Detecting Changes in Data Streams

Shai Ben-David, Johannes Gehrke, Daniel Kifer
Cornell
VLDB 2004

Laurent Charlin

November 30, 2005

Outline

- Problem Definition and setting
- Related Work
- The paper with background
  - Distance Measures & statistical guarantees
  - Constructing a statistical test
  - Algorithm for Kolmogorov-Smirnov
  - Experiments
- Critique & Discussion

Overall

- Data Streams: They want to detect changes in a stream
  - Reliable change detection (bounding the error)
  - Comprehensible description of the nature of the change

Overall

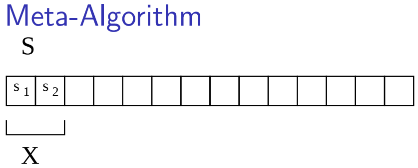
- Data Streams: They want to detect changes in a stream
  - Reliable change detection (bounding the error)
  - Comprehensible description of the nature of the change
- Statistics: Test whether two samples are generated by different distributions

Overall

- Data Streams: They want to detect changes in a stream
  - Reliable change detection (bounding the error)
  - Comprehensible description of the nature of the change
- Statistics: Test whether two samples are generated by different distributions
  - Tuples in a stream are independent
  - Stream is larger than available memory
  - No modeling assumption (non-parametric stats)

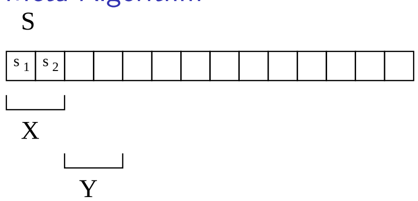
Basic setting and example

Meta-Algorithm



Basic setting and example

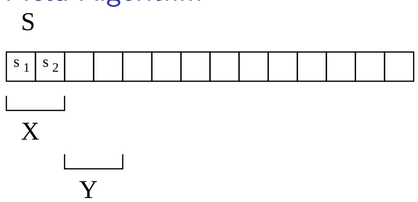
Meta-Algorithm



Problem Definition

Basic setting and example

Meta-Algorithm

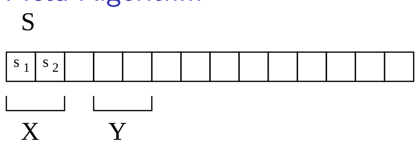


K(X,Y) > threshold

Problem Definition

Basic setting and example

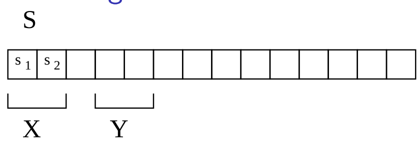
Meta-Algorithm



K(X,Y) > threshold

Basic setting and example

Meta-Algorithm

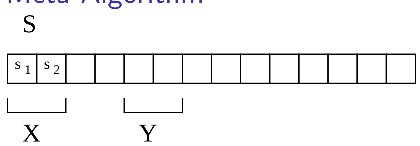


K(X,Y) > threshold

NO.

Basic setting and example

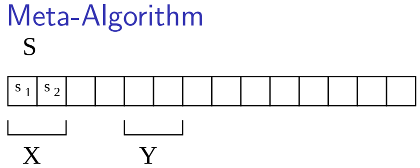
Meta-Algorithm



K(X,Y) > threshold

Basic setting and example

Meta-Algorithm

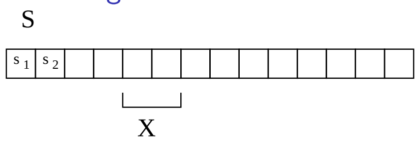


K(X,Y) > threshold

YES.

Basic setting and example

Meta-Algorithm



K(X,Y) > threshold

YES.

Their solution

Proposed Solution

1. Some (new) distance measure
2. A bound on errors w.r.t. the sample sizes
3. A full statistic test

Related Work

- The problem of detecting changes in streams has not be studied at such a level
- In-between database and statistics
  - Statistics usually assume you have access to all data and some model
  - Database no work that general and formal

Definition

$\mathcal{A}$-distance

- P, P' two probability distribution over a fixed space
- $\mathcal{A}$ a collection of measurable sets ($A \in \mathcal{A}$)

$$d_{\mathcal{A}}(P, P') = 2 \sup_{A \in \mathcal{A}} |P(A) - P'(A)|$$

Distance measure

Definition

$\mathcal{A}$-distance

- P, P' two probability distribution over a fixed space
- $\mathcal{A}$ a collection of measurable sets ($A \in \mathcal{A}$)

$$d_{\mathcal{A}}(P, P') = 2 \sup_{A \in \mathcal{A}} |P(A) - P'(A)|$$

Empirically

- S a finite domain subset

$$S(A) = \frac{|S \cap A|}{|S|}$$
$$d_{\mathcal{A}}(S_1, S_2) = 2 \sup_{A \in \mathcal{A}} |S_1(A) - S_2(A)|$$

Distance measure

Definition

$\mathcal{A}$-distance

- P, P' two probability distribution over a fixed space
- $\mathcal{A}$ a collection of measurable sets ($A \in \mathcal{A}$)

$$d_{\mathcal{A}}(P, P') = 2 \sup_{A \in \mathcal{A}} |P(A) - P'(A)|$$

Empirically

- S a finite domain subset

$$S(A) = \frac{|S \cap A|}{|S|}$$
$$d_{\mathcal{A}}(S_1, S_2) = 2 \sup_{A \in \mathcal{A}} |S_1(A) - S_2(A)|$$

- Interpretability : the user can choose the "shape" of A

Example

- $\mathcal{A}$ the collection of all intervals
- A is a subset of that collection
- S is a sample

Kolmogorov-Smirnov Statistic

- The $\mathcal{A}$-distance is a generalization of this statistic if we set $\mathcal{A}$ to be initial segments $(-\infty, x)$

$$\sup_x |F_1(x) - F_2(x)|$$
$$\text{where } F_i(x) = P_i(\{y : y \leq x\})$$

Relativized Discrepancy

- Normalized $\mathcal{A}$-distance

$\Pi_{\mathcal{A}}$ function

- Now that we have defined a proper distance measure how good does it behave ?
- Let's Introduce tools to help us out.

$\Pi_{\mathcal{A}}$ function

- Now that we have defined a proper distance measure how good does it behave ?
- Let's Introduce tools to help us out.

$$\Pi_{\mathcal{A}}(n) = \max\{|\{A \cap B : A \in \mathcal{A}| : B \subseteq X \text{ and } |B| = n\}$$

- Maximum number of subsets of B that can be intersected by A.
- $\Pi_{\mathcal{A}} \leq 2^n$

VC-Dimension

VC-Dim

- Work by Vapnik and Chervonenkis in the '70s
- Describes the complexity of a collection of sets.
- Root of Statistical Learning Theory

$$VC - Dim(\mathcal{A}) = \sup\{n : \Pi_{\mathcal{A}}(n) = 2^n\}$$

VC-Dimension

VC-Dim

- Work by Vapnik and Chervonenkis in the '70s
- Describes the complexity of a collection of sets.
- Root of Statistical Learning Theory

$$VC - Dim(\mathcal{A}) = \sup\{n : \Pi_{\mathcal{A}}(n) = 2^n\}$$

Sauer's Lemma

- n is the number of samples and d the $VC - Dim$

$$\Pi_{\mathcal{A}} \leq n^d$$

Example

The Bounds

- You want strong guarantees on the number of false alarms and missed detections
- You want to bound the probability of making a mistake by a function of your sample size

$$P[\exists A \in \mathcal{A} || P_1(A) - P_2(A)| - |S_1(A) - S_2(A)|| \geq \epsilon] < \Pi_{\mathcal{A}}(2n)4e^{-n\epsilon^2/4}$$

- This translates directly into the $\mathcal{A}$-distance and similar results are shown for the relativized discrepancies.

The components of a test

1. The null hypothesis ("The two samples are generated by the same distribution")

The components of a test

1. The null hypothesis ("The two samples are generated by the same distribution")
2. The statistics (Kolmogorov-Smirnov, $\mathcal{A}$-distance)

The components of a test

1. The null hypothesis ("The two samples are generated by the same distribution")
2. The statistics (Kolmogorov-Smirnov, $\mathcal{A}$-distance)
3. The critical region
   - For which values of the statistic do we reject the null hypothesis

Finding the critical region

- Find the appropriate α

$$K(< s_1, \dots, s_{m_1} >, < s_{i+m_1}, \dots, s_{i+m_1+m_2} >) > \alpha$$
$$\text{for } i = 1, 2, \dots, n - m_1 - m_2$$

- Let $F_{K,m_1,m_2,n(S)}$ be the maximum of these values
- This is a random variable (and it's shown to be independent of the generating function)
- So you can approximate its distribution and take the value of a (small) quantile to be α

The idea

- Can we find an algorithm to run these statistics efficiently

The idea

- Can we find an algorithm to run these statistics efficiently

The setting

- They restrict the domain of $\mathcal{A}$ to intervals or initial segments.
- They only show the full algorithm for the Kolmogorov-Smirnov statistic

KS structure

Definition

- A is a KS structure if
  - It's a finite array of elements $< a_1, \ldots, a_m >$ in R^2
    1. Value - $v(a_i)$
    2. Weight - $w(a_i)$
  - The array is sorted in increasing order of values.
  - $|A|$ is the length of the array.

KS structure

Definition

- A is a KS structure if
  - It's a finite array of elements $< a_1, \ldots, a_m >$ in $\mathbb{R}^2$
    1. Value - $v(a_i)$
    2. Weight - $w(a_i)$
  - The array is sorted in increasing order of values.
  - $|A|$ is the length of the array.

$G_A(k)$ function

- $G_A(k) = \sum_{i=1}^{k} w(a_i)$

Intuition for the Algorithm

Creating a *KS* structure Z

- Z is a *KS* structure which aggregated of the two windows X and Y

Algorithms

Intuition for the Algorithm

Creating a *KS* structure Z

- Z is a *KS* structure which aggregated of the two windows X and Y
  - $w(z_i) = -1/|X|$ if z_i comes from X.

Intuition for the Algorithm

Creating a *KS* structure Z

- Z is a *KS* structure which aggregated of the two windows X and Y
  - $w(z_i) = -1/|X|$ if z_i comes from X.
  - $w(z_i) = 1/|Y|$ if z_i comes from Y.

Algorithms

Intuition for the Algorithm

Creating a *KS* structure Z

- Z is a *KS* structure which aggregated of the two windows X and Y
  - $w(z_i) = -1/|X|$ if z_i comes from X.
  - $w(z_i) = 1/|Y|$ if z_i comes from Y.

Reduction of a statistic to the max of a difference

- They show that calculating the Kolmogorov-Smirnov statistic can be reduce to an aggregation of the $G_A(k)$ functions.

Laurent Charlin () Detecting Changes in Data Streams November 30, 2005 18 / 23

Algorithms

Intuition for the Algorithm

Creating a *KS* structure Z

- Z is a *KS* structure which aggregated of the two windows X and Y
  - $w(z_i) = -1/|X|$ if z_i comes from X.
  - $w(z_i) = 1/|Y|$ if z_i comes from Y.

Reduction of a statistic to the max of a difference

- They show that calculating the Kolmogorov-Smirnov statistic can be reduce to an aggregation of the $G_A(k)$ functions.
  - Intervals $(a, b) : \max_{a<b} |G_Z(b) - G_Z(a)| = \max_c G_Z(c) - min_d G_z(d)$

Laurent Charlin () Detecting Changes in Data Streams November 30, 2005 18 / 23

Intuition for the Algorithm

Creating a *KS* structure Z

- Z is a *KS* structure which aggregated of the two windows X and Y
  - $w(z_i) = -1/|X|$ if z_i comes from X.
  - $w(z_i) = 1/|Y|$ if z_i comes from Y.

Reduction of a statistic to the max of a difference

- They show that calculating the Kolmogorov-Smirnov statistic can be reduce to an aggregation of the $G_A(k)$ functions.
  - Intervals (a, b) : $\max_{a<b} |G_Z(b) - G_Z(a)| = \max_c G_Z(c) - min_d G_z(d)$
  - Initial segments $(-\infty, a)$: $\max |G_Z(a)|$

Example

Two Experiments

1. 2 million points without a change in the generating function (Figure 1)
   - Detect false alarms
2. Distribution drifts every 20K points (still 2 million points in total). (Figures 2-8)
   - Able to detect late or no detections and false alarms.

Results

- Different statistics still have different properties
- No one is always better than the other one
- Relativized Discrepancies seem to be the less prone to errors

The paper

- Very elegant way of dealing with a streams problems
  - Formal analysis using combinatorial properties
- Efficient algorithm (which might be run online with prior offline calculations).

Discussion

- Is Independency of tuples reasonable ?
  - They say they want to relax this but it seems that some of the theory they are using won't allow it.
- How easy is it to understand the result of the statistics ?
  - You can set the shape of $\mathcal{A}$.
  - But that shape is contrained to initial segments and intervals right now.
  - Still seems that some intuition is present.
- How can we push these ideas into higher dimension (right now we were algorithmically restricted to using interval or initial segments) ?
- In real-life which statistic do we use ? Could we combine them ?
- Since the samples are pretty large (ie: > 100) which not use parametric (normal distribution) statistics ?
  - The central limit theorem basically says that the mean of these samples should behave like a normal distribution.