On tabular constraint learning

March 9, 2016

1 Headers

Can be in columns or in row, see examples. Needs to be inferred using type information.

2 General Algorithm: Infer-Learn-Suggest – ILS

Infer Datatypes, headers, i.e. variable names, and table layouts, borders of tables, the number of tables etc

Learn On each level, i.e., table-wise or column-row-wise learn applicable constraints

Suggest If there is a row or a column with missing data (not-completed cells), then suggest values satisfying the constraints

Types of Constraints

Most of the constraints I have seen are numeric (finance, accounting) and occasionally strings, also emails and dates are present but their quantity is negligible.

Types of constraints exact or noisy(way harder)

- Ascending/Descending (all types)
- Ordered by (all types)
- All different (all types, similar to the key detection)
- Arithmetics (only for number type + possible ranges over two, N, all columns-rows or their consecutive subset)
- Concatenation (strings + any other type)
- Common substring?
- Column row-inclusion, sort of FD, all values of a row or a column are present in the other row or a column
- Aggregates? Counting
- Key – foreign or primary