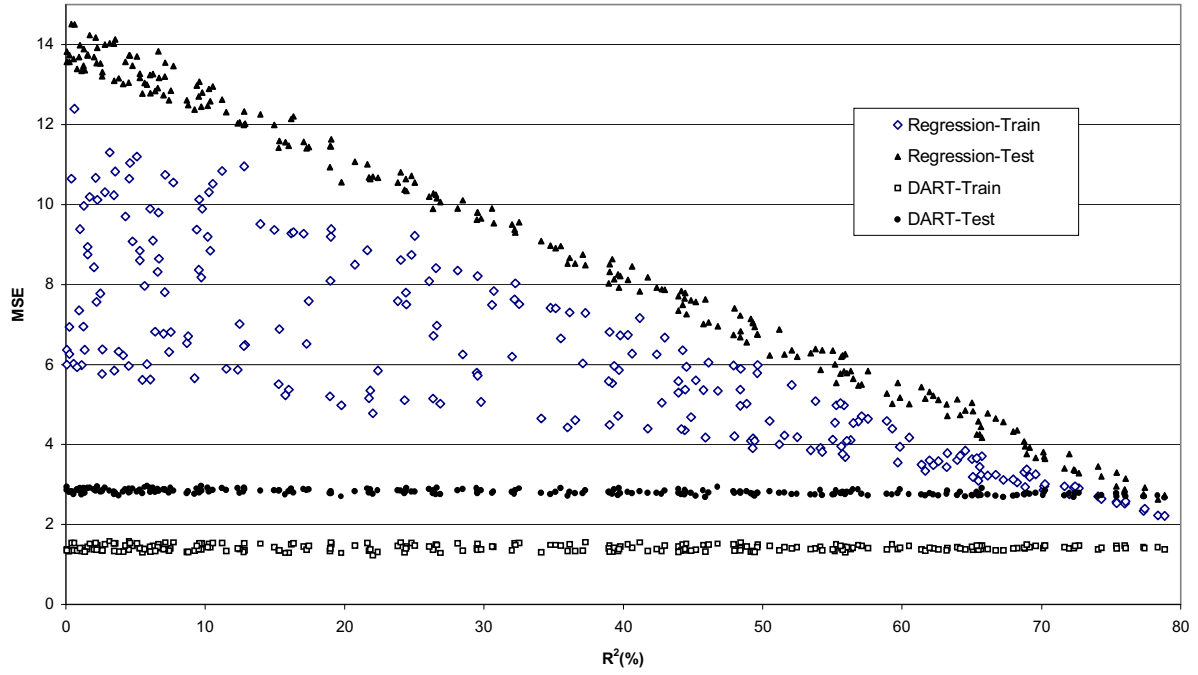
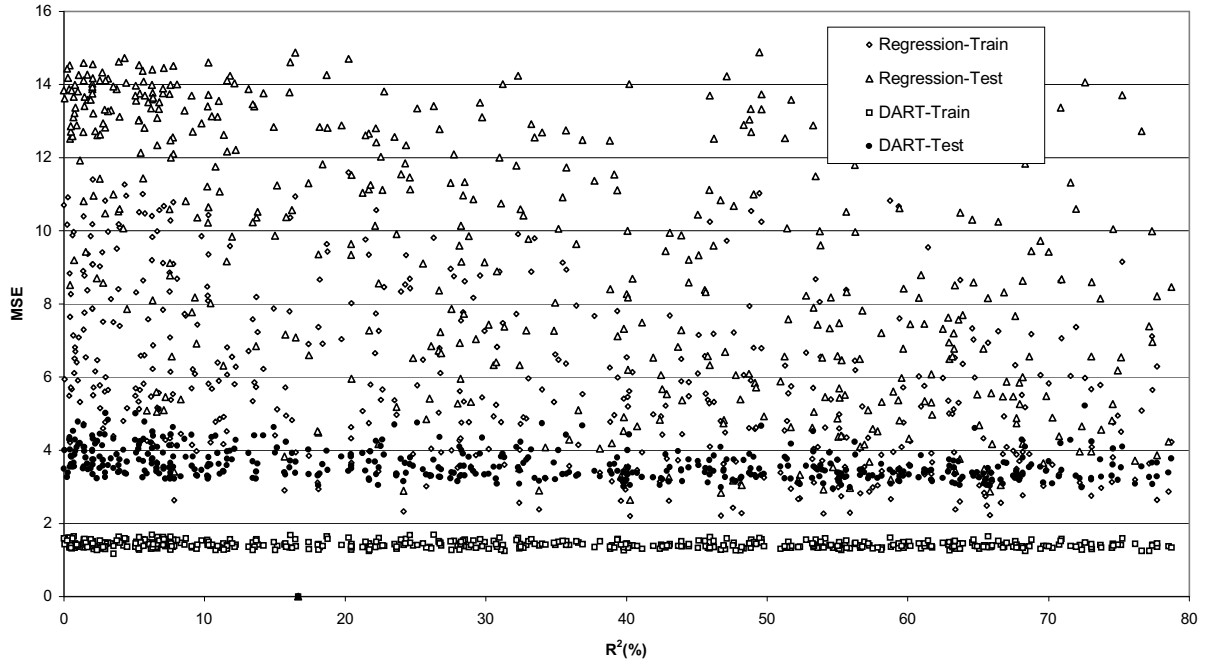


Figure 4: *MSE for linear regression and DART for training (200 cases) and test units (5000 cases), with respect to the R^2 values. Biased sampling with respect to Y .*



Only when non-linearity is completely absent, the linear regression model is competitive. As noted previously, most of the MSE is due to the bias while the variance of both estimators is negligible. Figure 4 refers to the smaller sample size ($n=200$) and 500 populations of size 5000. DART is clearly more effective than the linear regression prediction even when the relation between Y and X is linear. These outcomes also explain the results reported in section 4: the good predictive performance of DART in estimating the mean of Y is due to the good predictive performance for the unit values of Y .

6. Final remarks

In this paper, by means of a very large simulation, we analysed the problem of estimating the unit values or the mean of a variable Y when the sample is biased and non-linear relationships are present in the data. As outlined in sections 1 and 2, this is a very common problem in Data Mining applications. In this case, the results we obtained suggest the use of non-parametric methods. The tested method, DART, has been very effective in all simulations apart when the relationships are perfectly linear. This outcome suggests the use of this method in numerous situations, for example when we have to impute a large amount of missing data, which is again quite common in Data Mining. For the future we are planning to extend the study in order to include also qualitative data.

References

- Borra S., Di Ciaccio A. (2002) Improving nonparametric regression methods by bagging and boosting, *CSDA (guest eds Borra & Di Ciaccio)*, vol. 38 n. 4.
- Borra S., Di Ciaccio A. (2001) Performance evaluation of Bagging and Boosting in nonparametric regression, *Metron*, vol. LIX n. 3-4.
- Breidt F.J., Opsomer J. (2000) Local Polynomial Regression Estimators in Survey Sampling, *The Annals of Statistics*, 28, 1026-1053.
- Coppi, R. (2002) A theoretical framework for Data Mining: the “Informational Paradigm”, *CSDA (guest eds. Borra & Di Ciaccio)*, v. 38 n. 4, 501-515.
- Di Ciaccio A., Montanari G. E. (2001) “A nonparametric regression estimator of a finite population mean”, *Book of Short Papers, CLADAG 2001*, Palermo, 173-176.
- Friedman J.H.(1996a) *Local Learning Based on Recursive Covering*, Tech. Rep., Dept. of Statistics, Stanford University, available at <http://www-stat.stanford.edu/~jhf>.
- Friedman J.H.(1996b) DART/HYESS User Guide, Technical Report, Dept. of Statistics, Stanford University, available at <http://www-stat.stanford.edu/~jhf.60>
- Hastie T., Tibshirani R., Friedman J. (2001) *The elements of statistical learning: data mining, inference, and prediction*, Springer Verlag, New York.
- Särndal C.E., Swensson B., Wretman J.H.(1992) *Model assisted survey sampling*, Springer-Verlag, New York.
- Sudgen R.A., Smith T.M.F. (1984) Ignorable and informative designs in survey sampling inference, *Biometrika*, 71, 495-506.