PClean: Bayesian Data Cleaning at Scale via Domain-Specific Probabilistic Programming

Alexander K. Lew, Monica Agrawal, David Sontag, Vikash K. Mansinghka
MIT

1. Overview

PClean is a domain-specific probabilistic programming language for inferring ground-truth relational databases from flat, dirty datasets.

<table>
<thead>
<tr>
<th>Name</th>
<th>Specialty</th>
<th>Degree</th>
<th>School</th>
<th>Address</th>
<th>City</th>
<th>State</th>
<th>Zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>K. Ryan</td>
<td>Family Medicine</td>
<td>MD</td>
<td>PCOM</td>
<td>6317 York Rd</td>
<td>Baltimore</td>
<td>MD</td>
<td>21212-2310</td>
</tr>
<tr>
<td>K. Ryan</td>
<td>Family Medicine</td>
<td>MD</td>
<td>UMD</td>
<td>100 Walter Ward Blvd</td>
<td>Abingdon</td>
<td>MD</td>
<td>21009-1285</td>
</tr>
<tr>
<td>E. Evans</td>
<td>Internal Medicine</td>
<td>MD</td>
<td>UMD</td>
<td>100 Walter Ward Blvd</td>
<td>Abingdon</td>
<td>MD</td>
<td>21009-1285</td>
</tr>
<tr>
<td>M. Grady</td>
<td>Physical Therapy</td>
<td>PT</td>
<td>Other</td>
<td>3401 Marriott Blvd</td>
<td>Abingdon</td>
<td>MD</td>
<td>21009-2000</td>
</tr>
</tbody>
</table>

2. PClean Modeling Language

```
class School begin
  name = string_prior(1,100)
  preferring school_names
  degree_direct = dirichlet(one(100 degrees))
end

class Physician begin
  name = string_prior(100)
  preferring
  observed_cities
  state = uniform(states)
end

class Practice begin
  city = City
  bad_city = topos(city)
end

class Record begin
  physician = Physician
  location = Practice
end
```

```
Inferred relational database R
```

```
name  | degree | school
K. Ryan | Family Medicine | PCOM
K. Ryan | Family Medicine | UMD
E. Evans | Internal Medicine | UMD
M. Grady | Physical Therapy | Other

Dirty observations

PClean program

3. PClean Inference Engine

1) Initializes latent database with per-observation SMC; fixes mistakes with per-object rejuvenation
   - Non-parametric prior admits a sequential representation (right), enabling SMC that incorporates one observation at a time
   - Exchangeability of CRPs enables per-object rejuvenation moves: choose any object, and propose new parameters and reference slots — possibly creating new objects of other classes as their targets

2) Fast, data-driven SMC / block rejuvenation proposals created just-in-time by a proposal compiler
   - Translates subproblem to a Bayes net
   - For each pattern of missingness in the data, compiles efficient enumeration-based SMC proposals
   - As needed, compiles enumeration-based rejuvenation proposals
   - When possible exploits conjugacy for continuous parameters

3) User-specified inference hints help scale to large datasets and variable domains
   - subproblem begin ... end:
     Group adjacent statements into a subproblem which becomes an intermediate target in SMC. Smaller subproblems lead to more scorable enumerative proposals, at some cost to proposal quality
     
```
x ← d_E(e, δ, E) preferring E:
```

4. Experiments & Results

PClean is expressive enough to model data cleaning benchmarks in <50 LoC, and unlike generic PPL inference (left), achieves SOTA accuracy + runtime compared to weighted logic + machine learning baselines.

Table 1: Results of PClean and various baseline systems on three diverse cleaning tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>Flights</th>
<th>Hospital</th>
<th>Rents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PClean</td>
<td>HoloClean (Unpublished)</td>
<td>HoloClean</td>
<td>NADEF</td>
</tr>
<tr>
<td>F1</td>
<td>0.90</td>
<td>0.64</td>
<td>0.41</td>
<td>0.07</td>
</tr>
<tr>
<td>Time</td>
<td>3.1s</td>
<td>45.4s</td>
<td>32.6s</td>
<td>9.1s</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.91</td>
<td>0.83</td>
<td>0.27s</td>
</tr>
<tr>
<td>Time</td>
<td>4.5s</td>
<td>1m 10s</td>
<td>1m 32s</td>
<td>22.8s</td>
</tr>
<tr>
<td>F1</td>
<td>0.69</td>
<td>0.48</td>
<td>0.48</td>
<td>0</td>
</tr>
<tr>
<td>Time</td>
<td>1m 29s</td>
<td>20m 16s</td>
<td>13m 43s</td>
<td>13s</td>
</tr>
</tbody>
</table>

```
Comparison of Inference Algorithms
```

---

Left: A PClean program defines a relational schema for a database of objects underlying the dirty data, along with a probabilistic relational model over object attributes.

Below: PClean uses a domain-general non-parametric structure prior over the number of objects of each class, and over their relationships.

- GenerateSkeleton(C, D): Create C, each C ∈ D per observed record
- Generate a class C for each class in the dirty dataset
- Select references to class C: Ref(C) ∈ (r, y) ∈ r, δₕ(C'
- Assign reference slots pointing to C for object r ∈ C
- For object r ∈ C, do
  - r = y
  - return Ref(C)