电子科技大学
University of Electronic Science and Technology of China

A Survey Of Synopsis Construction In Data Streams

TanYue

Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

1.Introduction

数据挖掘实验室
Data Mining Lab

Computational & Storage costs &The large volume

Desiderata:

- Broad Applicability
- One Pass Constraint
- Time and Space Efficiency
- Robustness
- Evolution Sensitive

Methods:

- Sampling methods
- Histograms
- Wavelets
- Sketches
- Micro-cluster based summarization

2. RoadMap:

数据挖掘实验室
Data Mining Lab

2. Sampling Methods:

2.1 Random Sampling with a Reservoir
2.2 Concise Sampling

2. Sampling Methods

数据挖掘实验室
Data Mining Lab

Advantages:

- Easy and Efficient
- Any data mining application or database operation & Provable error guarantees
- Multi-dimensional

2. Sampling Methods

数据挖掘实验室
Data Mining Lab

Some properties:

- Markov inequality: $P(X > a) \le E[X]/a = \mu/a$
- Chebychev inequality(the random variable $(X-\mu)^2/\sigma^2$)

$$P(|X-u| > a) \le \sigma^2/a^2$$

X independent & identical Bernoulli random variables

$$P(X < (1-\delta)\mu) \le e^{-\mu\sigma^2/2}$$

$$P(X > (1+\delta)\mu) \le \max\{2^{-\delta\mu}, e^{-\mu\delta^2/4}\}$$

- Hoeffding inequality:

a set of k independent random variables the range [a, b]

$$P(|X-\mu| > \delta) \le 2e^{-2k\cdot\delta^2/(b-a)^2}$$

2.1 Random Sampling with a Reservoir

数据挖掘实验室
Data Mining Lab

One-pass & *N* are not known & Dynamically

Algorithm:

A reservoir of size k.

- Initialization :Chosen $1\sim k$ points in the data stream.
  The first k points in the data streams are added to the reservoir .
- Subsequent : Form $(k+1)th$ to N
  a) Each point has a k/ i probability of being selected.
  (i is the order of points)
  a) Current points in the reservoir are sampled with equal probability ($1/ k$) and subsequently removed.

2.1 Random Sampling with a Reservoir

数据挖掘实验室
Data Mining Lab

eg: If $k = 1000$, form 1001 to N ;

The probability of the 1001*th* point being included in the reservoir is 1000/1001.

Assumption:

a) The $(i + 1)th$ point is added to the reservoir with probability $k/(i + 1)$.

b) First i points have equal probability of being included in the reservoir and have probability equal to $k/(i+1)$.

This sampling approach feasible.

2.

Rese

Proc

The

1. If $i = k+1$,the probability of the $(k+1)th$ point being included in the reservoir is $k/k+1$.And first k points are included in the reservoir and have probability equal to $k/(k+1)$.
2. If j=i,the hypothesis is right, the probability of the i th point being included in the reservoir is k / i.And first $i-1$ points are included in the reservoir and have probability equal to k / i.
3. If j=i+1? we need to proof :The $(i + 1)th$ point is added to the reservoir with probability $k/(i + 1)$.First i points have equal probability of being included in the reservoir and have probability equal to $k/(i+1)$.

数据挖掘实验室 Data Mining Lab

Proof (Induction):

The probability of first *i* points being included in the reservoir have two components.

① Appearing in the reservoir before the *(i+1)th* selection.

② To ensure the first i choice is not replaced.

- Known from the 2. Before the *(i+1)th* selection,first *i* points are included in the reservoir and have probability equal to *k / i*.
- The probability of replacing:

the *(i+1)th* point is selected and the probability is *k/i+1*,the point in the reservior is choosed as the probability 1/k,so the probability of the first i points being replaced is *k/(i+1) *1/k=1/i+1*

the point in the reservoir in not replaced: 1 - 1/(i+1)=i / i+1

so the first i points included in the reservoir is k/i *i/(i+1) = k/i+1

2.2 Concise Sampling

数据挖掘实验室 Data Mining Lab

- **Increasing the sample size**
- **The available main memory restrictions**
- **The fact**:the number of distinct values of an attribute is often significantly smaller than the size of the data stream.

Definition 1:

A concise sample is a uniform random sample of the data set such that values appearing more than once in the sample are represented as a value and a count.

Most applicable while performing univariate sampling along a single dimension.

2.2 Concise Sampling

Maintained as a set S of <*value, count*> pairs.

- count = 1.(do not maintain the count explicitly)
- the value as a singleton.

Definition 2:

Let $S = \{ <v_1, c_1>, \ldots , <v_j, c_j>, v_{j+1}, \ldots , v_l \}$ *be a concise sample.*

Then $\textit{sample-size}(S) = e - j + \sum_{i=1}^{j} c_i$, *and* $\textit{footprint}(S) = e + j$.

Some conclusions:

- The footprint size <= the true sample size.
- If the count of any distinct element is larger than 2, then the footprint size is strictly smaller than the sample size.

2.2 Concise Sampling

数据挖掘实验室
Data Mining Lab

The algorithm :

For extracting a concise sample of footprint m .

- First, repeat m times: select a random tuple from the relation and extract its value .
- Next, replace every value occurring multiple times with a *<value, count>* pair.
- Then, continue to sample until either adding the sample point would increase the concise sample footprint to m + 1 or n samples have been taken.

2.2 Concise Sampling

数据挖掘实验室
Data Mining Lab

- if the corresponding value count-pair is already included in the set S, then only increment the count by 1. the footprint size does not increase.
- if the value of the current point is distinct from all the values encountered so far, or it exists as a singleton then the foot print increases by 1. Either a singleton needs to be added, or a singleton gets converted to a value-count pair with a count of 2.

2.2 Incremental of Concise Sampling

数据挖掘实验室
Data Mining Lab

The increase in footprint size may potentially require the removal of an element from sample S in order to make room for the new insertion.

- Seting up an entry threshold τ (initially 1) for new tuples to be selected for the sample with the probability $1/\tau$
- Picking a new (higher) value of the threshold τ'
- Reducing the count of a value with probability τ/τ' ,until at least one value-count pair reverts to a singleton or a singleton is removed.

Subsequent points from the stream are sampled with probability $1/\tau'$

Thanks !

Tan Yue
Data Ming Lab
tanxiangyueer@foxmail.com