

Mining Massive Data Streams

Michael Hahsler

Computer Science and Engineering
Southern Methodist University

January 23, 2012



SMU | BOBBY B. LYLE
SCHOOL OF ENGINEERING

Table of Contents

- 1 Introduction
- 2 Properties of Data Stream
- 3 Time Windows, Sampling and Sketches
- 4 Clustering
- 5 Classification
- 6 Conclusion

Data Streams

Definition

A data stream is an ordered and potentially infinite sequence of data points:

$$\langle \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \dots \rangle$$

where \mathbf{y}_i is a tuple (e.g., a vector)

Such streams of constantly arriving data are generated by many types of applications including:

- web click-stream data
- computer network monitoring data
- telecommunication connection data
- readings from sensor nets
- stock quotes

Example: HTTP Server Log

```
208.76.226.148 - - [15/Jan/2012:04:02:42 -0600]
  "GET /MMSA/destroysession.php HTTP/1.0" 302 -
208.76.226.148 - - [15/Jan/2012:04:02:42 -0600]
  "GET /MMSA/index.php HTTP/1.0" 200 11339
129.119.113.115 - - [15/Jan/2012:04:03:43 -0600]
  "GET / HTTP/1.1" 200 1227
208.76.226.148 - - [15/Jan/2012:04:03:48 -0600]
  "GET /PIIH/2011/hurricanes/AL122011/11090118AL1211_PIIH.txt
  HTTP/1.0" 304 -
```

Data stream mining algorithms

- Clustering
- Classification
- Frequent Pattern Mining
- Change Detection
- Database Operations: indexing streams for trend and aggregation queries
- Mining multiple streams

Table of Contents

- 1 Introduction
- 2 Properties of Data Stream**
- 3 Time Windows, Sampling and Sketches
- 4 Clustering
- 5 Classification
- 6 Conclusion

Properties of data streams

- Unbounded size of stream
 - ▶ Transient (stream might not be realized on disk)
 - ▶ Single pass over the data
 - ▶ Only summaries can be stored
 - ▶ Real-time processing (in main memory)
- Data streams are not static
 - ▶ Incremental updates
 - ▶ Concept drift
 - ▶ Forgetting old data
- Temporal order may be important

Why can we not use the standard algorithms?

- Why can we not use a regular relational DB and SQL?
- Why not a k -nearest neighbors classifiers?
- Why not k -means/hierarchical clustering?
- Why not Apriori to find frequent itemsets?

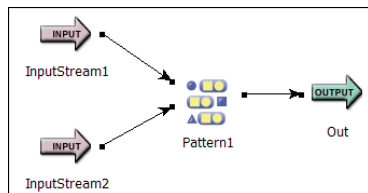
Relational DB vs. Data Streams

| Relational DBMS | DSMS (Stream) |
|---------------------------------|----------------------|
| persist ant relations | transient streams |
| only current state is important | history matters |
| not real-time | real-time |
| low update rate | stream! |
| one time queries | continuous queries |

Source: Babcock et al, Models and Issues in Data Stream Systems, POTS, 2002

DSMS typically offer SQL-like languages with stream extensions to create continuous queries.

Example: Pattern matching in StreamSQL

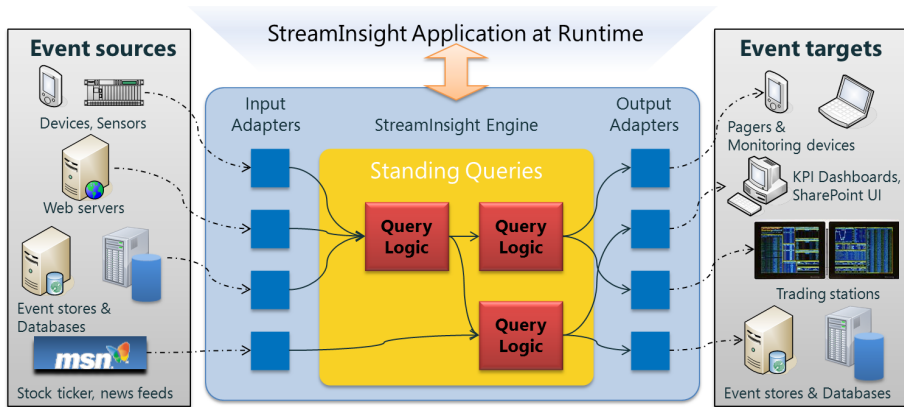


```
CREATE INPUT STREAM InputStream1 (stock string, value double);  
CREATE INPUT STREAM InputStream2 (stock string, value double);  
CREATE OUTPUT STREAM Out;
```

```
SELECT InputStream1.stock AS stock,  
       InputStream1.value AS value1,  
       InputStream2.value AS value2  
FROM PATTERN (InputStream1 THEN InputStream2) WITHIN 20 TIME  
WHERE (InputStream2.value > InputStream1.value)  
      AND (InputStream1.stock = InputStream2.stock)  
INTO Out;
```

Source: StreamBase, <http://www.streambase.com/>

Example: Microsoft StreamInsight



Source: Introducing Microsoft StreamInsight, 2009

Traditional algorithms vs. DS algorithms

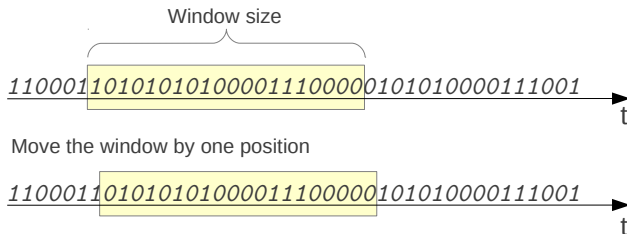
| | Traditional | Stream |
|------------------------|--------------------|---------------|
| passes | multiple | single |
| processing time | unlimited | restricted |
| memory | disk | main memory |
| results | typically accurate | approximate |
| distributed | typically not | often |

Source: Joao Gama, Data Stream Mining Tutorial, ECML/PKDD, 2007

Table of Contents

- 1 Introduction
- 2 Properties of Data Stream
- 3 Time Windows, Sampling and Sketches**
- 4 Clustering
- 5 Classification
- 6 Conclusion

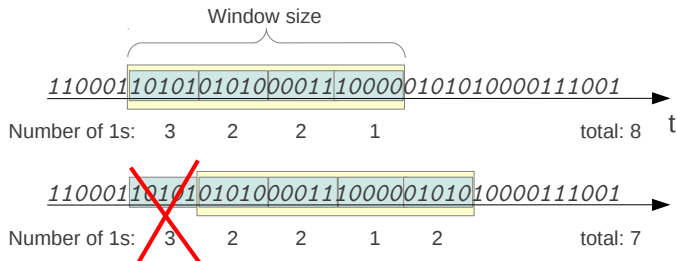
Time window



- Keep the most recent data points.
- Reconstruct a regular model from the window when it changes.
- Typically updated as a sliding window. Sometimes landmark or titled windows.

This is typically expensive!

How many 1s are within the window?



- Use buckets
- Models need to be additive (works for count, mean, variance, etc.)
- Can also be used to detect change

Sampling

- Reduce the amount of data to process and store.
- Updating an unbiased sample is tricky since new data is arriving constantly!

What is the problem with the following approach to create a sample of size k :

- 1 Insert first k elements into sample
- 2 Add each new element to the sample with a fixed probability p .
- 3 If a new element was inserted then delete the oldest element in the sample.

Reservoir Sampling

Vitter, J., Random Sampling with a Reservoir, ACM, 1985

Create a sample of size k :

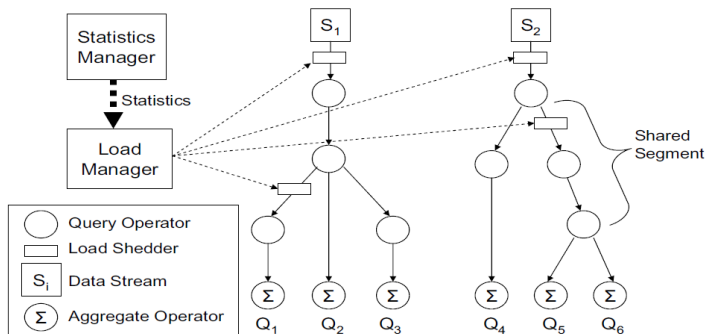
- 1 Insert first k elements into sample
- 2 Then insert i th element with probability $p_i = k/i$.
- 3 If a new element was inserted then delete an instance at random.

Load shedding

Many data streams have bursts. Similar situation is faced by utility companies with bursts in demand for electricity.

As a result some of the data cannot be processed and needs to be dropped.

Solution for sliding window aggregate queries:



Source: Babcock, Datar and Motwani (Load Shedding Techniques for Data Stream Systems, Data Streams - Models and Algorithms, 2007)

Sketches

A sketch is a small data structure which can be easily updated and helps with estimating frequency moments of a data stream (typically with an error guarantee).

Sketches exist to approximate:

- Count unique values in a stream
- Identify heavy hitters (most frequent items)
- Finding quantiles
- Finding the difference between streams

Sketches: Count distinct values

- Maintain a Hash Sketch *BITMAP* which is an array of L bits, where $L = O(\log(M))$, initialized to 0.
- Assume a hash function $h(x)$ that maps incoming values $x \in [0, M - 1]$, uniformly across $[0, 2^L - 1]$.
- Let $lsb(y)$ denote the position of the least-significant 1 bit in the binary representation of y .
- A value x is mapped to $lsb(h(x))$. For each incoming value x , set $BITMAP[lsb(h(x))] = 1$.

Example

$x = 5 \rightarrow h(x) = 101100 \rightarrow lsb(h(x)) = 2$

BITMAP: 0 0 0 0 0 0 0 0 0 0 1 0 0

Source: Flajolet and Martin, Probabilistic Counting Algorithms for DataBase Applications, JCSS, 1983. Adapted from Joao Gama, Data Stream Mining Tutorial, ECML/PKDD, 2007

Sketches: Count distinct values

Example

BITMAP: 0 0 0 0 1 0 1 1 0 1 1 1 1 1

Left most 0-bit is at position $R = 6$.

Flajolet and Martin proved that $E[R] = \log(\phi M)$ with $\phi = .77351$
Estimate of $M = 2^R / \phi$.

Example

$M = 2^6 / \phi = 82.7$ distinct values.

Wavelets

Idea: Concentrate on the important features of the data.

- Wavelet transforms (like Discrete Cosine and Fourier transforms) split the data up into components (e.g., basic trend and local variations)
- Retain only the most important components.
- For data stream summarization fast to compute Wavelets are used (e.g., Haar Wavelet)

Interactive Example:

<http://www.tomgibara.com/computer-vision/haar-wavelet>

Table of Contents

- 1 Introduction
- 2 Properties of Data Stream
- 3 Time Windows, Sampling and Sketches
- 4 Clustering**
- 5 Classification
- 6 Conclusion

Clustering Data Streams

Conventional clustering algorithms need several passes over the complete data set!

Main ideas:

- Strategies
 - ① **Time Window:** Split stream into time windows and cluster each window independently. Then combine the clusterings (STREAM).
 - ② **Micro-clusters:** A small set of statistics which can be iteratively updated (mean, variance, etc.). (CluStream, DenStream)
 - ③ **Density based:** Map each data point into a predefined grid. (D-Stream, MR-Stream)
- **Reclustering:** Use conventional clustering (e.g., k -means, DBSCAN) off-line to combine micro-clusters/grids.
- **Exponential decay** to decrease the influence of older data on the micro-clusters. This deal with concept drift.

A very simple algorithm

- 1 Start with an empty set of micro-clusters
- 2 For each new data point x
 - 1 Find for x the closest micro-cluster c
 - 2 If x is closer to c than a set threshold δ then
 - 1 add update x to absorb cotherwise
 - 1 create a new micro-cluster for c .

Some research questions

- Temporal structure?
- No off-line reclustering?
- How do we compare different algorithms?

Table of Contents

- 1 Introduction
- 2 Properties of Data Stream
- 3 Time Windows, Sampling and Sketches
- 4 Clustering
- 5 Classification**
- 6 Conclusion

Decision Trees

Problem: How do we decide on splits if new data is constantly arriving?

- **Solution 1:** Use a time window.
- **Solution 2 (Very Fast Decision Trees):** Uses the current best attribute to make a split once the number of examples satisfies the *Hoeffding* bound. This gives a guarantee on how different the tree will be from a tree built on all the data. (Domingos and Hulten, Mining High Speed Data Streams, KDD, 2000)

Problems with decision trees: Need to be rebuilt to adapt to concept drift.

Classification by Clustering

Idea: Cluster the data stream into groups and assign a label to each cluster. Find for a new data point the closest cluster and use its label. (Aggarwal et al., On Demand Classification of Data Streams, KDD, 2004)

Advantages:

- DS clustering is fast
- Takes care of concept drift
- Micro-clusters allow for an arbitrary decision boundary. Essentially is k -nearest neighbor with $k = 1$ and micro-clusters instead of data points)

Possible problems:

- What if micro-clusters contain points of several classes?
- How do we label a new developing micro-cluster?

Table of Contents

- 1 Introduction
- 2 Properties of Data Stream
- 3 Time Windows, Sampling and Sketches
- 4 Clustering
- 5 Classification
- 6 Conclusion**

Conclusion

- Data streams are everywhere
- Often a good approximation is all that is needed
- Some streaming algorithms produce results of similar quality as traditional algorithm at a fraction of the computational cost → apply them to large non-streaming data
- Data stream extensions for DBs are/will become available (e.g., MS SQL Server's StreamInsight)
- Distributed stream mining (cluster, grid, cloud) will become important