Managing and mining (streaming) sensor data

Petr Čížek

Artificial Intelligence Center
Czech Technical University in Prague

November 3, 2016

Stream data mining / stream data querying

Problem definition

- Data can not be stored
- Data arrive in stream or streams
- Random access to data impossible or very expensive $\rightarrow$ single scan algorithms

Challenges

- Queries are continuous
- Pre-defined vs. ad-hoc queries
- Answer update over time $\rightarrow$ anytime property

Practical example 1 - sensors

Sensor	MB/s	GB/h
Inertial Measurement Unit	0.1	0.3
Monocular camera (640x480@60fps MJPEG)	$\sim$1.73	$\sim$6.1
Monocular camera (640x480@60fps RAW)	17.5	63.2
Stereo camera (2x640x480@60fps RAW)	35	126.4
Velodyne 3D laser scanner	$\sim$100	$\sim$351

Practical example 2 - institutions

Institution	GB/s
CERN	
RAW data (sensors)	$\sim$600000
RAW data processed	$\sim$25
ALICE	4
ATLAS	1
CMS	0.6
LHCb	0.8
Network peer nodes	
NIX.cz	$\sim$37
AMS-IX	$\sim$500

Comparison to traditional data mining

	Traditional	Stream
No. of passes	Multiple	Single
Processing time	Unlimited	Restricted
Memory usage	Unlimited	Restricted
Type of result	Accurate	Approximate
Concept	Static	Evolving
Distributed	No	Yes

Stream data mining / data querying

Applications

Statistics, Classification, Clustering, Outlier (error) detection

Challenges

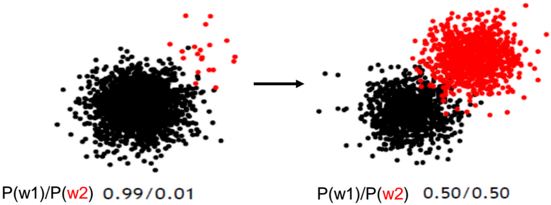
- Pre-defined vs. ad-hoc queries
- Concept-drift
- Concept-evolution
- Feature-evolution

Methods

- Random sampling
- Sketching
- Histograms
- Sliding windows (Fading windows)
- Multi resolution model (subsampling)
- Feature selection

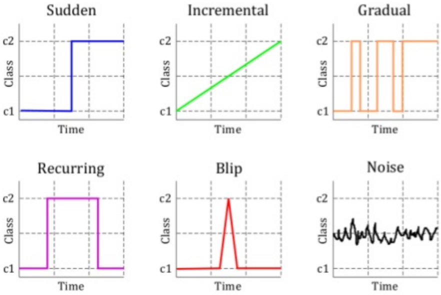
Challenges - concept-drift

Statistical properties of the target variable, which the model is trying to predict, evolve over time in unforeseen ways.



Challenges - concept-drift

Statistical properties of the target variable, which the model is trying to predict, evolve over time in unforeseen ways.



7

Image: D. Brzeziński thesis

Challenges - concept-drift

Active solutions

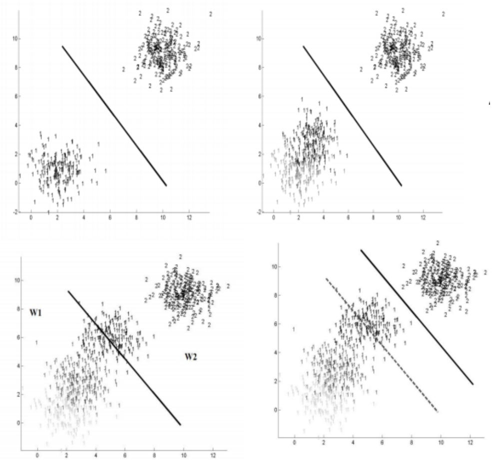
- Activated by triggers
- Can be used by any classification algorithm
- Need to completely relearn the model when triggered
- E.g. *n* latest decisions are monitored

Passive solutions

- Adaptive - continuously updating the model
- Don't detect changes

Challenges - concept-evolution

Misclassification of novel class in data



Challenges - feature-evolution

The features are evolving throughout the time

Methods for stream data processing - random sampling

Subsample the data in randomized way.

- Save only $1/n$ samples randomly
- Law of large numbers assure probability completeness

Methods for stream data processing - sketching

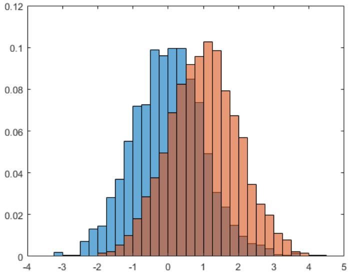
Extract frequency moments of the stream
The kth frequency moment of a set of frequencies $\mathbf{a}$ is
$F_k(\mathbf{a}) = \sum_{i=1}^{n} a_i^k$

- F_1 - total count of different frequencies
- F_2 - statistical properties - e.g. dispersion
- F_∞ - frequency of the most frequent items

Methods for stream data processing - histograms

Types of histograms

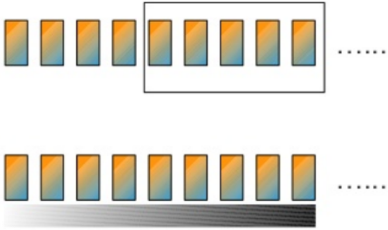
- V-optimal
  $v_1, v_2 \cdots v_n$ classes
  $\sum_i (v_i - \hat{v}_i)^2$
- Equal-width
- End-biased



Methods for stream data processing - sliding window

Forgetting mechanism

- Sliding window
- Fading factor



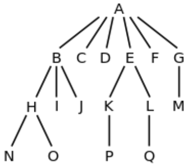
Methods for stream data processing - multi resolution model

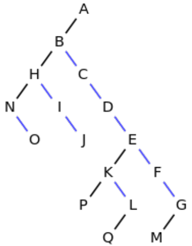
Using decision trees

- Use part of stream for choosing the root attribute
- Following examples pass to leaves

- Advantage - scalable
- Disadvantage - only for stationary distribution

→ Using context-drift aware decision trees





Methods for stream data processing - feature selection

Features are to characterize a particular sample $\rightarrow$ dimension reduction

- Artificial - (e.g. Image features)
- Learned - (e.g. using neural networks, reinforcement learning)

Mining data streams - research issues

- Mining sequential patterns
- Mining partial periodicity
- Mining notable gradients
- Mining outliers and unusual patterns
- Clustering

Thank you for your attention!

References

Pier Luca Lanzi, ***Course "Machine Learning and Data Mining"***, Politecnico di Milano
http://www.slideshare.net/pierluca.lanzi/18-data-streams

Anand Rajaraman and Jeffrey D. Ullman, ***Mining of Massive Datasets***, Cambridge University Press, 2011

Charu C. Aggarwal, ***Managing and Mining Sensor Data***, Springer Science Business Media, 2013.

Manuel Martín, ***Master's thesis: Handling concept drift in data stream mining***, University of Granada
http://www.slideshare.net/draxus/handling-concept-drift-in-data-stream-mining