

Statistical relational learning based automatic data cleaning

Weibang LI, Ling LI, Zhanhuai LI, Mengtian CUI

Frontiers of Computer Science, DOI: 10.1007/s11704-018-7066-4

Problems & Ideas

- Problems of automatic data cleaning
 - Most of existing approaches for data cleaning detect and repair dirty data by leveraging existing constraints or patterns.
 - Given dirty data, the problem is how to clean dirty data without existing constraints or patterns.
- Ideas: automatic data cleaning based on statistical relational learning and probabilistic inference
 - We present an unsupervised data cleaning framework based on statistical relational learning.
 - We propose an algorithm to generate first-order logic formulas based on the dependency relationships between attributes.

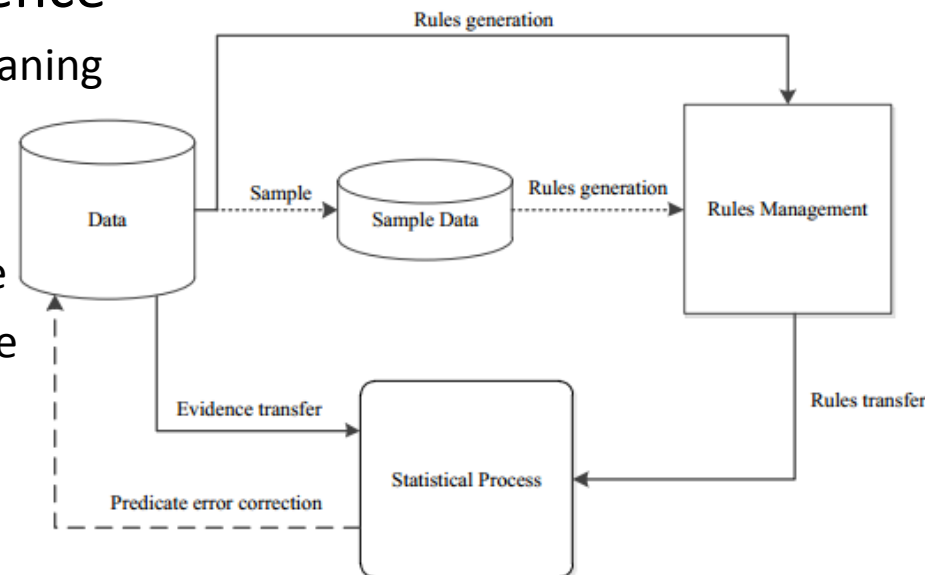


Fig. 1 Framework of the unsupervised data cleaning approach

Main Contributions

- ADCS outperforms BAYC and BYWP in terms of Precision, Recall and Fmeasure with the increase of the dataset size from 10K to 50K

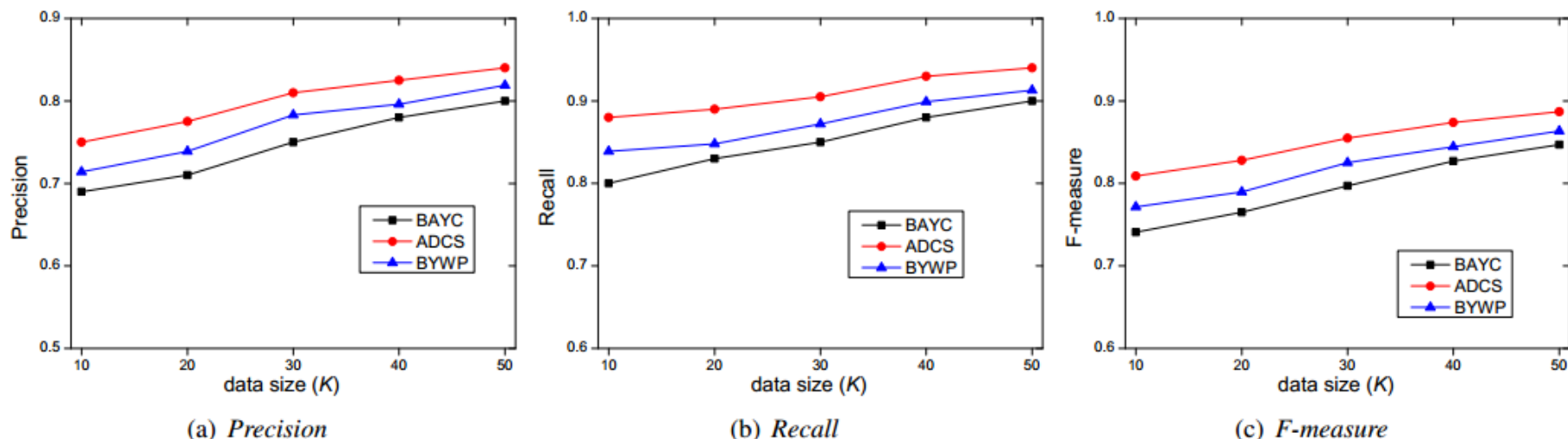


Fig. 6 Accuracy of Algorithm ADCS, BAYC and BYWP with $noi\%=6\%$ in terms of HPT datasets size.

- $ADCS_T$ surpasses $ADCS_R$ remarkably in accuracy on different datasets sizes

