Error localization

Mark van der Loo, Statistics Netherlands

CBS, Department of Methodology

Complutense University of Madrid, Spring 2019

UNIVERSIDAD COMPLUTENSE MADRID

Error localization

Data validation and error localization answer different questions.

Data validation

Which errors are there?

Error localization

Where do I need to make changes to fix the errors?

Example

Ruleset

```
age >= 0
age <= 120
if (drivers_licence == TRUE) age >= 18
```

Data

age	drivers_licence
10	TRUE

Question:

Which field or fields would you change?

Error localization

Definition

Error localization is a procedure that points out fields in a data set that can be altered or imputed in such a way that all validation rules can be satisfied.

Example

Ruleset

```
if (married == TRUE ) age >= 16
if (attends == "kindergarten") age <= 6
```

Data

age	married	attends
3	TRUE	kindergarten

Question

Which field or fields would you change?

UNIVERSIDAD COMPLUTENSE MADRID

Principle of Fellegi and Holt

Find the minimal (weighted) number of fields to adjust such that all rules, including implied rules, can be satisfied.

IP Fellegi and D Holt, JASA **71** 353 17–35 (1976).

Note

This should be used as a last resort, when no further information on the location of errors is available.

Implied rules?

```
turnover - total.cost == profit
              profit <= 0.6 * turnover
```

This implies (substituting profit):

```
          total.cost >= 0.4 * turnover
```

We need to take into account such *essentially new* rules (edits) —unstated relations between variables that can be derived from the explicitly defined rules.

UNIVERSIDAD COMPLUTENSE MADRID

Error localization is a set covering problem

Given a record $\boldsymbol{x} = (x_1, \ldots, x_n)$, restricted by a set of explicit and implicit validation rules. Minimize the sum

$$\sum_{j \in S} w_j, \; w_j > 0,$$

Where $S \subseteq \{1, \ldots, n\}$ such that all (explicit and implicit) validation rules can be satisfied by replacing $x_j, j \in S$ with new values.

Algorithms: De Waal *et al* (2011), John Wiley & Sons.

Solution types

(1) Feasible solution

Any set of variables that can be altered so that all rules, including implied ones can be satisfied. Example: change every variable.

(2) Feasible solution of minimal size

A set of variables that is a feasible solution and as small as possible.

(3) Feasible solution of minimal weight

A set of variables that is a feasible solution and minimizes the total weight.

Choosing weights

All weights equal (usually to one)

Least nr of variables adapted. In case of multiple solutions: choose randomly (e.g. by adding a small random perturbation to the weights).

Weights represent reliability

Heigher weight $\rightarrow$ variable is less likely chosen.

- Can be made to depend on ‘outlierness’, or expert judgement.
- Possible problem: minimal weights vs minimal nr of variables?

Choosing weights

Question

Is it possible to choose a set of weights, such that

1. The smallest number of variables is chosen
2. The weights are minimized

Intuition

If the weights do not differ too much, no extra variables will be introduced on top of the variables in a feasible solution of of minimal size.

Proposition

Given a set of weights $\boldsymbol{w} = (w_1, \ldots, w_n)$ and write

$$w_j = 1 + \delta_j \text{ with } 0 \leq \delta_j \leq \delta^{max}.$$

If $\delta^{max} < 1/(n-1)$ then a feasible solution of minimal weight is a feasible solution of minimal size.

Proof: van der Loo (2015) Rom. Stat. Rev. **2** 141–152; SDCR §7.5

UNIVERSIDAD COMPLUTENSE MADRID

Scaling rule

Given set of n weights $\boldsymbol{w} = \mathbf{1} + \boldsymbol{\delta}$. Then the following rescaling ensures that a solution of minimal weight also a solution of minimal size.

$$w'_j = 1 + \frac{w_j - w^{min}}{w^{max} - w^{min}} \times \frac{1}{n}$$

Procedure

1. Determine reliability weights (outlierness, expert judgement...)
2. Apply scaling rule