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

Structure-Aware Sampling on Data Streams

Yue Tan

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

2016/8/24 Wednesday

tanxiangyueer@foxmail.com

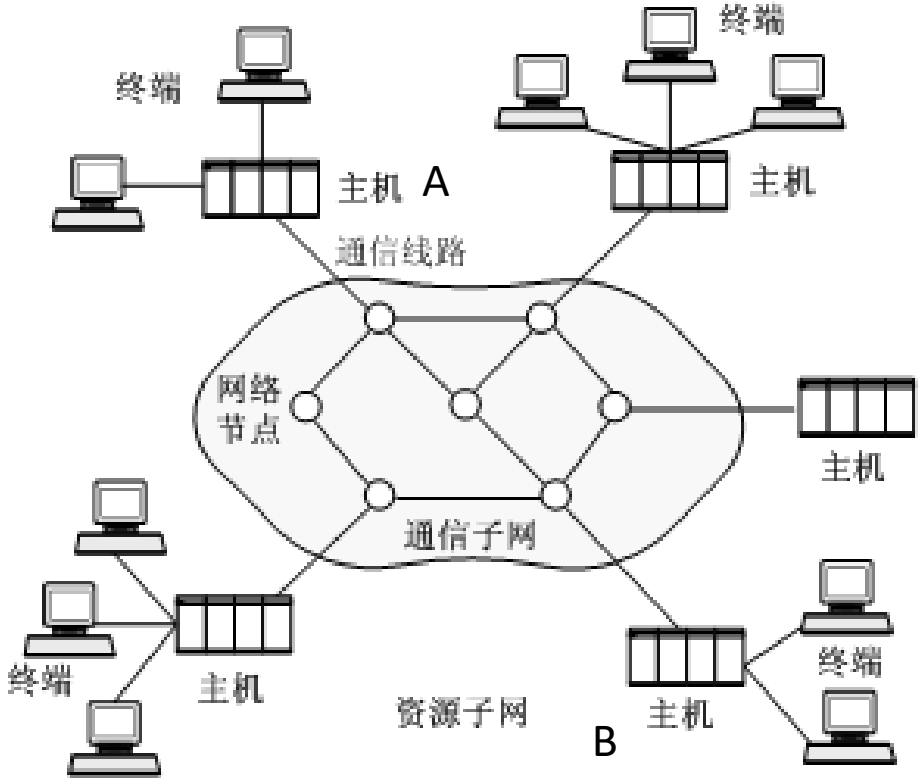
1.两句话说这篇文章讲啦什么！

两句话：

- 之前的采样方法没有考虑数据本事的结构特性就进行采样，这样做损失很多有价值信息，并且采样之后可能会有冗余数据。
- 针对这一问题作者提出了一种考虑数据局部结构特性的采样算法。

Cohen E, Cormode G, Duffield N. ***Structure-aware sampling on data streams*** *[C] // Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems. ACM, 2011: 197-208.*

2.一个现实中的小栗子



日志：{ 源地址、目标地址、中转设备地址、时间、大小 }。
作用：分析网络是否健康、探测异常的数据或者配置错误、网络是否有攻击

3. 一个为了说明问题的大栗子

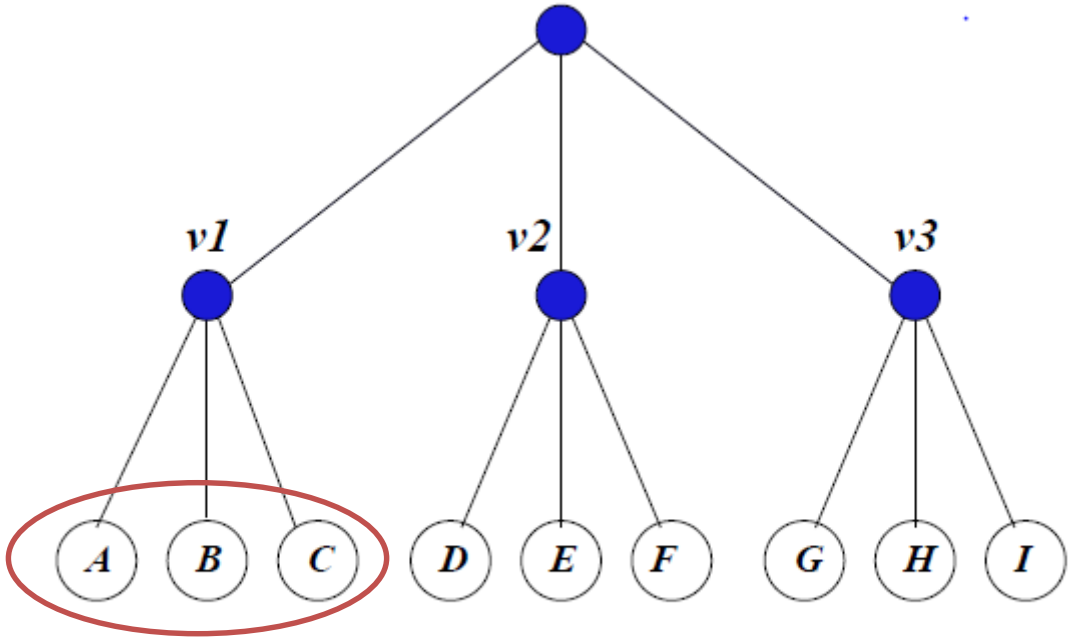


Figure 1: Hierarchy structure example.

4. 不成熟的小计算

- 每个点自带初始权重w，一个自定义的伸缩参数τ，更新之后的权重为a

	A	D	E	G	B	C	H	I	F	τ
3	1	1	1							1

采样数：$\sum_i \min\{1, w_i / \tau_k\} = k$　　τ_k

采样集 *A = {1 , 2, 3}*

- 判断数据G是否要加入采样数据集：

	A	D	E	G	B	C	H	I	F	τ
3	1	1	1	1						1
4	$\frac{4}{3}$	0	$\frac{4}{3}$	$\frac{4}{3}$						$\frac{4}{3}$

采样集 *A = {1，2，3，4}*

$M \leftarrow \frac{\sum_{i \in A} a_i}{|A| - 1}$　　$M = \tau$　　$p_i = w_i / M < 1$　　$a_i = \max\{w_i, \tau\}$

$a_i = \tau$

4.不成熟的小计算

- 判断B是否要加入采样数据集

	A	D	E	G	B	C	H	I	F	τ
3	1	1	1							1
4	$\frac{4}{3}$	0	$\frac{4}{3}$	$\frac{4}{3}$	1					$\frac{4}{3}$
5	0	0	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$					$\frac{5}{3}$

采样集：*A = {1，3，4，5}*　　*M =（4/3 + 4/3 + 4/3 + 1）/（4-1）= 5/3*

$\tau = \frac{5}{3}$　　$p_i = 1$　　$a_i = \frac{5}{3}$

- 。 。 。 。 。 。 。 。 。

- 之后重复迭代的计算，最后的结果为：

4.不成熟的小计算

	A	D	E	G	B	C	H	I	F	τ
3	1	1	1							1
4	$\frac{4}{3}$	0	$\frac{4}{3}$	$\frac{4}{3}$						$\frac{4}{3}$
5	0	0	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$					$\frac{5}{3}$
6	0	0	2	2	2	0				2
7	0	0	$\frac{7}{3}$	$\frac{7}{3}$	0	0	$\frac{7}{3}$			$\frac{7}{3}$
8	0	0	$\frac{8}{3}$	$\frac{8}{3}$	0	0	$\frac{8}{3}$	0		$\frac{8}{3}$
9	0	0	3	3	0	0	3	0	0	3

5. 更成熟的小计算

- 一个伸缩函数(tightness parameter)*c >= 1*。c越大，灵活性越大。
- 定义 *k/c* 是选择范围

C = 3/2, k = 3, k/c = 2,

- 采样集合*S = { A, D, E }*

	X	A	D	E	G	B	C	H	I	F	τ
3		1	1	1							1

- Range Cost

$$\rho\left(\{u,v\}\right)=\frac{w_u w_v}{a(s)}\left(\frac{w_u+w_v}{3}+a\left(S_M\right)\right)$$

a(S) : the total adjusted weight.

$S_M=\{i:\mathrm{u}<i<v\}$ *is the subset of all keys in S that lie strictly between u and v.*

Range Cost is minimized when the set X is a pair of keys

5. 更成熟的小计算

- Hierarchy

$$\rho\left(\{u,v\}\right)=\frac{w_u w_v}{a(s)}\left(\frac{w_u+w_v+a(S_M)}{3}+\frac{a(S_C)}{2}\right)$$

a(S) : the total adjusted weight.

$S_M=\{i:\mathrm{u}<i<v\}$ *is the subset of all keys in S that lie strictly between u and v.*

Sc : the remaining leaf nodes in the same subtree.

- Adjusted weight of candidate set

$$M(\mathrm{a},\mathrm{X})=\frac{\sum_{i\in CAND(\mathrm{X})}a_i}{\left|CAND(\mathrm{x})\right|-1}$$

$$M=\tau=a_i$$

5. 更成熟的小计算

- *C = 3/2, k = 3, k/c = 2,*

	X	A	D	E	G	B	C	H	I	F	τ
3		1	1	1							1
4	$\{D, E\}$	1	~~1~~ 0	2	1						$\frac{4}{3}$

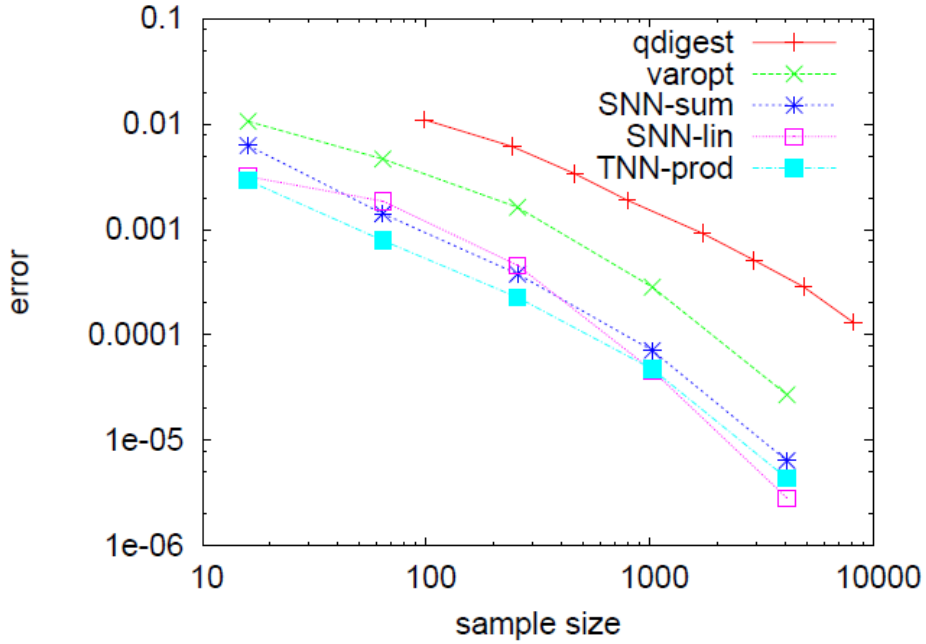
- 通过计算发现{D，E}之间的range cost最小
- M(a,X) = (1+1)/(2-1) = 2　　$a_i = 2$
- $\rho(\{D,E\}) = \rho(\{E,F\})$ 选择点 *E* 作为采样点

	X	A	D	E	G	B	C	H	I	F	τ
3		1	1	1							1
4	$\{D, E\}$	1	0	2	1						$\frac{4}{3}$
5	$\{A, B\}$	0	0	2	1	2					$\frac{5}{3}$

5. 更成熟的小计算

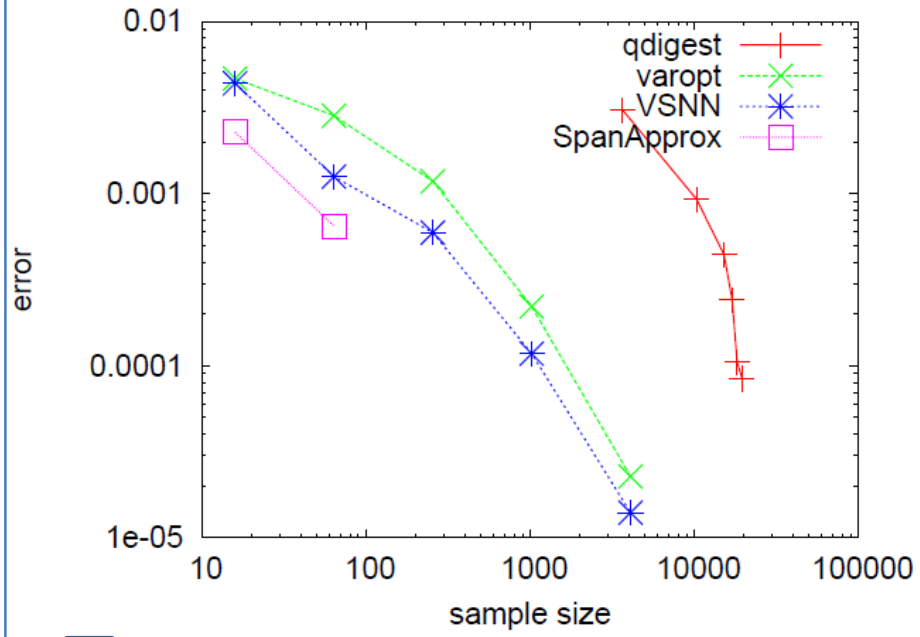
	X	A	D	E	G	B	C	H	I	F	τ
3		1	1	1							1
4	$\{D, E\}$	1	0	2	1						$\frac{4}{3}$
5	$\{A, B\}$	0	0	2	1	2					$\frac{5}{3}$
6	$\{B, C\}$	0	0	2	1	3	0				2
7	$\{G, H\}$	0	0	2	0	3	0	2			$\frac{7}{3}$
8	$\{H, I\}$	0	0	2	0	3	0	3	0		$\frac{8}{3}$
9	$\{E, F\}$	0	0	3	0	3	0	3	0	0	3

6. 实验小结果：



Overall MRE as a function of sample size k.

Accuracy: 1-dimensions



Overall MRE as a function of sample size k.

Accuracy: 2-dimensions

7. 一个小总结：

- 优点：
  结构灵活；低计算开销；。。。。。

- 缺点：
  初始权重的问题；。。。。

Thanks !

Tan Yue

Data Ming Lab

tanxiangyueer@foxmail.com