The design of algorithms for the production of training data

Élie de Panafieu, Quentin Lutz, Alexander Scott, Maya Stein

Nokia Bell Labs, Oxford University, University of Chile, Randnet project

Reading group network theory Lincs 2021

Optimal complexity

All optimal algorithms have the same distribution on the number of queries. The generating function of this distribution

$$P(z,u) = \sum_{\text{partition } p} u^{\text{queries}(p)} \frac{z^{|p|}}{|p|!}$$

is characterized by $P(0,u) = 1$ and

$$\partial_z P(z,u) = P(zu,u) e^{zu}.$$

Asymptotic optimal average query number and standard deviation

$$E \sim \frac{\binom{n}{2}}{\log(n)}, \quad \sigma \sim \frac{n^{3/2}}{\sqrt{\log(n)}}.$$

We expect a Gaussian limit law.

Proof

Analysis of a particular chordal algorithm.

Clique Algorithm.

- Ask queries between the largest vertex and all others
- The block of the largest label is now discovered
- Remove this block and continue with the remaining vertices.

A partition is then decomposed as a pair

$$(\text{partition, block of the largest vertex}).$$

Example: $\{\{1,3\},\{4\},\{2,5,6\}\} \mapsto (\{\{1,3\},\{4\}\}, \quad \{2,5\})$

Proof (exact results)

Example: $\{\{1,3\},\{4\},\{2,5,6\}\} \mapsto (\{\{1,3\},\{4\}\}, \quad \{2,5\})$

Symbolic method:

$$P(z) := \sum_n B_n \frac{z^n}{n!}, \qquad \qquad \partial_z P(z) = P(z) e^z$$

or recurrence

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k .$$

In fact, we know $P(z) = e^{\exp(z)-1}$.

Proof (exact results)

Example: $\{\{1,3\},\{4\},\{2,5,6\}\} \mapsto (\{\{1,3\},\{4\}\}, \quad \{2,5\})$

Symbolic method:

$$P(z) := \sum_n B_n \frac{z^n}{n!}, \qquad \partial_z P(z) = P(z) e^z$$

or recurrence

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k.$$

In fact, we know $P(z) = e^{\exp(z)-1}$.
An addtional variable is added to mark the query number.

Proof (exact results)

Example: $\{\{1,3\},\{4\},\{2,5,6\}\} \mapsto (\{\{1,3\},\{4\}\}, \quad \{2,5\})$

Symbolic method:

$$P(z,u) := \sum_p u^{\text{queries}(p)} \frac{z^{|p|}}{|p|!}, \qquad \partial_z P(z,u) = P(zu,u)e^{zu}.$$

or recurrence

$$B_{n+1,q} = \sum_{k=0}^{n} \binom{n}{k} B_{k,q-n}.$$

In fact, we know $P(z) = e^{\exp(z)-1}$.
An addtional variable is added to mark the query number.

Proof (asymptotics)

The asymptotics number of partitions is classicaly obtained using a saddle-point method on the Cauchy integral representation of the coefficient extraction

$$B_n = n![z^n]e^{\exp(z)-1} = \frac{n!}{2i\pi}\oint e^{\exp(z)-1}\frac{dz}{z^{n+1}}$$

Probability generating function

$$\mathrm{PGF}_n(u) := \sum_{q\geq 0}\mathbb{P}(q \text{ queries} \mid |p| = n)u^q = \frac{n!}{B_n}[z^n]P(z,u).$$

Expected number of queries

$$E_n = \partial_{u=1}\,\mathrm{PGF}_n(u) = \frac{n!}{B_n}[z^n]\partial_{u=1}P(z,u)$$

Solve the differential equation in $\partial_{u=1}P(z,u)$

$$\partial_z\partial_{u=1}P(z,u) = \partial_{u=1}(P(z,u)e^{uz}) = (\partial_{u=1}P(z,u))e^z + zP(z)e^z.$$

Open problems

Characterize optimal active clustering algorithms for non-uniform distributions on set partition.

If the queries are chosen randomly (but not trivial), what is their asymptotic expected number?

Given a (non-chordal) agregated graph, how hard is it to find the optimal queries? (NP-complete?)

Best use of a heuristic distance between elements.