2009 Ninth International Conference on Intelligent Systems Design and Applications

A Robust Prediction Method for Interval Symbolic Data

Roberta A.A. Fagundes
Universidade Federal de Pernambuco
Centro de Informática
[1]Av. Prof. Luiz Freire, s/n - Cidade Universitária
CEP 50740-540 - Recife PE - Brazil
raaf@cin.ufpe.br
[2] Universidade de Pernambuco
Campus Gov. Miguel Arraes de Alencar
Polo Comercial, BR 104, Km 62 Caruaru PE - Brazil

Renata M.C.R. de Souza
Universidade Federal de Pernambuco
Centro de Informática
Av. Prof. Luiz Freire, s/n - Cidade Universitária
CEP 50740-540 - Recife (PE) - Brazil
rmcrs@cin.ufpe.br

Francisco José A. Cysneiros
Universidade Federal de Pernambuco
Departamento de Estatística - CCEN
Av. Prof. Luiz Freire, s/n - Cidade Universitária
CEP 50740-540 - Recife (PE) - Brazil
cysneiros@de.ufpe.br

Abstract

This paper introduces a robust prediction method for symbolic interval data based on the simple linear regression methodology. Each example of the data set is described by feature vector, for which each feature is an interval. Two classic robust regression models are fitted, respectively for range and mid-points of the interval values assumed by the variables in the data set. The prediction of the lower and upper bounds of the new intervals is performed from these fits. To validate this model, experiments with a synthetic interval data set and an application with a cardiology interval-valued data set are considered. The fit and prediction qualities are assessed by a pooled root mean square error measure calculated from learning and test data sets, respectively.

1 Introduction

Most statistic methods for data analysis have been designed in a relatively simple way: the statistical unit is an individual (a person or an object) described by a well defined set of (qualitative or quantitative) variables that assume in just one single value. However, there are situations in which uncertainty or variability must be taken into account to faithfully represent the real word and the classical variables are not able to represent these nuances. In these cases, other kinds of variables, such as interval variables, are required. Interval data can arise through of natural aggregation of repeated measures or bounds of the set of possible values of an item or variation range of a variable through time. However, it is common that point outliers are present and therefore interval outliers can also be identified.

The statistical treatment of interval data has been considered in the context of Symbolic Data Analysis (SDA)[1], which is a domain in the area of knowledge discovery and data management related to multivariate analysis, pattern recognition and artificial intelligence. The aim of SDA is to provide suitable methods (clustering, factorial techniques, decision trees, etc.) for managing aggregated data described by multi-value variables, for which the cells of the data table contain set categories, intervals or weight (probability) distributions in [1].

In the framework of regression models for symbolic interval data, several approaches have been introduced ([2] - [5]). In these works, the parameters are estimated by the minimization of the squared error criterion function and this function is highly influenced by unusual data values. The main contribution of this paper is to introduce a robust prediction method for symbolic interval-valued data based on the robust linear regression methodology [6]. The proposed model consists of fitting two classic linear robust regression models to, respectively, the mid-point and the range of the

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.36

IEEE computer society

intervals. The prediction of an interval is based on a combination between fitted models.

The structure of the paper is as follows: Section 2 shows the real and synthetic interval data sets considered in the work. Section 3 describes the robust regression model for interval data. Section 4 presents a performance analysis of the proposed method regarding synthetic and real interval data sets. The assessment of the proposed method is based on the estimation of a pooled root mean square error measure. Finally, Section 5 gives the concluding remarks.

2 Interval-valued data

In classical data analysis, the items to be grouped are usually represented as vectors of quantitative or qualitative measurements where each column represents a variable. However, this model is too restrictive to represent complex data, which may, for instance, comprehend variability and/or uncertainty. Interval variables also allow consideration of imprecise data, coming from repeated measures or confidence interval estimation.

As this paper introduces a robust regression model for interval valued-data that can be potentially used when hypercube or rectangle outliers are present, a synthetic interval-valued data set and a cardiology interval-valued data set in $\Re^2$ containing rectangle outliers are here adopted in the context of a regression problem.

Let I_Y be a response interval variable that is related to a predictor interval variable I_X. Let $E = \{e_1, \ldots, e_n\}$ be an example set where each example e_i $(i = 1, \ldots, n)$ is represented as a interval quantitative feature vector $\mathbf{z} = (I_X(i), I_Y(i))$ with $I_X(i) = [l_X(i), u_X(i)] \in \Im = \{[a, b] : a, b \in \Re, a \leq b\}$ and $I_Y(i) = [l_Y(i), u_Y(i]] \in \Im$.

Let Y^c and X^c be, respectively, standard quantitative variables that assume as their values the mid-points of the intervals assumed by the symbolic interval-valued variables I_Y and I_X. Also, let Y^r and X^r, respectively, quantitative variables that assume as their values the ranges of the intervals assumed by the symbolic interval-valued variables I_Y and I_X.

2.1 Synthetic interval-valued data sets containing rectangle outliers

A synthetic interval-valued data set in $\Re^2$ is generated from a synthetic standard quantitative data set of 250 points in $\Re^2$ such that each point belonging to the standard quantitative data set is a mid-point (seed) for a rectangle in $\Re^2$ and this rectangle is built from a randomly selected range value. Here, the configuration for mid-point and range assumes that the mid-point and range of the intervals are simulated independently from uniform distributions according to [5].

Let $X^c \sim U[20, 40]$ and $X^r \sim U[1, 5]$ be, respectively, mid-point and range variables associated to a independent interval-valued variable I_X. Let $Y^r \sim U[1, 5]$ be the range variable associated to the dependent interval-valued variable I_Y. The mid-point variable Y^c is related to mid-point variable X^c as $Y^c = \boldsymbol{\beta_0}$+ $\boldsymbol{\beta_1} X^c + \epsilon$ where $\beta_0 \sim U[-10, 10]$ and $\beta_1 \sim U[-10, 10]$ and $\epsilon \sim U[-5, 5]$ is an error.

Here, an rectangle outlier is an rectangle that is remote in the *y* coordinate of the mid-point of this rectangle. The effect that this rectangle has on the regression model depends on the *x* coordinate of its mid-point and on the general disposition of the other rectangles in the data set.

Interval-valued outliers are created based on the mid-point data set $(Y^c(i),\ \mathbf{X}^c(i))$ $(i = 1, \ldots, 250)$. First, this set in $\Re^2$ is sorted ascending by the dependent variable Y^c and a small cluster containing the m first points of the sorted set $(Y^c(i), \mathbf{X}^c(i))$ are selected. The observations of this cluster are transformed into point outliers by $Y^c(i) = Y^c(i) - 3 * S_{Y^c}$ $(i = 1, \ldots, m)$ where S_{Y^c} is the standard deviation of the values $Y^c(i)$ $(i = 1, \ldots, n)$ of the data set.

After that, the lower and upper bounds of the intervals $I_X(i)$ and $I_Y(i)$ $(i = 1, \ldots, n)$ of the rectangle set are obtained by $I_X(i) = [X^c(i) - X^r(i)/2, X^c(i) + X^r(i)/2$ and $I_Y(i) = [Y^c(i) - Y^r(i)/2, Y^c(i) + Y^r(i)/2]$.

Figure 1 illustrates the interval-valued data set containing rectangle outliers. This interval-valued configuration take into account low variability on the ranges of the rectangles. Note in this figure that, due to the low variability on the ranges, there are rectangle ouliters that are remote in the *y* coordinate.

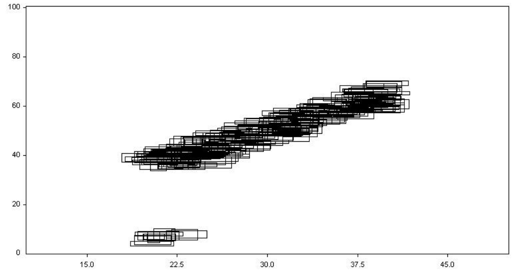

Figure 1. Interval-valued data set containing rectangle oultliers

2.2 Cardiology interval-valued data set

The cardiology interval-valued data set consists of a set of 59 patients described by 2 interval variables. In this data set, the independent interval variable - *Systolic Blood Pressure* has been considered for predicting the dependent interval variable *Pulse Rate*. These cardiology data are presented in [7] and [8]. They were collected by the department of Nephrology of the Hospital *Valle del Nalón*, in the city of *Langreo*, Spain.

Table 1 displays part of this data set. In this data set, two intervals for each one of the 59 patients w_u $(u = 1, \ldots, 59)$ are recorded.

Table 1. Cardiology data set with two interval variables

w_u	Pulse Rate	Systolic Blood Pressure
	I_Y	I_X
w_1	[58,90]	[118,173]
w_2	[47,68]	[104,161]
w_3	[32,114]	[131,186]
w_4	[61,110]	[105,157]
w_5	[62,89]	[120,179]
w_6	[63,119]	[101,194]
w_7	[51,95]	[109,174]
w_8	[49,78]	[128,210]
$\ldots$	$\ldots$	$\ldots$
w_{56}	[70,105]	[120,188]
w_{57}	[40,80]	[95,166]
w_{58}	[56,97]	[92,173]
w_{59}	[37,86]	[83,140]

Figure 2 displays rectangles for the interval-valued pairs: (*Systolic Blood Pressure* (I_X),*Pulse Rate* (I_Y)). Note that there are rectangle outliers that are unusual rectangles on the ranges of the intervals of this data set.

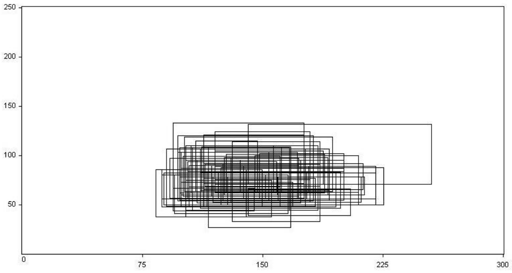

Figure 2. *Systolic Blood Pressure* **(x coord.) and** *Pulse Rate* **(y coord.)**

3 Robust regression model for interval-valued data

In classic data analysis, there are many situations where there are outliers that affect the regression model. A case of considerable practical interest is one in which the observations follow a distribution that has longer or heavier tails than normal. These heavy-tailed distributions tend to generate outliers, and these outliers may have strong influence on the least squares method in the sense that there is no longer an optimal estimation technique [9].

The importance of taking into account the mid-point and range information in a linear regression model for predicting interval-valued data were demonstrated [5]. In this model, the estimation procedure is based on the least square method that does not change probabilistic hypothesis on the response variable. However, this model may also suffer strong influence when there are interval outliers.

This section presents a robust regression model for interval-valued data that considers two independent classic robust regressions on the mid-point and range of the intervals of a learning data set, respectively. The prediction of the lower and upper bounds of new intervals is based on information on the mid-point and range estimates.

3.1 Constructing the model

In this model, each example e_i $(i = 1, \ldots, n)$ is represented by two vectors $\mathbf{z}_c = (X^c(i), Y^c(i))$ and $\mathbf{z}_r = (X^r(i), Y^r(i))$ where $X^c(i) = [l_X(i) + u_X(i)]/2$, $X^r(i) = u_X(i) - l_X(i)$, $Y^c(i) = [l_Y(i) + u_Y(i)]/2$ and $Y^r(i) = u_Y(i) - l_Y(i)$.

The predictor interval variable I_X is related to the response interval variable I_Y according to two linear regression equations, respectively, on their mid-point and range values

$$\begin{aligned} Y^c &= \beta_0^c + \beta_1^c X^c \\ Y^r &= \beta_0^r + \beta_1^r X^r \end{aligned} \tag{1}$$

The values of β_0^c, β_1^c , β_0^r and β_1^r are estimated minimizing a criterion function based on a function ρ of the residuals e_i^c and a function ρ of the residuals e_i^r

$$\sum_{i=1}^{n} \rho(e_i^c) + \rho(e_i^r) \tag{2}$$

where

$$\begin{aligned} \rho(e_i^c) &= \rho(Y^c(i) - \beta_0^c + \beta_1^c X^c) \\ \rho(e_i^r) &= \rho(Y^r(i) - \beta_0^r + \beta_1^r X^r) \end{aligned} \tag{3}$$

The equation (2) yields two minimization problems:

1. to find β_0^c and β_1^c that minimizes

$$\sum_{i=1}^{n} \rho(Y^c(i) - \beta_0^c + \beta_1^c X^c) \tag{4}$$

2. to find β_0^r and β_1^r that minimizes

$$\sum_{i=1}^{n} \rho(Y^r(i) - \beta_0^r + \beta_1^r X^r) \tag{5}$$

To minimize the equation (4) equate the first partial derivatives of ρ with respect to β_0^c and β_1^c to zero, yielding a necessary condition for a minimum. This gives the system of two equations

$$\sum_{i=1}^{n} X^c(i)\psi\left(\frac{Y^c(i) - \beta_0^c + \beta_1^c X^c}{s}\right) = 0$$

$$\sum_{i=1}^{n} \psi\left(\frac{Y^c(i) - \beta_0^c + \beta_1^c X^c}{s}\right) = 0$$

$\psi = \rho'$ and s is a scale parameter This system must be solved by iterative methods. The iteratively reweighted least squares is most widely used.

The robust estimate of β_0^c and β_1^c which minimizes the equation (4) is in matrix notation the solution to

$$\boldsymbol{\beta} = (\beta_0^c, \beta_1^c)^T = \mathbf{A}^{-1}\mathbf{b}$$

where

$$\mathbf{A} = \begin{pmatrix} \sum_{i=1}^{n} w_i^c & \sum_{i=1}^{n} w_i^c X^c(i) \\ \sum_{i=1}^{n} w_i^c X^c(i) & \sum_{i=1}^{n} w_i^c (X^c(i))^2 \end{pmatrix}$$

and

$$\mathbf{b} = \begin{pmatrix} \sum_{i=1}^{n} w_i^c Y^c(i) \\ \sum_{i=1}^{n} w_i^c X^c(i) Y^c(i) \end{pmatrix}$$

with w_i^c being the weight given to the residual e_i^c ($i = 1, \ldots, n$).

The same procedure is applied to the minimization problem 2. and the robust estimate of β_0^r and β_1^r which minimizes the equation (5) is in matrix notation the solution to

$$\boldsymbol{\beta} = (\beta_0^r, \beta_1^r)^T = \mathbf{A}^{-1}\mathbf{b}$$

where

$$\mathbf{A} = \begin{pmatrix} \sum_{i=1}^{n} w_i^r & \sum_{i=1}^{n} w_i^r X^r(i) \\ \sum_{i=1}^{n} w_i^r X^r(i) & \sum_{i=1}^{n} w_i^r (X^r(i))^2 \end{pmatrix}$$

and

$$\mathbf{b} = \begin{pmatrix} \sum_{i=1}^{n} w_i^r Y^r(i) \\ \sum_{i=1}^{n} w_i^r X^r(i) Y^r(i) \end{pmatrix}$$

with w_i^r being the weight given to the residual e_i^r ($i = 1, \ldots, n$).

There are a number popular robust criterion functions and the robust regression method can be classified by the their ψ function that controls the weight given to each residual [6]. For example, Tukey's biweight criterion function $\rho(x)$ has a monotone $\psi(x)$ function and does not weigh large residuals as heavily as least squares.Tukey's biweight function $\rho(x)$, its corresponding $\psi(x)$ function and its corresponding weight function $w(x)$ are, respectively, given as:

$$\rho(x) = \begin{cases} \frac{c^2}{6}(1 - [1 - (x/c)^2]^3) \text{ for } |x| \leq c \\ \frac{c^2}{6} \text{ for } |x| > c \end{cases}$$

$$\psi(x) = \begin{cases} x[1 - (x/c)^2]^2 \text{ for } |x| \leq c \\ 0 \text{ for } |x| > c \end{cases}$$

$$w(x) = \begin{cases} [1 - (x/c)^2]^2 \text{ for } |x| \leq c \\ 0 \text{ for } |x| > c \end{cases}$$

3.2 Rule of prediction

The prediction of the lower and upper bounds $\hat{I}_Y(v) = [\hat{l}_Y, \hat{u}_Y]$ of a new example v is based on the prediction of $\hat{Y}^c(v)$ and $\hat{Y}^r(v)$. Given the interval $I_X(v) = [l_X, u_X]$ with $X^c(v) = (l_X + u_X)/2$ and $X^r(v) = (u_X - l_X)/2$, the interval $\hat{I}_Y(v) = [\hat{l}_Y, \hat{u}_Y]$ is obtained as follows:

$$\hat{l}_Y = \hat{Y}^c(v) - \hat{Y}^r(v)/2 \text{ and } \hat{u}_Y = \hat{Y}^c(v) - \hat{Y}^r(v)/2$$

where

$$\begin{aligned}\hat{Y}^c(v) &= \hat{\beta}_0^c(v) + \hat{\beta}_1^c X^c(v) \\ \hat{Y}^r(v) &= \hat{\beta}_0^r(v) + \hat{\beta}_1^r X^r(v)\end{aligned}$$

4 Performance Analysis

To show the usefulness of the robust model presented in this paper, experiments with a synthetic interval-valued data set and an application with a real interval-valued data set in $\Re^2$ are considered in this section. These interval-valued data sets contain rectangle outliers. Moreover, the proposed robust regression model is compared with the non-robust regression model for interval-valued data introduced in [5].

4.1 Results for the synthetic interval data set

Here, the analysis was performed in the framework of a Monte Carlo experiment with 100 replications of the data set. Test and learning sets are are randomly selected from each synthetic interval data set.The learning set corresponds to 75% of the original data set and the test data set corresponds to 25%.

The performance assessment of the robust linear regression model presented in this paper is based on the *pooled root mean-square error* ($PRMSE$). This measure is obtained from the observed interval values $I_Y(i) = [l_Y(i), u_Y(i)]$ $(i = 1, \ldots, n)$ of I_Y and from their corresponding predicted interval values $\hat{I}_Y(i) = [\hat{l}_Y(i), \hat{u}_Y(i)]$ and it is estimated in the framework of a Monte Carlo simulation with 100 replications in two ways.

For each learning synthetic interval-valued data set the $PRMSE$ measure is given by

$$PRMSE_1 = \sqrt{\frac{\sum_{i=1}^{250} \omega(i) error(i)}{250}}$$

where,

$$error(i) = [(l_Y(i) - \hat{l}_Y(i))^2 + (u_Y(i) - \hat{u}_Y(i))^2]$$

and $\omega(i)$ is the weight of the residual $r_i = Y^c(i) - \hat{Y}^c(i)$ $(i = 1, \ldots, n$ controlled by a influence function corresponding to robust criterion function adopted to fit $(Y^c(i)$, $\mathbf{X}^c(i))$. Here, the robust fit is obtained using the Tukey's biweight criterion function. In the non-robust linear model, the least squares criterion function weights all residuals equally to 1.0.

For each test synthetic interval-valued set the $PRMSE$ measure is given by

$$PRMSE_2 = \sqrt{\frac{\sum_{i=1}^{125} error(i)}{125}}$$

The $PRMSE_1$ and $PRMSE_2$ measures are estimated for each fixed configuration. At each replication of the Monte Carlo method, a robust linear regression model to the learning data set is fitted. Thus, the fitted model is used to predict the interval values of the dependent interval-valued variable I_Y in the test and learning data sets and these are calculated.

For each $PRMSE_k$ $(k = 1, 2)$, the average and standard deviation over the 100 Monte Carlo simulations is calculated and a statistical Student's t-test for paired samples at a significance level of 1% is then applied to compare the robust regression model proposed in this paper with the non-robust regression model for interval-valued data introduced by [5]. Let μ_k^R and μ_k^{NR} be the average of the $PRMSE_k$ for robust and non-robust models, respectively. The null and alternative hypotheses are, respectively:
$H_0 : \mu_k^R = \mu_k^{NR}$
$H_1 : \mu_k^R < \mu_k^{NR}$.

For further consistency in the results, this procedure is repeated considering 100 different values for the vector $\boldsymbol{\beta}$. At each iteration, the comparison between models is accomplished by a statistical Student's t -test applied to each measure. For specific values of the parameters $\beta_0 = 5$ and $\beta_0 = 1.5$, the values of the t-test statistic for learning and test data sets are, respectively, -141.109 and -38.184. Regarding the 100 different values for the vector $\boldsymbol{\beta}$, the rejection ratio of H_0 is equal to 100% for both learning and test data sets. These results show clearly that the robust regression model for interval-valued data is superior to the regression model proposed in [5].

4.2 Results for Cardiology interval-valued data set

Below are presented two regression equations $\hat{Y}^c$ and $\hat{Y}^r$ for the cardiology interval-valued data set according to the robust linear model using Tukey's biweight criterion function and the non-robust linear model using least squares criterion function, respectively.

- *Robust Linear Model*
  $\hat{Y}^c = 64.6044 + 0.0645X^c$
  $\hat{Y}^r = 39.6555 - 0.0019X^r$

- *Non-Robust Linear Model*
  $\hat{Y}^c = 65.6057 + 0.0620X^c$
  $\hat{Y}^r = 41.3228 - 0.0051X^r$

The fitted values for the interval-valued variable I_Y are computed from $I_Y(i) = [\hat{Y}^c(i) - \hat{Y}^r(i)/2, \hat{Y}^c(i) + \hat{Y}^r(i)/2]$ $(i = 1, \ldots, 59)$.

The performance of the models is also evaluated through the $PRMSE^2$. This measure is estimated by the leave-on-out method and the results are 19.09 and 18.37 for non-robust and robust linear models, respectively. The comparison between the methods shows that, for cardiology interval-valued data set, the robust model proposed in this paper is the best option.

5 Conclusions

A robust linear prediction model for symbolic interval-valued data is introduced in this paper. The input data set is described by feature vectors, for which each feature is an interval. The relationship between a dependent interval variable (response variable) and an independent interval variable is modeled by information contained in the range and mid-point of intervals. Two classic robust regression models are fitted independently for range and mid-points, respectively, and the prediction of the lower and upper bounds of the intervals is performed from these fits.

In order to validate the introduced robust model for interval data, experiments with a synthetic interval data data set and an application with a cardiology interval-valued data set containing interval outliers are considered. The fit and prediction qualities are assessed by on a *pooled root mean square error* and the results provided by the proposed method are compared with the correspondence results provided by non-robust regression model for interval data presented in [5]. The results showed that the robust model outperformed the non-robust model. This fact indicates, according to used interval data sets, the introduced robust linear model is not sensitive in the presence of interval-valued outliers.

Acknowledgments: The authors would like to thank CNPq (Brazilian Agency) for its financial support.

References

[1] E. Diday, and M. Noirhomme-Fraiture: Symbolic Data Analysis and the SODAS Software. Wiley, 2008.

[2] L. Billard, and E. Diday: Regression Analysis for Interval-Valued Data. In: Data Analysis,Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies (IFCS'00), Springer-Verlag, Belgium,2000, 369-374.

[3] L. Billard, and E. Diday: Symbolic Regression Analysis. In: Classification, Clustering and Data Analysis: Proceedings of the Eighenth Conference of the International Federation of Classification Societies (IFCS'02),Springer, Poland, 2002 , 281–288.

[4] L. Billard, and E. Diday: Symbolic Data Analysis: Conceptual Statistics and Data Mining,Wiley,2006.

[5] E.A. Lima Neto,and F.A.T De Carvalho:Centre and Range method for fitting a linear regression model to symbolic 3 interval data. Computational Statistics and Data Analysis 52, 2008, 1500-1515.

[6] P.J. Huber: Robust Statistics. Wiley, 1981 (republished in paperback 2003)

[7] A.M. Gil, M.A. Lubiano,M. Montenegro, and M.T. Lpez-Garcia:Least square fitting of an affine function and strength of association for interval-valued data, Metrika 56,2002,97-111.

[8] A.M. Gil, G. Gonzlez-Rodriguez, and M. Montenegro:Testing linear independence in linear models with interval-valued data, Computing Statistic and Data Analysis 51,2007,3002-3015.

[9] D.C. Montgomery, and E.A. Peck:Introduction to Linear Regression Analysis. Wiley, New York,1982.