Building Adaptive Data Mining Models on Streaming Data in Real-Time, an Outlook on Challenges, Approaches and Ongoing Research

Marine Perception Research Department of the German Research Center for Artificial Intelligence (DFKI)

Frederic Stahl
Data never sleeps!

- Forbes: 2.5 quintillion bytes of data created every day.
- That’s about 100 million Blue-ray discs or about 530 million DVD discs.

Sources:
https://www.theregister.co.uk/2008/01/23/us_hd_player_sales/
https://www.domo.com/data-never-sleeps
How much Data Is created Every day?

- That’s about 100 million Blue-ray each 25 GB discs.
- Each disc is 1.2mm thick
  ⇒ This stacks to **120 km**!
  ⇒ Distance Oldenburg to Hamburg!

- Or in DVDs (4.7 GB each disc)
- Each disc is 1.2mm thick
  ⇒ This stacks to **630 km**!
  ⇒ Distance Oldenburg to London/Reading!
To make sense of this real-time data, analytics methods that never sleep are required!
Internet of Things

- By year-end 2039, IoT devices worldwide are forecasted to almost triple from 9.7 billion in 2020 to 29 billion in 2030 [1]

[1] Statista. (2020). Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030
Sources of Data Streams

Internet of Things
- By year-end 2039, IoT devices worldwide are forecasted to almost triple from 9.7 billion in 2020 to 29 billion in 2030 [1]

Personalisation
- Facebook:
  - 1.91 billion active users every day [2]
  - 4.75 billion pieces of content shared

---

[1] statistica. (2020). Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030
Sources of Data Streams

Internet of Things
- By year-end 2039, IoT devices worldwide are forecasted to almost triple from 9.7 billion in 2020 to 29 billion in 2030 [1]

Personalisation
- Facebook:
  - 1.91 billion active users every day [2]
  - 4.75 billion pieces of content shared

Marine Sciences
- Distribution of ocean science data acquired in the past decade, based on publicly available data from the internet (CC BY 4.0) [3]
- Expected to reach almost 500 Exabytes by the year 2025

[1] statistica. (2020). Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030
A data stream is a continuous, rapid flow of data that challenges our state-of-the-art processing and communication infrastructure.

**Static Data**
- Historical data
- Randomly accessible
- Secondary storage
- No/low processing latency criticality
- Assumption of pre-processed dataset

**Streaming Data**
- Often live, real-time data feed
- Sequentially accessed
- Limited memory requirements
- High processing latency criticality
- Assumption of inaccurate raw data
Concept Drift

- Underlying concept defining the knowledge being learned, begins to shift over time.
- Concept change is unforeseen and unpredictable.
- Concepts from the past may re-occur in the future.
- Concept drift exists in real-life problems:
  - Seasonal weather
  - Stock market rallies because of breaking news
  - etc.

Sudden Drift
Gradual Drift
Recurring Drift
Incremental Drift

Time
Concept Drift (cont.)

Concept shift/drift: changes mining set statistics

- A model should always reflect the time-changing concept.
- Render previously learned models inaccurate or invalid.
- Robustness and adaptability: quickly recover/adjust after concept changes.
Concept Drift (cont.)

Concept shift/drift: changes mining set statistics

• A model should always reflect the time-changing concept.
• Render previously learned models inaccurate or invalid.
• Robustness and adaptability: quickly recover/adjust after concept changes.
Concept Drift (cont.)

Concept shift/drift: changes mining set statistics
- A model should always reflect the time-changing concept.
- Render previously learned models inaccurate or invalid.
- Robustness and adaptability: quickly recover/adjust after concept changes.
What are Data Streams?

Challenges & Barriers

Building Data Mining Models from Data Streams

Research Activities in Data Stream Mining

Outlook and Conclusions
## The Data Tsunami

### Challenges

1) Data generated at a fast rate (Velocity), at potentially large and unknown quantities (Volume)
2) Concept Drift (changes of pattern encoded in the data over time)
3) Modelling real-time analytics workflows from streaming data
4) Multi-modality of data sources (text, video/images, unstructured)
5) Class label sparsity: adapting predictive models
6) Explaining Concept Drift

### Barriers

1) Limited scalable (parallel) real-time high throughput data stream mining algorithms
2) Different and changing types of concept drift
3) Lack of customisable pre-processing techniques
4) Different time stamps but co-occurring data items
5) Supervised algorithms not applicable in many cases
6) Lack of drift detectors explaining concept drift
What are Data Streams?

Challenges & Barriers

Building Data Mining Models from Data Streams

Research Activities in Data Stream Mining

Outlook and Conclusions
Methods: Windowing approaches to induce data mining models

1) Create time windows

Methods: Windowing approaches to induce data mining models

1) Create time windows

- Landmark window from \( t_1 \)
- Damped window
- Sliding window

2) Detect concept drift

Methods: Windowing approaches to induce data mining models

1) Create time windows

- Landmark window from $t_1$
- Damped window
- Sliding window


2) Detect concept drift

3) Adapt data mining model

What are Data Streams?

Challenges & Barriers

Building Data Mining Models from Data Streams

Research Activities in Data Stream Mining

Outlook and Conclusions
Objective: Develop a scalable predictive Data Stream classification
Objective: Develop a scalable predictive Data Stream classification

1) Initialising Micro-Clusters and maintenance statistics

- Initially a fixed number of Micro-Clusters is randomly initialised.
- Only components outlined in the table are stored.
- These can be used to calculate the clusters centroid and boundary (variance).
Objective: Develop a scalable predictive Data Stream classification

1) Initialising Micro-Clusters and maintenance statistics

- Initially a fixed number of Micro-Clusters is randomly initialised.
- Only components outlined in the table are stored.
- These can be used to calculate the clusters centroid and boundary (variance).

\[
<CF^x, CF^1, CF^1, n, CL, \epsilon, \Theta, \bar{v}, \Omega>
\]

\[
\text{centroid}(x) = \frac{CF^1}{n}
\]

\[
\text{Variance}[x] = \sqrt{\left(\frac{CF^2}{n}\right) - \left(\frac{CF^1}{n}\right)^2}
\]
Objective: Develop a scalable predictive Data Stream classification

1) Initialising Micro-Clusters and maintenance statistics

- Initially a fixed number of Micro-Clusters is randomly initialised.
- Only components outlined in the table are stored.
- These can be used to calculate the clusters centroid and boundary (variance).

\[
<CF_{x}, CF_{1x}, CF_{1l}, n, CL, \epsilon, \Theta, \hat{u}, \Omega>
\]

\[
\text{centroid}(x) = \frac{CF_{1x}}{n}
\]

\[
\text{Variance}[x] = \sqrt{\left(\frac{CF_{2x}}{n}\right) - \left(\frac{CF_{1x}}{n}\right)^2}
\]

2) Absorbing new data stream instances

3) Splitting and removing of Micro-Clusters

MC-NN: Results

Adapting to Concept Drift

Incremental Naive Bayes (NB)

MC-NN (MC)

Hoeffding Tree (HOFF)
MC-NN: Results

Adapting to Concept Drift

Scalability through parallelisation

Hoeffding Tree (HOFF)

Incremental Naive Bayes (NB)

MC-NN (MC)
Adapting to Concept Drift

Hoeffding Tree (HOFF)

Incremental Naive Bayes (NB)

MC-NN (MC)

Scalability through parallelisation

Long term adaptation

Using MC-NN for Explaining Concept Drift through Feature Tracking

Measuring centroid velocity

Measuring split & death rate
Using MC-NN for Explaining Concept Drift through Feature Tracking

Measuring centroid velocity

Measuring split & death rate

Feature tracking and ranking

Results

MC-NN: Unsupervised Classification

The presented work falls in this intersection.
MC-NN: Unsupervised Classification

The presented work falls in this intersection.
The presented work falls in this intersection

Example Problem: Real-time Network Alarm Forecasting

• Increasing reliance on Telecommunication services for business and personal use
• Telecommunication Networks have a great deal of redundancy (99.999%) availability), however, the “last mile” is often a single point of failure
• Network devices emit different events data at different frequencies under different conditions. Yet they may be linked.

Systems Development: ChESS (ongoing)

ChESS: Change Event based Sensor Sampling

BUT

Interesting events always happen when nobody’s there

Solution

- real-time data stream mining
- AI based event detection
- part automation

Multi-Sensor System
- underwater monitoring
- normal maintenance & sampling by Natural Scientist

advances natural science

resource efficiency

sustainable monitoring

support of natural scientist

Agile Perception Measure when and where it matters

Applications: Intelligent Maintenance of Coastal Environments (just starting)

Ad-hoc data acquisition mesh for enhanced versatile explorations of waters

Real-time anomaly detection in aquatic system (RADiuS)

@ Prof. Jan Schulz
What are Data Streams?

Challenges & Barriers

Research Activities in Data Stream Mining

Building Data Mining Models from Data Streams

Outlook and Conclusions
### Challenges
1. Data generated at a fast rate (Velocity), at potentially large and unknown quantities (Volume)
2. Concept Drift (changes of pattern encoded in data over time)
3. Modelling real-time analytics workflows from streaming data
4. Multi-modality of data sources (text, video/images, unstructured)
5. Class label sparsity: adapting predictive models
6. Explaining Concept Drift

### Barriers
1. Limited scalable (parallel) real-time high throughput data stream mining algorithms
2. Different and changing types of concept drift
3. Lack of customisable pre-processing techniques
4. Different time stamps but co-occurring data items
5. Supervised algorithms not applicable in many cases
6. Lack of drift detectors explaining concept drift
Utility of Adaptation: when is it worth updating your model?

ROI is return on employing an adaptive predictor as compared to keeping a fixed nonadaptive model.

Thank you!

- Marine Perception Team at DFKI
- Prof. Oliver Zielinski
- Prof. Lars Nolle
- The team of the Department of Computer Science at the University of Reading
- Prof. Giuseppe Di Fatta
- UK and ISG Teams
- Prof. Mohamed Gaber
- My wife Laura
- And others.
Questions?