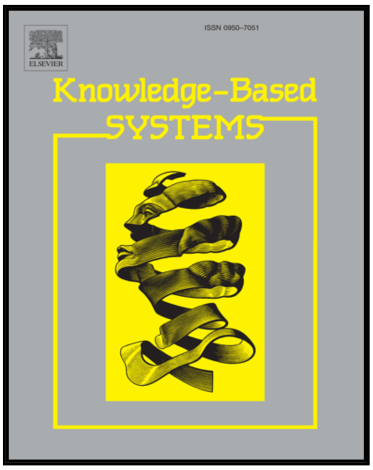

A Parametrized Approach for Linear Regression of Interval Data

Leandro C. Souza, Renata M.C.R. Souza, Getúlio J.A. Amaral,
Telmo M. Silva Filho

PII: S0950-7051(17)30286-1
DOI: 10.1016/j.knosys.2017.06.012
Reference: KNOSYS 3940

To appear in: *Knowledge-Based Systems*

Received date: 19 January 2017
Revised date: 11 April 2017
Accepted date: 6 June 2017

Please cite this article as: Leandro C. Souza, Renata M.C.R. Souza, Getúlio J.A. Amaral, Telmo M. Silva Filho, A Parametrized Approach for Linear Regression of Interval Data, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.06.012

A Parametrized Approach for Linear Regression of Interval Data

Leandro C. Souza[a], Renata M. C. R. Souza[b], Getúlio J. A. Amaral[c], Telmo M. Silva Filho[b]

[a] *Centro de Ciências Exatas e Naturais - CCEN/UFERSA, Av. Francisco Mota, 572 - Costa e Silva 59.625-900, Mossoró - RN, Brazil*
[b] *Centro de Informática - Cin/UFPE, Av. Jornalista Anibal Fernandes, s/n - Cidade Universitária 50.740-560, Recife - PE, Brazil*
[c] *Departamento de Estatística - DE/UFPE, Av. Jornalista Anibal Fernandes, s/n - Cidade Universitária 50.740-560, Recife - PE, Brazil*

Abstract

Interval symbolic data is a complex data type that can often be obtained by summarizing large datasets. All existing linear regression approaches for interval data use certain fixed reference points to model intervals, such as midpoints, ranges and lower and upper bounds. This is a limitation, because different datasets might be better represented by different reference points. In this paper, we propose a new method for extracting knowledge from interval data. Our parametrized approach automatically extracts the best reference points from the regressor variables. These reference points are then used to build two linear regressions: one for the lower bounds of the response variable and another for its upper bounds. Before the regressions are applied, we compute a criterion to verify the mathematical coherence of predicted values. Mathematical coherence means that the upper bounds are greater than the lower bounds. If the criterion shows that the coherence is not guaranteed, we suggest the use of a novel interval Box-Cox transformation of the response variable. Experimental evaluations with synthetic and real interval datasets illustrate the advantages and the usefulness of the proposed method to perform interval linear regression.

Keywords: Interval Linear Regression, Symbolic Data Analysis, Interval Parametrization

1. Introduction

Linear regression is related to the construction of models that explore linear dependency between variables. Two types of variables are involved: the response (or dependent) variable and the regressor (or independent) variables. The main goal is to

Email addresses: leandro.souza@ufersa.edu.br (Leandro C. Souza), rmcrs@cin.ufpe.br (Renata M. C. R. Souza), gjaa@de.ufpe.br (Getúlio J. A. Amaral), tmsf@cin.ufpe.br (Telmo M. Silva Filho)