Prediction of pollution levels from atmospheric variables
A study using clusterwise symbolic regression

Nikhil Suresh[1]
up201801861@fep.up.pt
Paula Brito[1]
mpbrito@fep.up.pt
Sónia Dias[2]
sdias@estg.ipvc.pt

[1] Faculdade de Economia
University of Porto & LIAAD INESC TEC,
Portugal

[2] Escola Superior de Tecnologia e Gestão
Instituto Politécnico de Viana do Castelo
Viana do Castelo, & LIAAD INESC TEC, Portugal

Abstract

This work performs statistical analysis of "Big data", considering the recent approach of Symbolic Data Analysis (SDA). The practical situation under study concerns the prediction of pollution levels in Senegal from atmospheric variables (meteorological indicators). The large number of records leads to the need of data aggregation. A temporal aggregation (by day) is made, where to each new unit (day) corresponds the interval of recorded values (minimum and maximum) in a given day. The symbolic data studied in this work is therefore interval data.

The objective was then to obtain symbolic regression models that allow explaining an objective interval-valued variable, the pollution level, as a function of explanatory interval-valued variables - the atmospheric variables. However, a single regression model is often not sufficient to adequately model the phenomenon under study. Thus, it was necessary to identify classes in the observed set and obtain a specific model appropriate for each class. To solve this problem, clusterwise regression for interval-valued data was developed.

1 Introduction

In classical data analysis, data is usually represented as an array where rows represent individuals and columns represent the variables (or attributes) describing them. It is possible to represent the data in a two dimensional array of n rows and p columns since a single value, numerical or categorical, is recorded for each variable and for each individual. However, when data is grouped to a higher level, the classical solution which is to use the mean, median or mode to represent each group leads to a loss of information, especially as concerns the variability present in each group. In such situations, SDA [1, 2] provides a framework to represent data with inherent variability, by using variables of special types. Among these representations, the focus in this work is on interval-valued data. A combination of existing dynamic clustering techniques and regression models for interval-valued data is proposed.

2 Problem: Predicting the levels of pollution in Senegal

The data under study consists of records of observations of atmospheric variables (meteorological indicators) and levels of pollution in Senegal, recorded from January 2006 to December 2010. The explanatory variables are wind speed, wind direction, air temperature and relative humidity, and the response variable is the particules concentration. The data was aggregated by day to form interval-valued variables recording the minimum and maximum values for each day. From the microdata, Table 1, the aggregation per day allows building an interval data array, as in Table 2.

Year	Month	Day	Hour	Min	Air Temp	Humitidy	…
2006	1	1	0	0	20.34	18.07	…
2006	1	1	0	5	20.30	18.09	…
2006	1	1	0	10	20.18	18.23	…
2006	1	1	0	15	20.14	18.30	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	…

Table 1: A snippet of Senegal meteorological indicators

The objective of this study is to predict the response variable, i.e., the particules' concentration, from the meteorological variables. However, a single regression model is often not sufficient to adequately model such relations. With the application of a clusterwise regression model for the interval data, we expect to obtain better results, by considering a partition of the time periods (days).

Year	Month	Day	Air Temp	Humitidy	…
2006	1	1	[20.13;20.34]	[18.07;18.30]	…
2006	1	2	[20.04;21.3]	[18.1;18.95]	…
⋮	⋮	⋮	⋮	⋮	…

Table 2: Senegal data snippet aggregation

3 The method

3.1 Interval Distribution (ID) regression model

Dias and Brito [3] proposed a new linear regression method for interval-valued variables known as the Interval Distribution (ID) regression model. In this approach, the intervals are represented by quantile functions taking into account the distribution within them. As it is usually the case in the literature, the Uniform distribution is assumed within each interval. Therefore, the quantile function that represents each interval is a linear non-decreasing function with domain $[0,1]$.

For each observation i of an interval-valued variable Y, $Y(i)$ is a interval $I_{Y(i)} = \left[\underline{I}_{Y(i)}, \overline{I}_{Y(i)}\right]$ where $\underline{I}_{Y(i)}, \overline{I}_{Y(i)}$ are the respective lower and upper bounds; $I_{Y(i)}$ may also be written as $I_{Y(i)} = \left[c_{Y(i)} - r_{Y(i)}, c_{Y(i)} + r_{Y(i)}\right]$, where now $c_{Y(i)}, r_{Y(i)}$ are the center and half range of the interval.

The quantile function that represents the interval $I_{Y(i)}$, when the Uniform distribution is assumed is written as $\Psi^{-1}_{Y(i)}(t) = \underline{I}_{Y(i)} + (\overline{I}_{Y(i)} - \underline{I}_{Y(i)})t$ or $\Psi^{-1}_{Y(i)}(t) = c_{Y(i)} + r_{Y(i)}(2t-1), \quad t \in [0,1]$.

Figure 3.1 represents the interval $I = [1,3]$ and the respective quantile function $\Psi^{-1}(t) = 1 + 3t, \quad t \in [0,1]$.

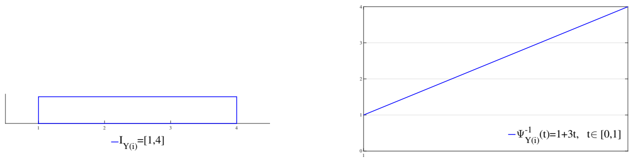

Figure 1: Graphical representation of the interval $[1,3]$ and respective quantile function.

The set of quantile functions defined from $[0,1]$ into R, with the usual operations of addition (between two quantile functions) and product of a quantile function by a real number, is not a subspace of the functions' vector space, but only a semi-vector space. The addition of two quantile functions raises no problem since the result is always a non-decreasing function. But the multiplication of a quantile function by a negative real number produces a function that is not non-decreasing, and hence cannot be a quantile function. Therefore, a problem arises when we multiply a quantile function representing an interval by -1, since we obtain a function that does not represent an interval.

As a result, when using quantile functions to represent intervals, the linear relation between interval-valued variables cannot be a direct adaptation of the classical linear regression model. That is not possible because if the parameters of the model were negative, the quantile function predicted for the response variable Y could well turn out to be a decreasing function, i.e., not a quantile function. Applying non-negativity constraints on the model would guarantee a quantile function, but that would compel a direct linear relationship between the explanatory variables and the response variable, a too strict limitation. To allow for both direct and inverse linear relations between the response and the explanatory variables, Dias and Brito [3] proposed a method that considers not only the quantile function that represents the interval observation of each explanatory variable but also the quantile function that represents the respective symmetric interval. Therefore, the ID regression model allows predicting, for each unit i, the quantile function $\Psi^{-1}_{Y(i)}(t)$ from the linear combination of $\Psi^{-1}_{X_j(i)}(t)$ and $-\Psi^{-1}_{X_j(i)}(1-t)$, as follows:

$$\Psi^{-1}_{\hat{Y}(i)}(t) = a_0 + \sum_{j=1}^{p} (a_j - b_j) c_{X_j(i)} + \sum_{j=1}^{p} (a_j + b_j) r_{X_j(i)} (2t-1) \qquad (1)$$

with $t \in [0,1]$; $a_j, b_j \geq 0$, $j \in \{1,2,\ldots,p\}$ and $a_0 \in R$.

The non-negative parameters in the model are obtained by solving a quadratic optimization problem using the Mallows distance (see, e.g., [3]), used to measure the difference between the observed and the predicted quantile functions, for each unit $i, i \in \{1,\ldots,n\}$.

A measure Ω, similar to the classical coefficient of determination, was deduced for the ID regression model:

$$\Omega = \frac{\sum\limits_{i=1}^{n} D_M^2\left(\widehat{Y}(i), \overline{Y}\right)}{\sum\limits_{i=1}^{n} D_M^2\left(Y(i), \overline{Y}\right)} \qquad (2)$$

where $\bar{Y}$ is the symbolic mean of Y; $\widehat{Y}(i)$ and $Y(i)$ are the estimated and observed intervals of the interval-valued variable Y for each observation i. This measure, based on the Mallows distance D_M, measures the goodness of fit of the model, and ranges between 0 and 1.

3.2 Clusterwise Regression

The Clusterwise Regression model proposed in this work combines the dynamic clustering algorithm [4], with the ID regression model, considering a Uniform distribution within the intervals, in order to identify both a partition of the data units and the relevant regression models, one for each cluster. The steps of the algorithm to be followed are:

Step 1: Represent the interval data by quantile functions.

Step 2: Consider an initial partition of the given units.

Step 3: Fit a regression for each cluster using the ID Model.

Step 4: Re-assign each unit to the cluster that provides the best fit, as measured by the squared Mallows distance.

Steps 3 and 4 are repeated until convergence is attained and a local minimum of the sum of squares of the errors (measured by the Mallows distance) is obtained (or the fixed maximum number of iterations is reached).

The process may be applied varying the number of clusters K; for each fixed K, the implemented algorithm allows for different initial partitions, and selects the solution with lowest Total Error:

$$W = \sum_{k=1}^{K} \sum_{i \in C_k} D^2(Y(i), \widehat{Y}^k(i)) \qquad (3)$$

To select the best solution, across different K, we use the Weighted Coefficient of Determination [3],

$$\Omega = \sum_{k=1}^{K} \frac{n_k}{n} \Omega_k \quad \text{with} \quad \Omega_k = \frac{\sum\limits_{i \in P_k} D_M^2\left(\widehat{Y}^k(i), \overline{Y}_k\right)}{\sum\limits_{i \in P_k} D_M^2\left(Y(i), \overline{Y}_k\right)} \qquad (4)$$

where n_k is the number of observations in class k; $\bar{Y}_k$ is the (local) symbolic mean of Y in class k and $\widehat{Y}^k(i)$ is the estimated interval of $Y(i)$ obtained by the (local) regression model in class $k, k \in \{1,\ldots,K\}$.

Another measure used is the (adapted) Silhouette coefficient [5]:

$$S = \sum_{i=1}^{n} \frac{S(i)}{n} \qquad (5)$$

where, for each $i \in P_k$

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (6)$$

where the $a(i) = D_M^2(Y(i), \widehat{Y}^k(i))$ is the squared Mallows distance from unit i to its local estimate and $b(i) = \min_{\substack{l=\{1,\ldots,K\} \\ l \neq k}} D_M^2(Y(i), \widehat{Y}^l(i))$ is the minimum squared Mallows distance from unit i to the estimate provided by another class.

The final clusters may then be used to predict target intervals for new observations.

4 Results and Conclusions

The Clusterwise Regression method presented above was applied to the dataset described in Section 2 multiple times for different parameters. For each value of the number of clusters, 15 different initial partitions were analyzed. The algorithm was repeteadly applied until there was no increase in the value of the evaluation measure, or until the increase in the evaluation measure became negligible. Table 3 presents the best assessment measures received for each value of number of clusters. It was expected that the weighted coefficient of determination would rise with the number of clusters K. But it is no surprise that the rise would plateau after a certain value of K, in this case 5, for which the value of the weighted Ω attains 92%.

Nb. of clusters	Weighted Ω	Silhouete Coef.
2	0.7709	0.7963
3	0.8604	0.7187
4	0.9025	0.7052
5	0.9179	0.6892
6	0.9181	0.6840
7	0.9277	0.6692
8	0.9323	0.6438
9	0.9340	0.6717
10	0.9337	0.6679

Table 3: Model evaluation measures

The advantages of using a clusterwise regression model is that it fits one regression model for each cluster. Each cluster seems to have its own set of relevant regressors, with different values for these regressors. This provides a lot more flexibility than to fit a model for the entire dataset, which could dilute the effect of one specific regressor over a subset of data. In this case, with a global model we indeed obtain a worse fit, with $\Omega = 0.5685$.

References

[1] H.-H. Bock and E. Diday (Eds.). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer, Heidelberg, 2000.

[2] P. Brito. Symbolic data analysis: Another look at the interaction of data mining and statistics. *WIREs Data Mining and Knowledge Discovery*, 4(4):281–295, 2014.

[3] S. Dias and P. Brito. Off the beaten track: a new linear model for interval data. *European Journal of Operational Research*, pages 47–94, 2017.

[4] E. Diday and J.C. Simon. Clustering analysis. In *Digital pattern recognition*, pages 47–94. Springer, 1976.

[5] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.