Efficient and Comprehensible Local Regression

Luís Torgo

LIACC-FEP, University of Porto
R. Campo Alegre, 823 – 4150 Porto – Portugal
ltorgo@ncc.up.pt URL : http://www.ncc.up.pt/~ltorgo

Abstract. This paper describes an approach to multivariate regression that aims at improving the computational efficiency and comprehensibility of local regression techniques. Local regression modeling is known for its ability to accurately approximate quite diverse regression surfaces with high accuracy. However, these methods are also known for being computationally demanding and for not providing any comprehensible model of the data. These two characteristics can be regarded as major drawbacks in the context of a typical data mining scenario. The method we describe tackles these problems by integrating local regression within a partition-based induction method.

1 Introduction

This paper describes a hybrid approach to multivariate regression problems. Multivariate regression is a well known data analysis problem that can be loosely defined as the study of the relationship between a target continuous variable and a set of other input variables based on a sample of cases. In many important regression domains we cannot assume any particular functional form for the model describing this relationship. This type of problems demand for what is usually known as non-parametric approaches. An example of such techniques is local regression modeling (*e.g.* [3]). The basic idea behind local regression consists of delaying the task of obtaining a model till prediction time. Instead of fitting a single model to all given data these methods obtain one model for each query case using only the *most similar* training cases. As a result of this methodology these techniques do not produce any visible and comprehensible model of the given training data. Moreover, for each query point its "neighbors" have to be found, which is a time-consuming task for any reasonably large problem. Still, these models are able to easily adapt to any form of regression surface, which leads to large advantages in terms of their ability to approximate a wide range of functions. In this paper we address the drawbacks of local models by integrating them with regression trees.

2 Local Regression Modeling

According to Cleveland and Loader [3] local regression modeling traces back to the 19th century. These authors provide a historical survey of the work done since then. In

this paper we focus on one particular type of local modeling, namely kernel regression. Still, the described methodology is applicable to other local models. Within kernel regression a prediction for a query case is obtained by an averaging process over the most similar training cases. The central issue of these models is thus the notion of similarity, which is determined using a particular metric over the multidimensional space defined by the input variables. Given a data set $\{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$, where $\mathbf{x}_i$ is a vector of input variable values, a kernel model prediction for a query case $\mathbf{x}_q$ is obtained by,

$$k(\mathbf{x}_q) = \frac{1}{SKs} \sum_{i=1}^{n} K\left(\frac{d(\mathbf{x}_i, \mathbf{x}_q)}{h}\right) \times y_i \qquad (\mathbf{1})$$

where,

$d(.)$ is the distance function between two instances;
$K(.)$ is a kernel (weighing) function;
h is a bandwidth (or neighbourhood size) value;
and SKs is the sum of all weights, *i.e.* $SKs = \sum_{i=1}^{n} K\left(\frac{d(\mathbf{x}_i, \mathbf{x}_q)}{h}\right)$.

In this work we have used an Euclidean distance function together with a gaussian kernel (see [1] for an overview of these and other alternatives).

A kernel prediction can be seen as a weighed average of the target variable values of the training cases that are nearer to the query point. Each of the training cases within a specified distance (the bandwidth h) enter this averaging. Their weight is inversely proportional to the distance to the query, according to the $K(.)$ gaussian function.

The classical definition of the knowledge discovery in databases [4] refers this process as striving to identify valid, novel, potentially useful, and ultimately *understandable* patterns in data. From the perspective of understandability the local regression framework described above is very poor. Another characteristic of a typical data mining problem is its high dimensionality, *i.e.* the large number of cases and/or variables. Local modeling has a very high computational complexity if applied as described above. In effect, the prediction for each query case demands a look-up over all training cases to search for the most similar instances. This process has a complexity of the order of $O(n \times v)$ for each test case, where n is the number of training cases, and v is the number of variables.

3 Local Regression Trees

Regression trees (*e.g.* [2]) are non-parametric models that have as main advantages a high computational efficiency and a good compromise between comprehensibility and predictive accuracy. A regression tree can be seen as a partitioning of the input space. This partitioning is described by a hierarchy of logical tests on the input variables.

Standard regression trees usually assume a constant target variable value within each partition.

The regression method we propose consists of using local regression in the context of the partitions defined by a regression tree. The resulting model differs from a regression tree only in prediction tasks. Given a query case we drop it down the tree until a leaf is reached, as in standard regression trees. However, having reached a leaf (that represents a partition) we use the respective training cases to obtain a kernel prediction for the query case. From the perspective of local modeling these *local regression trees* have two main advantages. Firstly, they provide a focusing effect, that avoids looking for the nearest training cases in all available training data. Instead we only use the cases within the respective partition, which has large computational efficiency advantages. Secondly, the regression tree can be seen as providing a rough, but comprehensible, description of the regression surface approximated by local regression trees.

4 Experimental Evaluation

This section describes a series of experiments designed with the goal of comparing local regression trees with kernel regression modeling. The goal of these experiments is to compare the predictive accuracy of kernel models and local regression trees, and also to assert the computational efficiency gains of the later. Regarding local regression trees we have used exactly the same local modeling settings as for kernel regression, the single difference being that one is applied in the leaves of the trees while the other uses the information of all training set. The experimental methodology used was a 10-fold cross validation (CV). The results that are shown are averages of 10 repetitions of 10-fold CV runs. The error of the models was measured by the mean squared error (MSE) between the predicted and truth values. Differences that can be considered statistically significant are marked by + signs (one sign means 95% confidence and two 99% confidence). The best results are presented in bold face.

Table 1 shows the results of these experiments with three different domains. *Close Nikkei 225* and *Close Dow Jones* consist of trying to predict the evolution of the Nikkei 225 and Dow Jones stock market indices for the next day based on information of previous days values and other indices. *Telecomm* is a commercial telecommunications problem used in a study by Weiss and Indurkhya [7]. The two former consist of 2399 observations each described by 50 input variables, while the later contains 15000 cases described by 48 variables.

Table 1. Comparing local regression trees with kernel models.

	Close Nikkei 225			***Close Dow Jones***			***Telecomm***		
	Local RT	*Kernel*		*Local RT*	*Kernel*		*Local RT*	*Kernel*	
MSE	140091.6	**125951.1**	++	**86.8**	214.5	++	**42.40**	57.19	++
CPU sec.	**4.4**	6.5	++	**2.47**	6.66	++	**63.57**	452.88	++

The results in terms of predictive accuracy are contradictory. In effect, both two methods achieve statistically significant (> 99% confidence) wins on different domains. However, local regression trees are able to significantly outperform kernel

models in terms of computation efficiency, in spite of the small size of both the training and testing samples. In effect, additional simulation studies with increasing sample sizes have shown a more significant efficiency advantage of local regression trees [6]. Further details on these and other experiments can be found in [5, 6].

5 Conclusions

Local regression is a well-known data analysis method with excellent modeling abilities in a large range of problems. However, these techniques suffer from a high computational complexity and by not obtaining any visible and comprehensible model of the data. These can be considered major drawbacks in a typical data mining scenario.

In this paper we have described local regression trees that can be regarded as a new type of regression models that integrate a partition-based technique with local modeling. Local regression trees provide the smoothing effects of local modeling within the efficiency and comprehensibility of partition-based methods. Through the use of kernel models in the leaves of a standard regression tree we are able to provide a focusing effect on the use of kernel models with large advantages in the computation necessary to obtain the predictions. At the same time, the partitioning obtained with the tree can be regarded as a comprehensible overview of the regression surface being used to obtain the predictions.

We have carried out a large set of experiments that confirmed that local regression trees have an overwhelming advantage in terms of computation time with respect to standard local modeling techniques. Moreover, we have observed significant advantages in terms of predictive accuracy in several data sets.

References

1. Atkeson,C.G., Moore,A.W., Schaal,S.: Locally Weighted Learning. *Artificial Intelligence Review*, **11**, 11-73. Special issue on lazy learning, Aha, D. (Ed.), 1997.
2. Breiman,L. , Friedman,J.H., Olshen,R.A. & Stone,C.J.: *Classification and Regression Trees*. Wadsworth Int. Group, Belmont, California, USA, 1984.
3. Cleveland,W., Loader,C.: Smoothing by Local Regression: Principles and Methods (with discussion). *Computational Statistics*, 1995.
4. Fayyad,U.,Shapiro,G.,Smyth,P.:From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining*, Fayyad *et al.*(eds). AAAI Press (1996).
5. Torgo, L.: *Inductive Learning of Tree-based Regression Models*. Ph.D. Thesis. Dept. of Computer Science, Faculty of Sciences. University of Porto, 1999. Available at http://www.ncc.up.pt/~ltorgo.
6. Torgo,L.: Efficient and Comprehensible Local Regression. LIACC, Machine Learning Group, Internal Report n.99.2 , 1999. Available at http://www.ncc.up.pt/~ltorgo.
7. Weiss, S. and Indurkhya, N.: Rule-based Machine Learning Methods for Functional Prediction. *Journal of Artificial Intelligence Research* (JAIR), **3**, pp.383-403, 1995