Detecting Changes in Data Streams

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

Presented by
Shen-Shyang Ho
Content:

1. Summary of the paper (abstract)
2. Problem Setting
3. Statistical Problem
4. Hypothesis Test: Wilcoxon and Kolmogorov-Smirnov
5. Meta-algorithm
6. Metrics over the space of distribution
7. Statistical Bounds for the changes
8. Critical Region
9. Characteristics of algorithm
10. Experiment
Summary of Paper (abstract)

1. Method for detection and estimation of change
2. Provide proven guarantees on the statistical significance of detected changes
3. Meaningful description and quantification of those changes
4. Nonparametric i.e. no prior assumption on the nature of the distribution that generate the data but must be i.i.d.
5. Method works for both continuous and discrete data
1. Assume that the data is generated by some underlying probability distribution, one point at a time, in an independent fashion.

2. When this data generating distribution changes, detect it.

3. Quantify and describe this change (comprehensible description of the nature of the change).
1. What is data stream and static data?
   - Static data: generated by a fixed process e.g. sampled from a fixed distribution.
   - Data stream: temporal dimension and underlying process generating the data stream can change over time

2. Impacts of changes: Data that arrived before a change can bias the model towards characteristics that no longer hold
Solution: Change-Detection Algorithm

1. Two-window paradigm.

2. Compare data in some “reference window” to the data in current window.

3. Both windows contain a fixed number of successive data points.

4. Current window slides forward with each incoming data point, and the reference window is updated whenever a change is detected.
**Statistical Problem:**

1. Detecting changes over a data stream is reduced to the problem of testing whether two samples were generated by different distribution.
2. Detecting a difference in distribution between two input samples.
3. Design a “test” that can tell whether two distributions $P_1$ and $P_2$ are different.
4. A solution that guarantees that when a change occurs it is detected and limits the amount of false alarm.
5. Extend the guarantees from two-sample problem to the data stream.
6. Non-parametric test that comes with formal guarantees.
7. Also describe change in a user-understandable way.
Change-Detection Test

We want the test to have the 4 properties:

1. Control false positives (spurious detection)
2. Control false negatives (missed detection)
3. Non-parametric
4. Description of the change.

What about classical nonparametric test?

1. Wilcoxon Test
2. Kolmogorov-Smirnov Test
1. Null and Alternative Hypothesis
   - $H_0$: The sample populations have identical distribution.
   - $H_1$: The distribution of population 1 is shifted to the right of population 2. (two-tailed test: either left or right)

2. Test Statistics

3. A Critical Region
Wilcoxon Test - (1)

1. Signed Rank Test: To test whether the median of a symmetric population is 0. (Rank without sign; Reattach sign; Compute One sample z statistic, $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$)

2. Rank Sum Test: To test whether two samples are drawn from the same distribution.

Algorithm:

1. rank the combined data set

2. divided the ranks into two sets according to the group membership of the original observations.

3. calculate a two-sample z statistics, $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$.
Wilcoxon Test - (2)

1. For large samples ($> 25 - 30$), the statistic is compared to percentiles of the standard normal distribution.

2. For small samples, the statistic is compared to what would result if the data were combined into a single data set and assigned at random to two groups having the same number of observations as the original samples.
Kolmogorov-Smirnov (KS-)Test

1. The KS-test is used to determine if two datasets differ significantly
2. Continuous random variables.
3. Given $N$ data points $y_1, y_2, \cdots, y_N$ the Empirical Cumulative Distribution Function (ECDF) is defined as
   
   \[ E_j(i) = \frac{n(i)}{N}, j = 1, 2 \]
   
   where $n(i)$ is the number of points less than $y_i$. This is a step function that increases by $1/N$ at the value of each data point.
4. Compare the two ECDF. That is,
   
   \[ D = \max |E_1(i) - E_2(i)| \]
5. The null hypothesis is rejected if the test statistic, $D$, is greater than the critical value obtained from a table.
Meta-Algorithm: Find_Change

1. for $i = 1 \cdots k$ do
   (a) $c_0 \leftarrow 0$
   (b) $\text{Windows}_{1,i} \leftarrow$ first $m_{1,i}$ points from time $c_0$
   (c) $\text{Windows}_{2,i} \leftarrow$ next $m_{2,i}$ points in stream
2. end for
3. while not at end of stream do
   (a) for $i = 1 \cdots k$ do
      i. Slide $\text{Windows}_{2,i}$ by 1 point
      ii. if $d(\text{Windows}_{1,i}, \text{Windows}_{2,i}) > \alpha_i$ then
         A. $c_0 \leftarrow$ current time
         B. Report change at time $c_0$
         C. Clear all windows and GOTO step 1
      iii. end if
   (b) end for
4. end while
Metrics over the space of distribution:  
Distance measure: \( L_1 \) norm (or total variation, TV)

The \( L_1 \) norm between any 2 distributions defined as

\[
||P_1 - P_2||_1 = \sum_{a \in \chi} |P_1(a) - P_2(a)|
\]

Let \( A \) be the set on which \( P_1(x) > P_2(x) \). Then

\[
||P_1 - P_2||_1 = \sum_{a \in \chi} |P_1(a) - P_2(a)|
\]

\[
= \sum_{x \in A} |P_1(x) - P_2(x)| + \sum_{x \notin A^c} |P_2(x) - P_1(x)|
\]

\[
= P_1(A) - P_2(A) + P_2(A^c) - P_1(A^c)
\]

\[
= P_1(A) - P_2(A) + 1 - P_2(A) - 1 + P_1(A)
\]

\[
= 2(P_1(A) - P_2(A))
\]

\[
TV(P_1, P_2) = 2 \sup_{E \in \mathcal{E}} |P_1(E) - P_2(E)|
\]

where \( P_1 \) and \( P_2 \) are over the measure space \( (X, \mathcal{E}) \).
1. $L_1$ distance (or total variation) between 2 distributions is too sensitive and can require arbitrarily large samples to determine whether 2 distributions have $L_1$ distance $> \epsilon$.

2. $L_P$ norm ($p > 1$) are too insensitive.
$A$ – distance - (1)

FIX a measure space and let $A$ be a collection of measurable sets ($A \subset \mathcal{E}$). Let $P$ and $P'$ be probability distributions over this space.

- The $A$ – distance between $P$ and $P'$ is defined as

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

$P$ and $P'$ are $\epsilon$- close with respect to $A$ if $d_A(P, P') \leq \epsilon$

- For a finite domain subset $S$ and a set $A \in A$ let the empirical weight of $A$ w.r.t. $S$ be

$$S(A) = \frac{|S \cap A|}{|S|}$$

- For finite domain subsets, $S_1$ and $S_2$, we define the empirical distance to be

$$d_A(S_1, S_2) = 2 \sup_{A \in A} |S_1(A) - S_2(A)|$$
$\mathcal{A}$ – distance - (2)

1. Relaxation of the total variation distance
2. $d_{\mathcal{A}}(P, P') \leq TV(P, P')$ (less restrictive)
3. help get around the statistical difficulties associated with the $L_1$ norm.
4. If $\mathcal{A}$ is not too complex (VC-dimension!!), then there exists a test that can distinguished with high probability if two distributions are $\epsilon$-close with respect to $\mathcal{A}$ using a sample size that is independent of the domain size.
1. Special Case: Kolmogorov-Smirnov Test: \( \mathcal{A} \) is the set of one-sided intervals \((-\infty, x), \forall x \in \mathbb{R}\).

2. if \( \mathcal{A} \) is the set of all intervals \([a, b], \forall a, b \in \mathbb{R}\), (or the family of convex sets for high dimensional data), then \( \mathcal{A} \)-distance reflects the relevance of locally centered changes.
Relativized Discrepancy

1. Variation of A-distance that takes the relative magnitude of a change into account.

2. Use to provide statistical guarantees that the differences that these measures evaluate are detectable from bounded size samples.
Statistical bound: change-detection estimator

Given a domain set, $X$ and $\mathcal{A}$ be a family of subsets of $X$.

1. n-th shatter coefficient of $\mathcal{A}$:

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

- Maximum number of different subsets of $n$ points that can be picked out by $\mathcal{A}$
- Measure the richness of the $\mathcal{A}$
- $\Pi_\mathcal{A} \leq 2^n$

2. VC-dimension (Complexity of $\mathcal{A}$):

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

3. Sauer’s Lemma: $\Pi_\mathcal{A}(n) \leq \sum_{i=0}^{d} \binom{n}{i} < n^d$

4. Vapnik-Chervonenkis Inequality: Let $P$ be a distribution over $X$ and $S$ be a collection of $n$ i.i.d. sampled from $P$. Then for $\mathcal{A}$, a family of subsets of $X$ and a constant $\epsilon \in (0, 1)$

$$P^n(\sup_{A \in \mathcal{A}} |S(A) - P(A)| > \epsilon) < 4\Pi_\mathcal{A}(2n)e^{-n\epsilon^2/8}$$
Statistical bound: change-detection estimator

Let $P_1, P_2$ be any probability distributions over some domain $X$ and let $A$ be a family of $X$ and $\epsilon \in (0, 1)$. If $S_1, S_2$ are i.i.d. $m$ samples drawn by $P_1, P_2$ respectively, then,

$$P(\exists A \in A || P_1(A) - P_2(A) - |S_1(A) - S_2(A)| \geq \epsilon) \leq 8\Pi_A(2m)e^{-m\epsilon^2/32}$$

[Proof:]

$$P(\exists A \in A || P_1(A) - P_2(A) - |S_1(A) - S_2(A)| \geq \epsilon) \leq P(\sup_{A \in A} |P_1(A) - P_2(A) - S_1(A) + S_2(A)| \geq \epsilon)$$

$$\leq P(\sup_{A \in A} (|P_1(A) - S_1(A)| + |P_2(A) - S_2(A)|) \geq \epsilon)$$

$$\leq P((\sup_{A \in A} |P_1(A) - S_1(A)| \geq \frac{\epsilon}{2}) \cup (\sup_{A \in A} |P_2(A) - S_2(A)| \geq \frac{\epsilon}{2}))$$

$$\leq P(\sup_{A \in A} |P_1(A) - S_1(A)| \geq \frac{\epsilon}{2}) + P(\sup_{A \in A} |P_2(A) - S_2(A)| \geq \frac{\epsilon}{2})$$

$$\leq 8\Pi_A(2m)e^{-m\epsilon^2/32}$$

It follows that

$$P(|d_A(P_1, P_2) - d_A(S_1, S_2)| \geq \epsilon) \leq 8\Pi_A(2m)e^{-m\epsilon^2/32}$$
Let $\mathcal{A}$ be a collection of subsets of a finite VC-dimension $d$. Let $S_1, S_2$ are samples of size $n$ each, drawn i.i.d. by the same distribution $P$ (over $X$), then,

\[ P^n(\phi_\mathcal{A}(S_1, P) > \epsilon) \leq 8\Pi_\mathcal{A}(2n)e^{-n\epsilon^2/4} \]

\[ P^{2n}(\phi_\mathcal{A}(S_1, S_2) > \epsilon) \leq 2\Pi_\mathcal{A}(2n)e^{-n\epsilon^2/4} \]

[Proof:] Use a result from Anthony and Shawe-Taylor:

\[ P(\sup_{A \in \mathcal{A}} \frac{S_1(A) - S_2(A)}{S_1(A) + S_2(A)}) \leq \Pi_\mathcal{A}(2n)e^{-n\epsilon^2/4} \]
Let $P_1$ and $P_2$ be probability distributions over $X$ and $S_1, S_2$ finite samples of sizes $m_1, m_2$ drawn i.i.d. according to $P_1, P_2$ respectively. Then

$$P^{m_1+m_2}(|\phi_A(S_1, S_2) - \phi_A(P_1, P_2)| > \epsilon) \leq (2m_1)^d e^{-m_1\epsilon^2/16} + (2m_2)^d e^{-m_2\epsilon^2/16}$$

In Addition, if $m_1 = m_2 = n,$

$$P^{2n}(|\phi_A(S_1, S_2) - \phi_A(P_1, P_2)| > \epsilon) \leq 16\Pi_A(2n)e^{-n\epsilon^2/16}$$
1. A statistical test over data streams is a size($n,p$) test if, on data that satisfies the null hypothesis, the probability of rejecting the null hypothesis after observing $n$ points is at most $p$.

2. Construct a critical region $\{x : x \geq \alpha\}$ such that if the null hypothesis were true, then the expected number of times (per $n$ points) that the test statistic falls in the critical region is small.

3. Reject the null hypothesis for large values of the test statistics.

4. In order to construct the critical regions, we must study the distributions of the test statistics under the null hypothesis (all $n$ points have the same generating distribution).
Computing the $\alpha$ for given $n$ and $p$

1. Theorem 4.1: The distribution of $F$ (given the test statistics, $n$ and the two set sizes: the maximum of all the tests values of a particular test statistics over all possible window locations) does not depend on the generating distribution $G$ of the $n$ points. (Hence, one-time cost to compute $\alpha$ for a particular $p$)

2. 3 ways: 1) Direct Computation 2) Simulation 3) Sampling

3. Theorem 4.3: Assures that the critical region constructed and the probability of falsely rejected the null hypothesis is $\le p$ even if $G$ is discrete.

4. They use simulation (500 runs) to find $\alpha$
Brief characteristics of algorithm

1. A balanced tree to maintain the samples from two windows, $O(\log(m_1 + m_2))$ for all the four tests.

2. Re-compute the Kolmogorov-Smirnov statistics over initial segments and intervals in $O(\log(m_1 + m_2))$ time.

3. A balanced tree and divide-and-conquer algorithm for the incremental algorithm in (2).
**Experiment**

- A stream of 2,000,000 points and change distribution every 20,000 points (99 true changes)
- run the change-detection algorithm on 5 control streams of 2 million points each and no distribution change.
- 2 critical regions: $S(50k, .05)$ and $S(20k, .05)$.
- 4 test statistics
- 4 window sizes: 200, 400, 800, 1600.
- Distribution change: data stream with distribution F with parameters $p_1, \cdots, p_n$ and rate of drift $r$. When it is time to change, choose a uniform r.v. $R_i$ in $[-r, r]$ and add it to $p_i$ for all $i$.
- $a/b$: $a$ is the number of change reports considered to be not late; $b$ represents the number of changes reports that are late or wrong.
# Experimental Result

<table>
<thead>
<tr>
<th>size(n,p)</th>
<th>W</th>
<th>KS</th>
<th>KS(int)</th>
<th>φ</th>
<th>Ξ</th>
</tr>
</thead>
<tbody>
<tr>
<td>S(20k, .05)</td>
<td>8</td>
<td>8</td>
<td>9.8</td>
<td>3.6</td>
<td>7.2</td>
</tr>
<tr>
<td>S(50k, .05)</td>
<td>1.4</td>
<td>0.6</td>
<td>1.8</td>
<td>1.6</td>
<td>1.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>S(20k,.05)</th>
<th>S(50k,.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>0/5</td>
<td>0/4</td>
</tr>
<tr>
<td>KS</td>
<td>31/30</td>
<td>25/15</td>
</tr>
<tr>
<td>KSI</td>
<td>60/34</td>
<td>52/27</td>
</tr>
<tr>
<td>φ</td>
<td>92/20</td>
<td>86/13</td>
</tr>
<tr>
<td>Ξ</td>
<td>86/19</td>
<td>85/9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>S(20k,.05)</th>
<th>S(50k,.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>0/2</td>
<td>0/0</td>
</tr>
<tr>
<td>KS</td>
<td>0/15</td>
<td>0/7</td>
</tr>
<tr>
<td>KSI</td>
<td>4/32</td>
<td>2/9</td>
</tr>
<tr>
<td>φ</td>
<td>16/33</td>
<td>12/27</td>
</tr>
<tr>
<td>Ξ</td>
<td>13/36</td>
<td>12/18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C</th>
<th>S(20k,.05)</th>
<th>S(50k,.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>10/27</td>
<td>6/16</td>
</tr>
<tr>
<td>KS</td>
<td>17/30</td>
<td>9/27</td>
</tr>
<tr>
<td>KSI</td>
<td>16/47</td>
<td>10/26</td>
</tr>
<tr>
<td>φ</td>
<td>16/38</td>
<td>11/31</td>
</tr>
<tr>
<td>Ξ</td>
<td>17/43</td>
<td>16/22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D</th>
<th>S(20k,.05)</th>
<th>S(50k,.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>12/38</td>
<td>6/34</td>
</tr>
<tr>
<td>KS</td>
<td>11/38</td>
<td>9/26</td>
</tr>
<tr>
<td>KSI</td>
<td>7/22</td>
<td>4/14</td>
</tr>
<tr>
<td>φ</td>
<td>7/29</td>
<td>5/18</td>
</tr>
<tr>
<td>Ξ</td>
<td>11/46</td>
<td>4/20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>S(20k,.05)</th>
<th>S(50k,.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>36/42</td>
<td>25/30</td>
</tr>
<tr>
<td>KS</td>
<td>24/38</td>
<td>20/26</td>
</tr>
<tr>
<td>KSI</td>
<td>17/22</td>
<td>13/15</td>
</tr>
<tr>
<td>φ</td>
<td>12/32</td>
<td>11/18</td>
</tr>
<tr>
<td>Ξ</td>
<td>23/33</td>
<td>15/23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F</th>
<th>S(20k,.05)</th>
<th>S(50k,.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>36/35</td>
<td>31/26</td>
</tr>
<tr>
<td>KS</td>
<td>23/30</td>
<td>16/27</td>
</tr>
<tr>
<td>KSI</td>
<td>14/25</td>
<td>10/18</td>
</tr>
<tr>
<td>φ</td>
<td>14/21</td>
<td>9/17</td>
</tr>
<tr>
<td>Ξ</td>
<td>23/22</td>
<td>17/11</td>
</tr>
</tbody>
</table>

1. A. Uniform on $[-p, p](p = 5)$ with drift = 1.
2. B. Mixture of standard normal and uniform $[-7, 7](p = 0.9)$ with drift = 0.05.
3. C. Normal $(\mu = 50, \sigma = 5)$ with drift = 0.6.
4. D. Exponential $(\lambda = 1)$ with drift = 0.1.
5. E. Binomial $(p = 0.1, n = 2000)$ with drift = 0.001.
6. F. Poisson $(\lambda = 50)$ with drift = 1.
Conclusion

For high dimensional data, let $\mathcal{A}$ be the family of convex sets for high dimensional data:

- Given $X$ be $R^2$ and $E$ be the set of all convex sets of the plane, VC-dimension $= \infty$ !!

- if $E$ is the set of convex sets with $d$ sides, VC-dimension $= 2d + 1$ but not measurable!!

...
