

Properties of Inconsistency Measures for Databases

Hernan Razo
Samantha Berg
Christopher Anthony

Introduction

- The study focuses on the inconsistency in databases and the methods used to quantify these inconsistencies.
- Specific inconsistency measures are listed and described.
- Authors also propose a new measure of inconsistency.
- Verify the proposed inconsistency measure through a case study of HoloClean System.
 - Compared results of inconsistency measures against the popular and established HoloClean System, a statistical inference engine to impute, clean, and enrich data.



Integrity Constraints and Inconsistent Databases

- **Integrity constraints** - sets of rules used to maintain the quality of data when performing operations.
 - example: A column for phone numbers should not have letters
- **Inconsistent database** - A database that violates its integrity constraints



Why Inconsistency in Databases?

There can be various reasons for inconsistency in a database:

- Data Generation
- Data Integration
- Imprecise Data
- Difference in integrity constraints across databases.



Measuring Inconsistency of a Database

- Authors focus on the following established inconsistency measures:
 - I_d - The indicator function of inconsistency
 - I_p - The count of facts that belong to the minimal inconsistent subset
 - I_{MI} - MI Shapely Inconsistency. The cardinality of the set.
 - I_{MC} - The set of all maximal consistent subsets of a database
 - I_R - The minimal cost of a sequence of operations that repairs a database



Properties of Inconsistency Measures

- Inconsistency measures have four properties:
 - **Positivity** - An inconsistency measure is strictly positive if and only if the database is inconsistent
 - **Monotonicity** - Inconsistency cannot be reduced if the constraint gets more strict
 - **Continuity** - A single operation can have a limited relative impact on inconsistency
 - **Progression** - An operation can always be found to reduce inconsistency



Proposed Measure

- The new measure is a linear relaxation of the Measure I_R (the minimal number of tuples to be removed to satisfy the minimum constraints)
- The ILP of I_R is used with a change in the second equation.
- In the first equation, we ensure that for every Minimal inconsistent subset that violate the constraint we remove at least one tuple to resolve the violation.
- The result of the ILP is the value of the measure I_R .

$$\text{Minimize : } \sum_{i \in ids(D)} x_i \cdot \kappa(\langle -i \rangle(\cdot), D) \text{ subj. to:}$$

$$\forall E \in Ml_{\Sigma}(D) : \sum_{i \in ids(E)} x_i \geq 1 \quad (1)$$

$$\forall i \in ids(D) : x_i \in \{0, 1\} \quad (2)$$

Proposed Measure

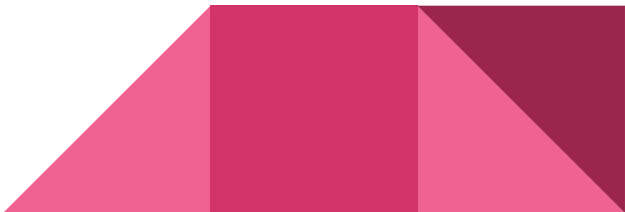
- In the second equation x_i is 0 or 1 based on which the tuple is removed or kept:
 $x_i = 1$: remove the tuple $x_i = 0$: keep the tuple.
- The second equation is changed to $0 \leq x_i \leq 1$
- In the New Measure the second condition is replaced with a new condition " $\forall i \in \text{ids}(D) : 0 \leq x_i \leq 1$ " it allows to use any value between 0 and 1 for the variable.

$$\text{Minimize : } \sum_{i \in \text{ids}(D)} x_i \cdot \kappa(\langle -i \rangle(\cdot), D) \text{ subj. to:}$$

$$\forall E \in \text{Ml}_{\Sigma}(D) : \sum_{i \in \text{ids}(E)} x_i \geq 1 \quad (1)$$

$$\forall i \in \text{ids}(D) : x_i \in \{0, 1\} \quad (2)$$

Repair Inconsistency Measures

- A repair operation can include deletion or insertion of a tuple, or an update of attribute value.
 - Utilize inconsistency measure for implementing a progress for data repairing.
 - A measure of inconsistency can be utilized for computing and suggesting actions in data repairing as well as address types that contain highest responsibility to the inconsistency level or the might result in the greatest reduction in inconsistency.
 - To effectively communicate progress evidence in repairing, the measure should include certain characteristics.
- 

Repair Inconsistency Measures

- To allure the repairing process, the measures should be acknowledged of the underlying repairing operations.
- A repair system is a group of repairing operations with an identical cost of applying to a given database.
- For a repair system R , denote by $R^* = (O^*, k^*)$ the repair system of all sequence operations from R , where the cost of a sequence is the sum of costs of the distinctive operations.



Experimental Study

8 Datasets are used for the experiment and random noise is added using one of two algorithms:

- Constraint Oriented Noise (CONoise)
 - Randomly Select a Constraint
 - Randomly Select 2 tuples from the Database
 - Modify value in one of the tuples to violate constraints
- Random Noise (RNoise)
 - Randomly select a database cell.
 - Change its value to another value from the active domain or to a typo.

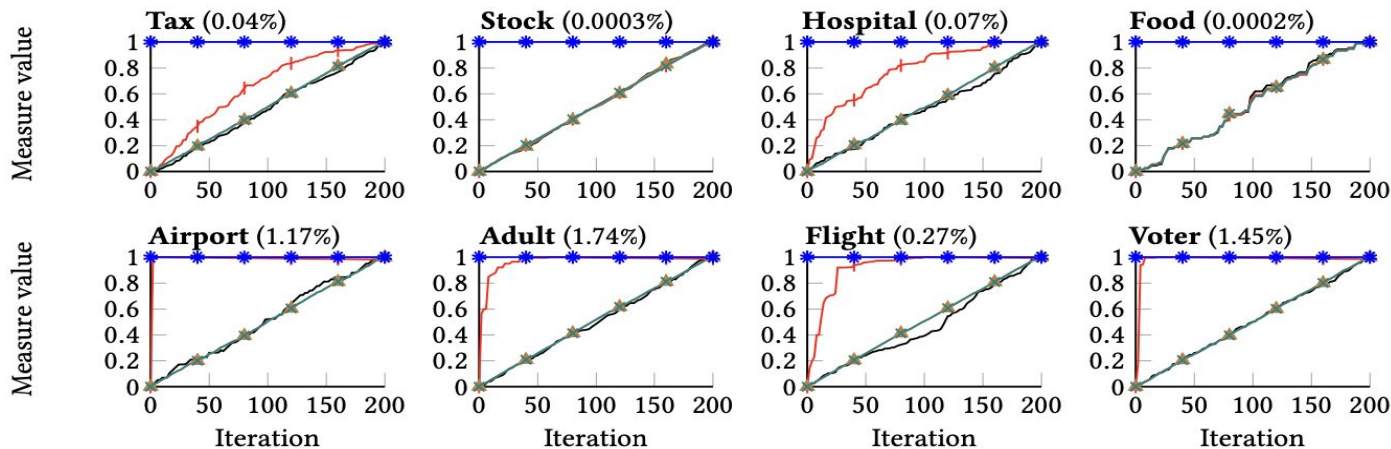


Experimental Study

- In each iteration, one tuple or cell is changed.
- After modifying the databases, the Measure values are then computed.

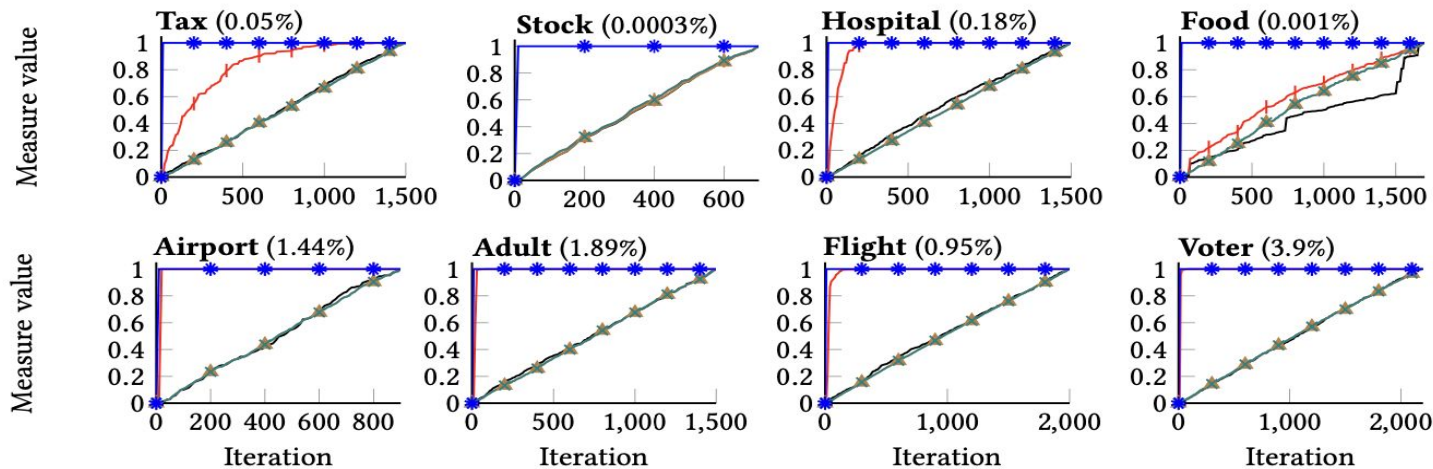


Comparing Measure Values for Noise added with CONoise



(a) Noise added with CONoise.

Comparing Measure Values for Noise added with RNoise



(b) Noise added with RNoise ($\alpha = 0.01$ and $\beta = 0$).

Case Study: HoloClean

- The HoloClean system is a popular and establishes cleaning system for imputing, cleaning, and enriching data. The authors compare HoloClean with the inconsistency measures listed previously.
- The HoloClean system features one-shot automatic cleaning. The system was simulated by providing 1 constraint at a time.
- The measure was computed after every step.
- It was observed that I_d and I_p fall short of effectively indicating progress. Contrarily, the other measures, particularly I_R and $\mathcal{I}_{\mathcal{R}}^{\text{lin}}$ are able to capture the reduction in the inconsistency level, and show an almost linear decay as desired.

