Polishing: Enhancing Data Quality by Repairing Imperfections

Choh Man Teng

Investigators

- Choh Man Teng
  - Institute for Human and Machine Cognition, University of West Florida
- Henry E. Kyburg, Jr.
  - Philosophy and Computer Science, University of Rochester
  - Institute for Human and Machine Cognition, University of West Florida
- Jeffrey Scargle
  - NASA Ames Research Center

Motivation

- Examples of data imperfections.
  - Transmission errors
  - Missing values
  - Typos
- Imperfect data is less useful than clean data but still contains valuable information.
- In many cases perfect data is unobtainable.

Objectives

- Preserve and extract information from imperfect data, rather than discarding the corrupted portions.
- Improve data quality automatically or semi-automatically.
- Identify and repair corrupted or imperfect data.
- A general “preprocessor”.
  - Polished data can be used by a host of data analysis algorithms that do not have to individually cope with noise.

Specification

- Input: (possibly) imperfect data.
- Polishing: coupled with modelling.
- Output:
  - (automated) repaired data; and/or
  - (semi-automated) analysis of corruption with suggested repairs.

Technical Approach

- Main idea
  - Capitalize on the dependency relationship between different parts of the data.
- Polishing
  - Choose a model.
  - Prediction: identify corrupted parts and suggest correction.
  - Adjustment: selectively carry out the suggested changes.

Technical Approach (Cont'd)

- Prediction
  - Predict data values based on models built from original data.
  - Candidate repairs: derived from disagreement between predicted and observed/recorded values.
- Adjustment
  - Identify from candidate repairs the corrections to be implemented.
  - Criteria: improvement in overall fitness/accuracy of the model.

Test Data

- Benchmarks
  - UCI machine learning repository.
- Simulated data
  - Known structure and noise properties.
  - e.g., market basket transactions based on distribution of customer behavior.
- Time series data
  - Hit rate of X-ray photons (NASA)
- Others
  - Navy recruiting data (ONR)

Preliminary Findings

- Comparison of different approaches.
  - Robust algorithms: allowing for noise by avoiding overfitting the model to the data.
  - Filtering: eliminating the corrupted portions.
  - Polishing: correcting the corrupted portions.
- Polished data resulted in significant improvement in classification accuracy for many of the benchmark data sets.
- A combined approach might be more beneficial.

Preliminary Findings (Cont'd)

- Treatment of continuous attributes.
  - Discretization: deciding the cut points.
    - Binning based on spread and/or weight.
    - Information entropy minimization.
  - Using statistical parameters within discretized subgroups to identify suspects and appropriate repairs.

Preliminary Findings (Cont'd)

- Alternative models examined
  - Decision trees
  - Linear regression
  - Association rules
    - Qualitative inference rules compiled from a database.

Expected Impact

- Simplify data analysis.
  - Methods can focus on scientific objectives rather than imperfection tolerance.
- Extract more information.
  - Higher quality data leads to higher quality results.
- Automated noise/anomaly detection.

References

- Choh Man Teng. Learning from dissociations. To appear in *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery*, 2002.

- Choh Man Teng. A comparison of noise handling techniques. In *Proceedings of the International Florida Artificial Intelligence Research Symposium*, pages 269-273, 2001.

- Choh Man Teng. Evaluating noise correction. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, Springer-Verlag, 2000.

- Choh Man Teng. Correcting Noisy Data. In *Proceedings of the International Conference on Machine Learning*, pages 239-248, 1999.

Publications (2001-2002)

- Choh Man Teng. Learning from dissociations. To appear in *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery*, 2002.
- Henry E. Kyburg, Jr. and Choh Man Teng. Randomization and uncertain inference. To appear in *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, 2002.
- Henry E. Kyburg, Jr. and Choh Man Teng. The logic of risky knowledge. To appear in *Proceedings of the Ninth Workshop on Logic, Language, Information and Computation*, 2002.
- Choh Man Teng and Rattikorn Hewett. Associations, statistics, and rules of inference. To appear in *Proceedings of the International Conference on Artificial Intelligence and Soft Computing*, 2002.
- Henry E. Kyburg, Jr. and Choh Man Teng. Random sampling and uncertain inference. To appear in *Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2002.
- Henry E. Kyburg, Jr. and Choh Man Teng. Evaluating interpretations of probability in the face of drifts. To appear in *Proceedings of the International Conference on Artificial Intelligence*, 2002.

Publications (Cont'd)

- Rattikorn Hewett, John Leuchner, Choh Man Teng, Sean D. Mooney and Teri E. Klein. Compression-based induction and genome data. To appear in *Proceedings of the International Florida Artificial Intelligence Research Symposium*, 2002.
- Henry E. Kyburg, Jr. and Choh Man Teng. Evaluating defaults. In *Proceedings of the Ninth International Workshop on Non-Monotonic Reasoning*, pages 257-264, 2002.
- Henry E. Kyburg, Jr. and Choh Man Teng. Induction and consistency. In *Common Sense, Reasoning, and Rationality*, Oxford University Press, 2002.
- Henry E. Kyburg, Jr. and Choh Man Teng. *Uncertain Inference*, Cambridge University Press, 2001.
- Choh Man Teng. A comparison of noise handling techniques. In *Proceedings of the International Florida Artificial Intelligence Research Symposium*, pages 269-273, 2001.
- Rattikorn Hewett and Choh Man Teng. On making data mining truly useful to end-users. In *Proceedings of the AFCEA Federal Database Colloquium*,pages 179-185, 2001.