Error localization as a mixed integer problem with the `editrules` package

package version 1.1.0

Edwin de Jonge and Mark van der Loo

August 29, 2011

Abstract

This vignette is far from finished. Version 2.0 fo the package will have the full vignette. At the moment, functionality for solving error localization problems with lp solvers is experimental.

Contents

1 Introduction

?, ?

2 Error localization as a mixed integer problem

2.1 Previous work

2.2 Numerical stability

3 Formulation as a mixed integer problem

3.1 Numerical edits

Numerical variables in practice are always bounded, e.g. age, income, height. In our formulation of the error localization problem as a mixed integer problem we use the boundaries explicitly: first of all it allows an easy to understand and to implement equations, but our formulation also improves numerical stability for suitable choosen boundaries. We use greek symbols to note constants.

The record to be checked has the following (numerical) values:

$$x_i = \chi_i \tag{1}$$

We assume each variable x_i is bounded by α_i and β_i

$$\alpha_i \leq x_i \leq \beta_i \tag{2}$$

For each variable x_i we introduce a binary variable $u_i \in \{0, 1\}$ and we add two edits to the original edit matrix that conditionally constrain the value of x_i to χ_i

$$\begin{aligned} x_i &\leq \chi_i + (\beta_i - \chi_i)u_i && (3)\\ (\alpha_i - \chi_i)u_i + \chi_i &\leq x_i && (4) \end{aligned}$$

Variable u_i signifies if x_i should be adapted: if $u_i = 0$ these edits reduce to equation 1, meaning that the reported χ_i is assumed correct. When $u_i = 1$ the edits reduce to equation 2, meaning that the value of x_i is unconstrained and can take any feasible value.

Felligi Holt, minimize the weigthed sum of adaptations.

The error localization problem can now be formulated as the mixed integer problem: Minimize $\sum_i w_i u_i$, with:

$$\begin{array}{rcrcl} a_i x_i & & & \odot_i & b_i \\ x_i & - & \beta_i^* u_i & \leq & \chi_i \\ -x_i & + & \alpha_i^* u_i & \leq & -\chi_i \end{array} \tag{5}$$

3.2 Implementation

3.3 categorical edits

3.4 mixed edits

4 Boundary heuristics

4.1 From data set

Minimum and maximum value of each variable in the data set to be checked gives reasonable boundaries for the dataset.

4.2 From observation, per record

Assume that maximum values for each observation differ at most a factor f (e.g. 1000) of the reported value.

5 Discussion

Is a usefull addition to `editrules`, finds quickly solutions to error localization problems with hundreds to thousands of variables.

Solutions given by current lp solvers can be numerical unstable, which may result in a false positive or a false negative solution. Luckily `editrules` contains `substValue` and `isFeasible` that can be used together to check the validity of a solution. Furthermore several heuristics can be used to increase the numerical stability by using smaller boundaries for the variables.

5.1 Comparison to `backtracker`

errorLocalizer is more complete, offers a more complete tool box for finding an optimal solution. It can also find more equivalent solutions, which is not possible or difficult with MIP solvers.

However when speed of finding a solution matters, it is