NISS

Decision-Theoretic Framework for Data Quality

Alan Karr
March 3, 2005

Summary

- Specific context and testbed database
- Specific set of DQ strategies
- Evaluation of strategies
- Predictive statistical models

Notation

- $\mathcal{D}^{\mathrm{true}}$ = true database (flat file of cases × attributes
- $\mathcal{D}^{\mathrm{pre}}$ = database prior to clean-up
- S = clean-up strategy
- $\mathcal{D}^{\mathrm{post}}(S)$ = database resulting from applying S to $\mathcal{D}^{\mathrm{prc}}$

Measuring Effectiveness

Conceptually,

$$\mathrm{Eff}(S) = d(\mathcal{D}^{\mathrm{post}}(S), \mathcal{D}^{\mathrm{true}}),$$

where d is a data quality metric

Inference-Based Effectiveness

More meaningfully,

$$\mathrm{Eff}(S, P, \mathcal{D}^{\mathrm{pre}}) = d_P(\mathcal{D}^{\mathrm{true}}, \mathcal{D}^{\mathrm{pre}}) - d_P(\mathcal{D}^{\mathrm{true}}, \mathcal{D}^{\mathrm{post}}(S)),$$

where

- P = inference procedure that can be applied to the data
- d_P = function measuring the difference in the results of P applied to two different databases

What if Truth is Not Known?

Use

$$\mathrm{Eff}^{\mathrm{naive}}(S, P, \mathcal{D}^{\mathrm{pre}}) = d_P(\mathcal{D}^{\mathrm{pre}}, \mathcal{D}^{\mathrm{post}}(S)).$$

Relevant points:

- $+$ sign for $\mathrm{Eff}^{\mathrm{naive}}(S, P, \mathcal{D}^{\mathrm{pre}})$ may not signal improvement
- Small values of $\mathrm{Eff}^{\mathrm{naive}}(S, P, \mathcal{D}^{\mathrm{pre}})$ mean no improvement

Prediction

- $\{S(\theta) : \theta \in \Theta\}$ = parameterized family of clean-up strategies
- Goal: solve

$$\theta^* = \arg\max_{\theta} \mathrm{Eff}(S(\theta), P, \mathcal{D}^{\mathrm{pre}})$$

- Problem: only know $\mathrm{Eff}(\theta)$ for a few values of θ

Predictive Models

Build statistical model

$$\widehat{\mathrm{Eff}}(\theta) = f(\theta) + \text{uncertainty}$$

Challenges:

- "Form" of the model
- Nature of the uncertainties
- What data are necessary to fit the model
- Validation