DETECTING CHANGE IN DATA STREAM: USING SAMPLING TECHNIQUE

By Wei Li , Xiaoming Jin and Xiaojun Ye
Presented by Mark Everline

Outline

- Overview
- A-Distance
- DCDDS Algorithm
- Experimental Results
- Schedule

Introduction

- Probability distribution as the key character of a data stream in detecting change
- Data stream changed PD has changed.
- Detecting Change in the Distribution (most common)
  - Willcoxom test
  - Lp distance
  - Jensen-Shannon Divergence (information distance)
- Using A-Distance.

A-Distance

- Definition 1 Change: $S<s_1,s_2,s_3,\ldots,s_t>$, *tc* (current time) at anytime *t*, $t < tc$ there are $S1<s_1,s_2,s_3,\ldots,s_t>$ and $S2<s_{t+1},s_{t+2},s_{t+3},\ldots,s_{tc}>$, if $f(S1, S2) > \varepsilon$ there is change a time *t*.
  - f is distance function
  - ε is threshold
  - R1 and R2 are the subset of the complete data stream.
- A-Distance[1] defined

$$f_A(P_1,P_2)=2\sup_{a\in A}\frac{|P_1(a)-P_2(a)|}{\{\min\{\frac{P_1(a)-P_2(a)}{2},1-\frac{P_1(a)-P_2(a)}{2}\}\}^{\frac{1}{2}}}$$

- Replace $P_i(a)$ with $S_i(a)=|S_i \wedge a| / |S_i|$

DCDDS Algorithm

```
Find_Change
    For I = 1 … k do
        C0 = 0
        S1,i = first m point from time C0
        S2,i = next m point in stream
    End for
    While not at end of stream do
        For I = 1…K do
            Sampling the new data into S1,i
            if ( f(S1,i, S2,i) > εi then
                C0 = current time
                Report change at time C0
                Clear all windows and GOTO 1
            end if
        End of
    End while
```

- f - Distance function
- m sample size.
- Set of Triples
  $\{(p_1,\varepsilon_1),(p_2,\varepsilon_2)\ldots(p_k,\varepsilon_k)\}$

- Meta Algorithm is running K independent algorithms
- Compare Random X with Sample probability p. (sample algorithm)

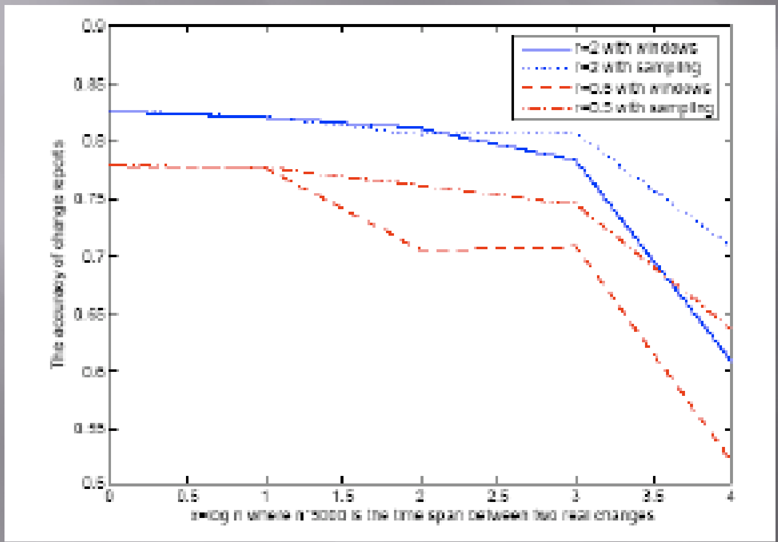
When Sample is full discard oldest point in sample size.

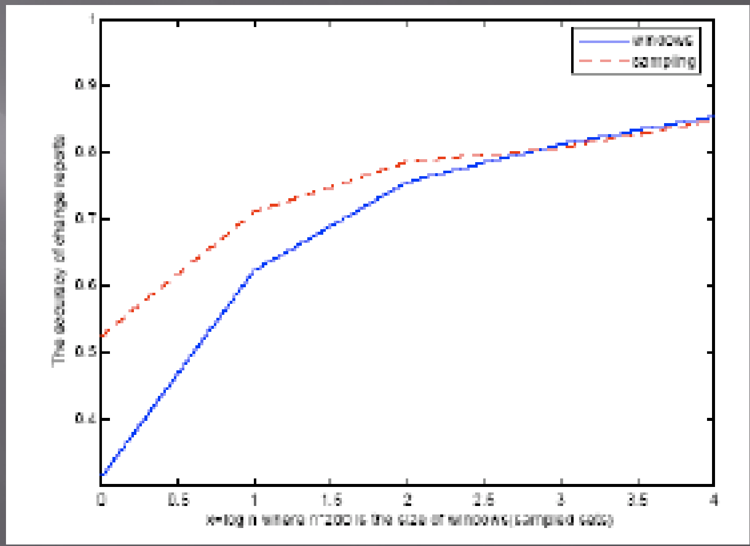
DCDDS Advantages

- Provide tighter statistical guarantees
- Less missing detections and false alarms
- Works better with sliding window model on detecting small changes
- Better time cost then sliding window
- Time cost is 1/p same as sliding window.

Experimental Results

- Experiment 2 Mill points, uniform distribution, time span=20,000, window size 200-300 drift r=2 and p=5

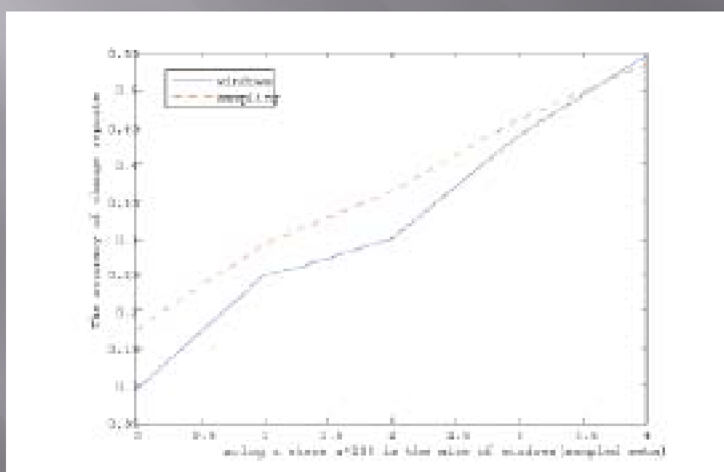


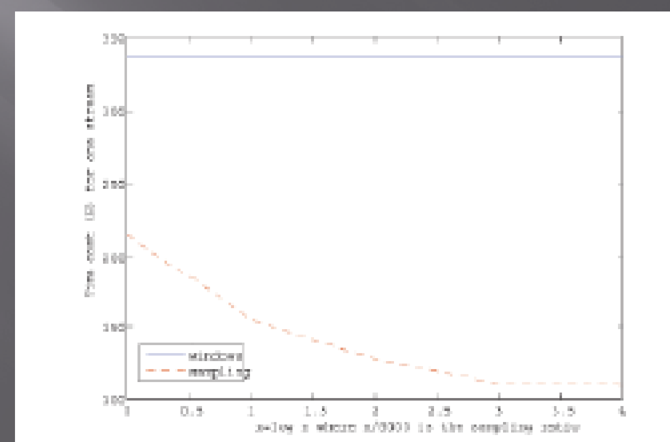


Experimental Results

- The Normal distribution with *μ=50, σ=5, with the* change drift *r=0.5*

- The time cost statistics using the uniform distribution with *p = 5 and r = 2.The time span is 20,000, and size of* windows(sampled-sets) is 1600.





Questions/Conclusion

- W. L. X. J. X. Ye, "Detecting Change in Data Stream: Using Sampling Technique," *Natural Computation, 2007. ICNC 2007. Third International Conference on* vol. 1, pp. 130-134, Aug 2007, 2007.

- [1] S. B.-D. Daniel Kifer, Johannes Gehrke "Detecting change in data streams," *Proceedings of the Thirtieth international conference on Very large data bases,* vol. 30, no. 13, pp. 180-191, 2004

Schedule

- Continue Lit Search
- Implementing Multi Variant KDE: #Crime, Lat Lon
- Scrubbing the Data:
  - Adding: Town of Herndon, Fairfax City Vienna