Beyond the non-probabilistic symbolic regression models for interval variables

Eufrásio de A. Lima Neto (DE/UFPB),
Gauss M. Cordeiro (DEINFO/UFRPE)
Francisco de A.T. De Carvalho (Cin/UFPE)

eufrasio@de.ufpb.br

Recife, May 2009.

Schedule

1. Introduction
2. Interval-valued Datasets
3. The non-probabilistic symbolic regression models
4. The probabilistic symbolic regression models
5. Future Works

Introduction

- The **regression models** are one of the most important statistical methods to study the relationship between variables.
- In **data mining**, the regression models can be used in situations involving numerical prediction or classification.
- In **Symbolic Data Analysis** (SDA) is common to record interval-valued data:
  - Monthly interval temperatures in meteorological stations;
  - Daily interval stock prices;
  - From the aggregation of huge databases into a reduced symbolic data set.

Interval-valued Dataset

- Let E = $\{e_1,\ldots, e_n\}$ a set of observations described by *p+1* symbolic interval variables.
- Let each $e_i \in E$ (i = 1,..., n) represented by a vector of intervals

$$\mathbf{z}_i = (\mathbf{x}_i, y_i)$$

- where: $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $x_{ij} = [a_{ij}, b_{ij}] \in \Im = \{[a, b]: a,b \in a \leq b\}$ representing the **independent or explanatory interval variables X_j** (j = 1,..., p) and $y_i = [y_{Li}, y_{Ui}] \in \Im$ representing the **response or dependent interval variable Y**.

An Interval-valued Dataset

e	Pulse Rate (Y)	Systolic Blood Pressure (X_1)	Diastolic Blood Preassure (X_2)
1	[44-68]	[90-100]	[50-70]
2	[60-72]	[90-130]	[70-90]
3	[56-90]	[140-180]	[90-100]
4	[70-112]	[110-142]	[80-108]
5	[54-72]	[90-100]	[50-70]
6	[70-100]	[130-160]	[80-110]
7	[63-75]	[60-100]	[140-150]
8	[72-100]	[130-160]	[76-90]
9	[76-98]	[110-190]	[70-110]
10	[86-96]	[138-180]	[90-110]
11	[86-100]	[110-150]	[78-100]

The non-probabilistic symbolic regression models

- The non-probabilistic symbolic regression methods visualize the problem from a optimization point of view.
  - Find the best parameters estimates that minimize a criterion, like the sum of squares of errors.
- **Center Method – Billard and Diday (2000)**
  - They were the first to propose a regression model to interval-valued dataset.
  - This method uses the information contained in the midpoints of the intervals to fit a symbolic linear regression model.

The non-probabilistic symbolic regression models

- **Center and Range Method – De Carvalho et. al (2004) and Lima Neto and De Carvalho(2008)**
  - They combined the midpoint and the range information, producing a new method with a best prediction performance.
  - They also compare this approach with two independent regression models over the limits of the intervals.
- **Constrained Regression Methods – Lima Neto et. al. (2005)**
  - The main contribution of these methods was guarantee mathematical coherence in the prediction of the intervals bounds ($\hat{y}_{Li} \leq \hat{y}_{Ui}$).

The non-probabilistic symbolic regression models

- **Nonlinear Symbolic Regression Method – Lima Neto and De Carvalho (2008)**
  - They proposed the first nonlinear regression method for symbolic interval variables.
  - The method uses the information of the midpoint and range of the intervals.
  - This feature allow to the analyst a large possibility of nonlinear models an can guarantee that $\hat{y}_{Li} \leq \hat{y}_{Ui}$ without the use of inequality constraints.

The non-probabilistic symbolic regression models

- **Another important contributions related to symbolic regression models**
  - Linear regression models to symbolic variables type histogram, hierarchical and taxonomic (Billard and Diday, 2006).
  - Maia and De Carvalho (2008) presented an approach based on the L1 regression model for interval-valued data.
  - Souza et. al. (2008) has studied logistic regression models taking into account explanatory interval variables.

The non-probabilistic symbolic regression models

- The non-probabilistic symbolic regression models attack the problem from an optimization point of view;
- The use of inferential techniques over the parameters estimates and predicted values it is not possible because these methods do not take into account the probabilistic nature of the response interval variable Y.

The probabilistic symbolic regression models

- Lima Neto et. al. (2009) proposed a probabilistic symbolic regression model, called **Bivariate Generalized Linear Model (BGLM)**, for symbolic interval variables.
- The model consider the interval variable

$$\mathbf{Y = [y_L;y_U]}$$

as **bivariate random vector** with joint density probability function belonging to **bivariate exponential family**, denoted by

$$f(y_1, y_2; \theta_1, \theta_2) = \exp[\phi^{-1}\{y_1\theta_1 + y_2\theta_2 - b(\theta_1, \theta_2, \rho)\} + c(y_1, y_2, \rho, \phi)] \quad (1)$$

The probabilistic symbolic regression models

- The **Bivariate Generalized Linear Model (BGLM)**, is divided in two components:
- The **random component**

$$\mathbf{Y = [y_1;y_2]}$$

that following the **bivariate exponential family** (1)

- The **systematic component** is denoted by

$$\eta_1 = g_1(\mu_1) = \mathbf{X}_1\beta_1 \text{ and } \eta_2 = g_2(\mu_2) = \mathbf{X}_2\beta_2,$$

where g_1 and g_2 are *link functions* that connect the systematic component to the averages μ_1 and μ_2 of the variables $\mathbf{y_1}$ and $\mathbf{y_2}$, respectively.

The probabilistic symbolic regression models

- For particular ρ, it is possible to estimate the vector of parameters β_1 and β_2 based on the iterative Fisher scoring method.
- Obtained the parameters estimates of $\beta_1$and β_2, we compute the goodness-of-fit measure **deviance**

$$
\begin{aligned}
D(\rho) &= 2\sum_{i=1}^{n}\{y_{1i}[q_1(y_{1i},\rho) - q_1(\hat{\mu}_{1i},\rho)] + \\
&+ y_{2i}[q_2(y_{2i},\rho) - q_2(\hat{\mu}_{2i},\rho)] + \\
&+ [b(q_1(\hat{\mu}_{1i}), q_2(\hat{\mu}_{2i}),\rho) - b(q_1(y_{1i}), q_2(y_{2i}),\rho)]\}.
\end{aligned}
$$

- and consequently, we estimate the parameter dispersion ϕ

$$\tilde{\phi} = \frac{D(\rho)}{2n - (p_1 + p_2)},$$

The probabilistic symbolic regression models

- Substituting the estimates of β_1, β_2 and ϕ in the log-likelihood function (2), obtained from (1), it is possible re-estimate the correlation parameter plotting ρ against $l(\rho)$

$$l_i(\rho) = \phi^{-1}\{y_{i1}\hat{\theta}_1 + y_{i2}\hat{\theta}_2 - b(\hat{\theta}_1, \hat{\theta}_2, \rho)\} + \mathrm{c}(y_{i1}, y_{i2}, \rho, \tilde{\phi})] \quad (2)$$

- This iterative process can continue until convergence of the all parameters estimates.

The probabilistic symbolic regression models

- **Important Features**
  - The **Bivariate Generalized Linear Model (BGLM)**, is an extension of the generalized linear model (Nelder and Wedderburn, 1972);
  - The bivariate random vector $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2]$ can be represented by the lower and upper limits or the midpoint and half-range of the intervals;
  - The use of link function give a more flexibility to fit the BGLM model to an interval dataset;
  - Goodness-of-fit measures were proposed and inferential techniques can be applied to analyzed the fitted model.

The probabilistic symbolic regression models

- **Application to Real Interval Dataset**
  - Soccer interval dataset (http://www.ceremade.dauphine.fr/~touati/foot2.htm)
  - Y = weight, X_1 = height and X_2 = age
  - 531 soccer players grouped in 20 teams

COMPARISON OF SYMBOLIC REGRESSION METHODS IN SOCCER DATA SET.

Method	$RMSE_L$	$RMSE_U$	r_L^2 (%)	r_U^2 (%)
CM	7.54	7.68	40.95	23.18
CRM	1.95	2.66	57.75	26.57
BGLM	2.24	2.57	65.47	55.11

Future Works (short time)

- **Some ongoing….**
  - A simulated study (Monte Carlo) for a more consistent conclusion about the BGLM models;
  - Residual and Diagnostic measures will need to be extending for the BGLM model.
    - Coming soon
  - Development of symbolic regression models for modal and/or multi-categorical variables
    - Including inference techniques

References

- BILLARD, L.; DIDAY, E. Regression Analysis for Interval-Valued Data. In: Data Analysis, Classification and Related Methods. Proceedings of the 7th Conference of the International Federation of Classification Societies, Springer-Verlag, Belgium, 369-374, 2000;
- BILLARD, L.; DIDAY, E. Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley, New York, 2006;
- De CARVALHO, F.A.T., LIMA NETO, E. A. ; TENORIO, C. P. A new method to fit a linear regression model for interval-valued data. In: 27th Germany Conference on Artificial Intelligence. Lecture Notes in Computer Science. Berlin-Heidelberg : Springer, Ulm, v. 3238. p. 295-306, 2004;
- LIMA NETO, E. A. ; De CARVALHO, F. A. T. ; FREIRE, E. S. Applying Constrained Linear Regression Models to Predict Interval-Valued Data. In: 28th German Conference on Artificial Intelligence. Lectures Notes in Computer Science. Berlin Heidelberg : Springer-Verlag, Koblenz, v. 3698. p. 92-106, 2005;
- LIMA NETO, E. A. ; De CARVALHO, F. A. T. Nonlinear Regression Model to Symbolic Interval-valued Variables. In: 2008 IEEE International Conference on Systems, Man, and Cybernetics, Cingapura. v. 1. p. 1247-1252, 2008;

References

- LIMA NETO, E. A. ; De CARVALHO, F. A. T. Centre and Range Method for Fitting a Linear Regression Model to Symbolic Interval Data. Computational Statistics & Data Analysis, v. 52, p. 1500-1515, 2008;
- LIMA NETO, E. A. ; CORDEIRO, G. M. ; De CARVALHO, F. A. T. ; ANJOS, U. U. ; COSTA, A. G. Bivariate Generalized Linear Model for Interval-valued Variables. In: International Joint Conference on Neural Networks, Atlanta (aceito p/ publicação), 2009;
- MAIA, A. L. S.; De CARVALHO, F. A. T. Fitting a Least Absolute Deviation Regression Model on Interval-Valued Data. In: SBIA 2008. Advances in Artificial Intelligence, Heidelberg: Springer Berlin, Salvador, v. 5249. p. 207-216, 2008;
- NELDER, J.A.; WEDDERBURN, R.A. Generalized Linear Models. Journal of the Royal Statistical Society A, v. 135, 370-384, 1977;
- SOUZA, R. M. C. R.; CYSNEIROS, F. J. A.; QUEIROZ, D. C. F.; Fagundes, R.A.A. A Multi-Class Logistic Regression Model for Interval Data. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Singapore, p. 1253-1258, 2008;