1. Motivation

Data analytics is challenging due to large scale and variety. Smart algorithms may cut cost but needs fast algorithms.

2. Thesis Statement and Contributions

We use VC-dimension to obtain (probabilistically-quantified) high-quality approximations for many data analytics tasks by processing only a small random sample of the data.

3. Approximations and Limitations of Classic Approach

Tradeoff between sample size and quality of approximation is well studied.

Given \( n, \alpha, \beta \) such that \( \alpha \cdot n + \beta \cdot n = \varepsilon \), classic bounds cannot handle Big Data Variety (i.e., high values of \( \varepsilon \)).

Our goal: compute sample size \( s \) to obtain \( (\varepsilon, \beta) \)-approximation.

Classical bounds such as Chernoff bound and union bound sample size is:

\[
|S| = O \left( \frac{1}{\beta} \left( \frac{\alpha + \beta}{\varepsilon} \right) \right)
\]

Theorem: Assume \( |C| \leq D \), let \( 0 < \alpha, \beta < 1 \) be a probabilistic distirbution on \( D \) and \( s \) be a collection of samples from \( \mathbb{X} \) w.r.t. \( \varepsilon \) with

\[
|S| = O \left( \frac{1}{\beta} \left( \frac{\alpha + \beta}{\varepsilon} \right)^2 \right)
\]

Then, with probability \( 1 - \varepsilon \),

\[
|S| = O \left( \frac{1}{\beta} \left( \frac{\alpha + \beta}{\varepsilon} \right)^2 \right)
\]

...-approximation of \( (\varepsilon, \beta) \)-approximation.

4. Vapnik-Chervonenkis (VC)-Dimension

Theorem: \( VC(D, F) \geq \log |F| \) for any \( D \subseteq \mathbb{F}_2 \).

Assume \( VC(D, F) \leq s \), let \( 0 < \alpha, \beta < 1 \) be a probabilistic distrbution on \( D \) and \( s \) be a collection of samples from \( \mathbb{X} \) w.r.t. \( \varepsilon \) with

\[
|S| = O \left( \frac{1}{\beta} \left( \frac{\alpha + \beta}{\varepsilon} \right)^2 \right)
\]

5. Mining Frequent Itemsets and Assoc. Rules (1)

6. Mining Frequent Itemsets and Assoc. Rules (2)

7. Estimating Betweenness Centrality

Betweenness centrality: measure of vertex importance in graphs.

Our goal: fast computation of \( (\varepsilon, \beta) \)-approximation using sampling algorithm.

Exact algorithm for \( (\varepsilon, \beta) \)-approximation takes time \( O(n^2 \log n) \).

Our goal: fast computation of \( (\varepsilon, \beta) \)-approximation using sampling algorithm.

8. Estimating the Selectivity of Database Queries

DBMS must choose plan with smallest execution time.

DBMS use histograms: independence, uniformity assumption — not true.

Selectivity of a query \( q \) is \( \sigma(q) \) is in output of \( q \) and \( \sigma(q) \) is in input of \( q \).

Exact value not available before execution, must approximate.

Theorem: Let \( \alpha, \beta \) be max no. of join operations in a query from \( q \).

Then, \( \sum_{q} \alpha \cdot \beta \) is \( O(m^2) \).

Evaluation: Sample can fit into main memory — estimation is fast. Estimate much closer to real value than guaranteed.

Beat PostgreSQL and SQLServer by orders of magnitude.