Data Stream Mining

GE461: Introduction to Data Science

FAZLI CAN

Computer Engineering Department

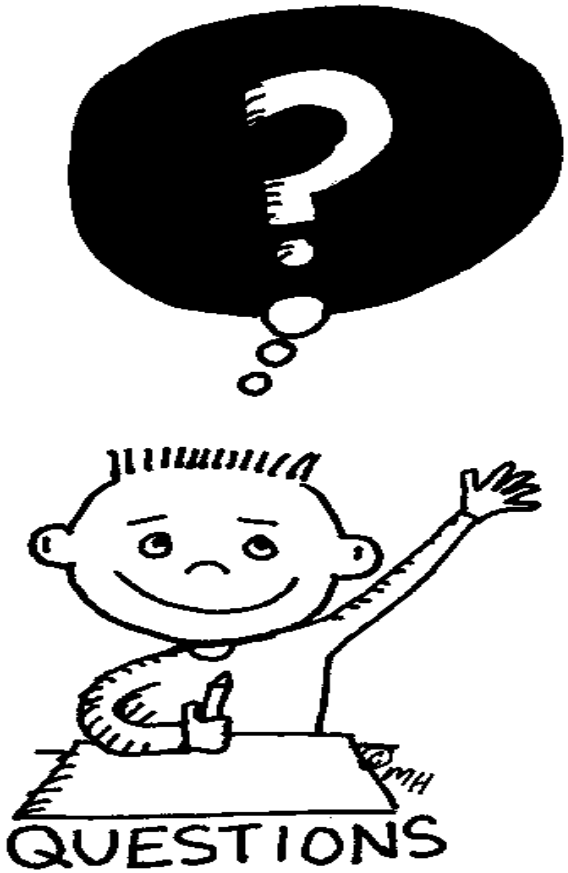
Bilkent University

April 28 & May 5, 2022

Outline

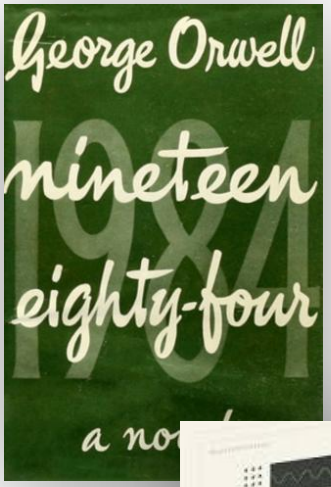
1. Is Future Today?: Current-state-of-art
2. Data stream mining
3. Concept drift
4. Single label data stream classification
5. Multi-label data stream classification
6. Ensemble approach
7. Measuring diversity in ensembles
8. Prediction effectiveness measures
9. Frameworks

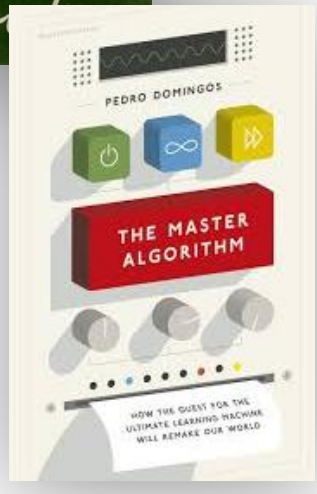
Questions?
Anytime is OK



Future: Dystopia?
Mad Max: Fury Road

Is Tomorrow Today?





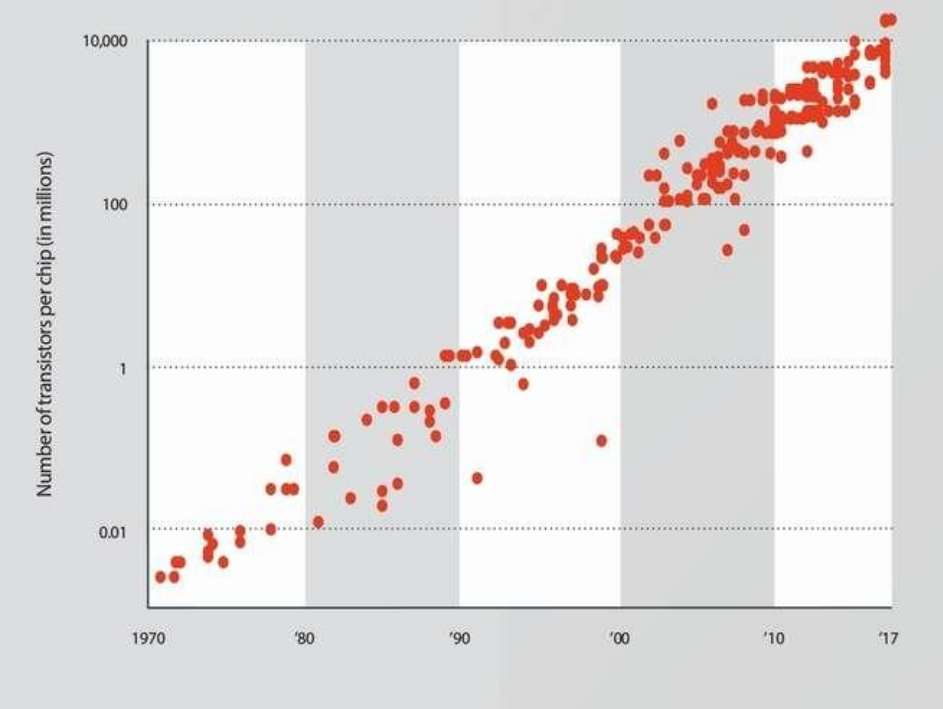
Personalization: Google

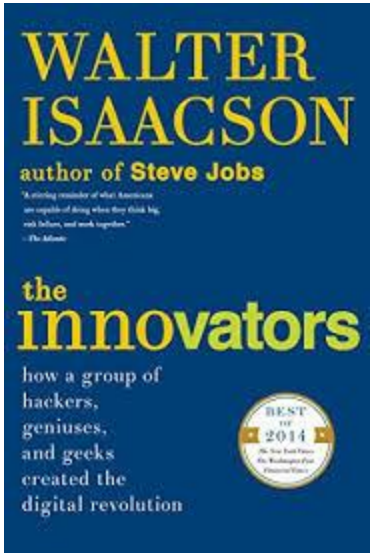
Eric Schmidt
Technical Advisor to Alphabet:

- “Google search will continue to become more personalized.”
- Google combines several signals: Page rank, click, similarity, …

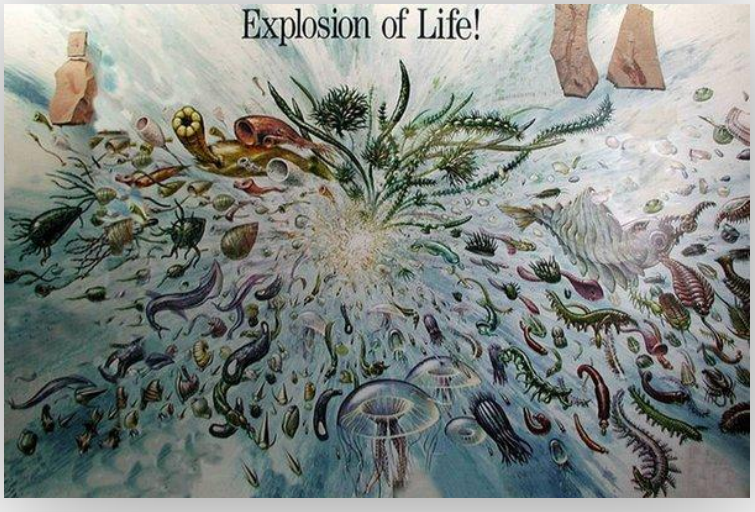
https://thenextweb.com/google/2012/02/28/eric-schmidt-google-search-will-continue-to-become-more-personalized/

von Neumann, Moore's Law… From Stored Program Concept To … ?





“A New Golden Age of Computer Architectures”: Hennessy & Patterson



Opabinia

cloud.google.com › ... › Cloud TPU › Documentation

Cloud Tensor Processing Units (TPUs) | Cloud TPU | Google ...

Cloud TPU enables you to run your **machine learning** workloads on **Google's** ... The TensorFlow server running on the host machine (the **CPU** attached to the ...

John L. Hennessy, David A. Patterson: A new golden age for computer architecture. *Commun. ACM* 62(2): 48-60 (2019).

Each Person Creates Multiple Data Streams

- I was talking about my foot pain 2 hours later I received a WebMD email about reasons for foot pain. We provide too much information about ourselves in our web searches, skype conversations, etc.

- Foot pain
- Face pain
- How old is …?

Three Vs of Data Streams: No Privacy

- Velocity
- Variety
- Volume

Big data what is big data 3 Vs of big data volume, velocity and variety day 2 of 21, https://blog.sqlauthority.com/2013/10/02/big-data-what-isbig-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/

What is Data Stream?

- Data stream is a sequence of temporal data.
- Sources : news portals, social media, stock market data (Twitter, Facebook, …)
- Data types : image, text, signal, ….

Data Stream Types

- Social media posts
- Stock tickers
- Sensor data
- News articles
- Intelligence reports
- Customer feedback
- Web clicks
- …

Data Stream Characteristics

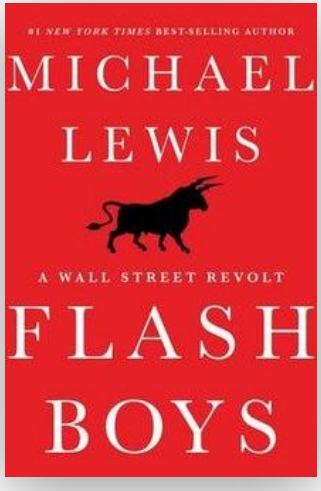
- Temporal
- Unbounded
- Time Constraint: See *Flash Boys*
- Memory Constraint
- Concept drift
- Evolving: Concept drift, new classes, new features, missing features, missing data items
- Experiments: Interleaved-Test-then-Train

Data Stream Characteristics: Revisited

- Continuous flow (data arrives over time)
- Huge amount of data (infinite data items)
- Data distribution may change over time
- Batch learning vs. incremental online learning
- Rapid arrival rate

Verizon
https://www.marketsmedia.com/

- Frequent Trades
- New York – Chicago: under15 ms
- With Light Speed? 13.3 ms



04.02.2013
By Terry Flanagan

Banks

Chicago-New York Link Edges Closer to Speed of Light

Share this on:

A direct high speed trading link between Chicago and New York will cut the round trip time by more than half a millisecond, to just under 15 milliseconds.

That's a significant boost, given that it takes light 13.3 milliseconds to travel from Chicago to New York and back again.

The link, constructed by Verizon, will connect CME Group's data center in in Aurora, Ill. to New York metro financial markets, offering an ultra-low latency round-trip route that bypasses traditional area "data hotels."

Data Stream Mining: Three Major Problems

- Class prediction: What is it?
- Concept drift: Evolving data items and relationships
- Evolving classes: New classes, Dying Classes

Data Stream Mining Problems
1. Class Prediction

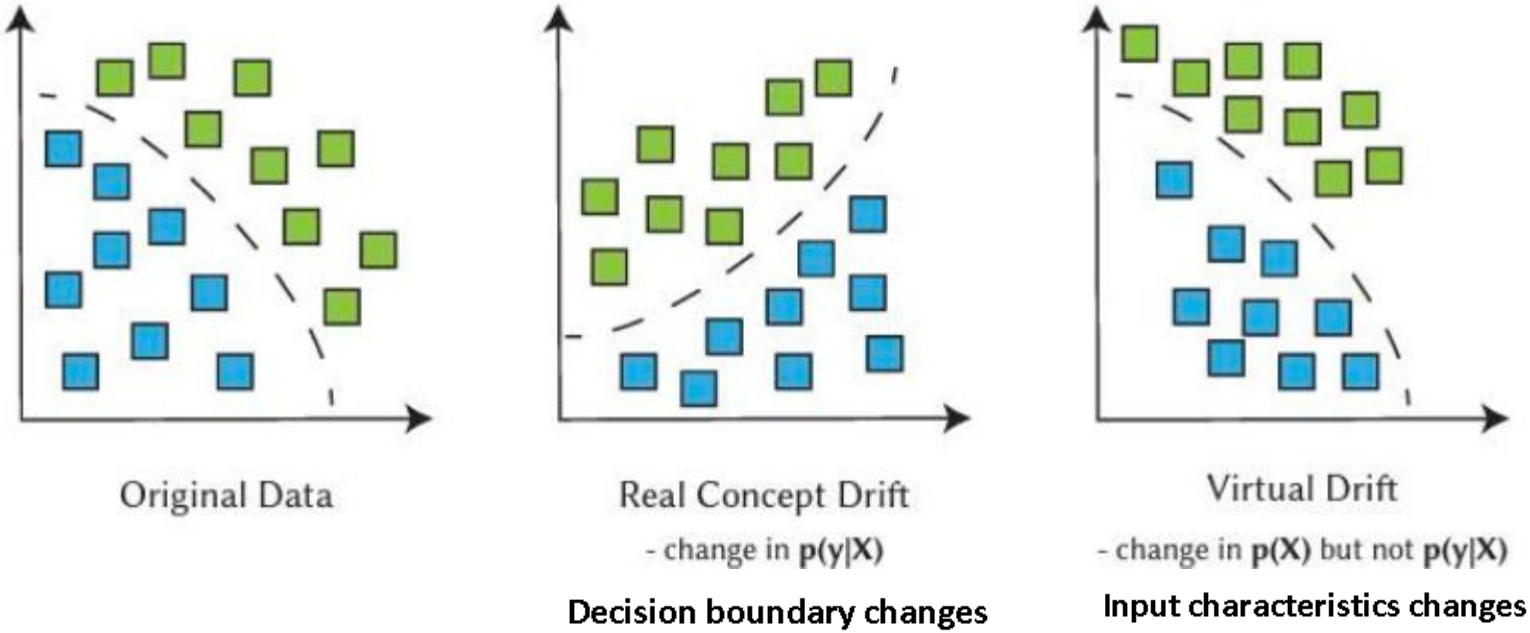
- Unary: From us vs. not from us (anomaly detection)
- Multi-class: Choose one out of many (comedy)
- Multi-label: Choose more than one out of many (comedy, action)

Data Stream Mining:
2. Concept Drift Detection

- Y: Output, X: input, P(Y/X): class prediction for X
- Real Concept Drift
  - Classification boundary changes (X remains the same, P(Y/X) changes
    - Esad ➔ Esed
    - David Lean movie: Bridge over the River Kwai

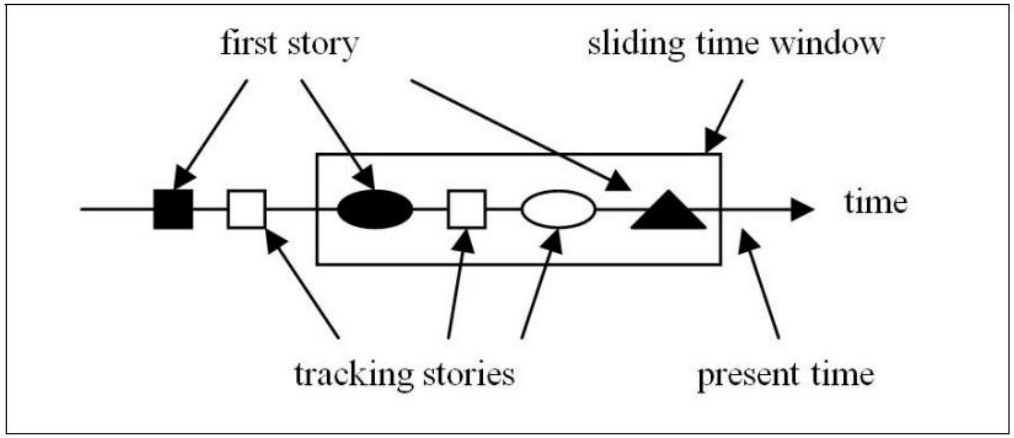
- Virtual Concept
  - Data items of a class change: Virtual Concept Drift
    - For people definition of beauty changes

Data Stream Mining:
2. Concept Drift Detection



Data Stream Mining:
3. Evolving Classes

- Discover new patterns: New classes
- New Becomes Old: New event detection and tracking
- Also considered as a type of concept drift (not considered in this course)



Data Stream Processing

- Destination: human or application
- Personalized vs. public
- What is the intended use: summarization...
- Flow size and time constraint
- Scalability: Centralized vs. distributed
- Congestion: Resource adaptive
- Dynamic/incremental processing

Data Stream Classification

- Updated Data Feeding
- Scheduled Feeding of Ensemble Members
- Add/Drop Classifiers
- Feature Regulation

How to Handle Data Streams

1. Process one example at a time and inspect it only once
2. Use a limited amount of memory
3. Be ready to predict at any time
4. Be able to react to concept drift in case of evolving data streams

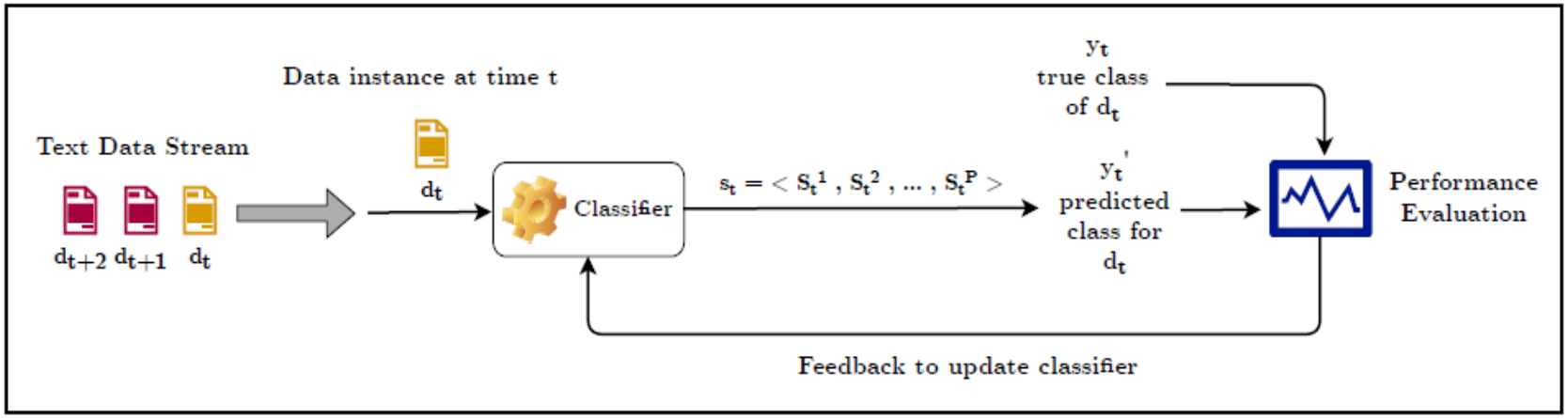
Bifet, Albert, et al. MOA: Massive online analysis. *Journal of Machine Learning Research* 11. 1601-1604 (2010).

Online Learning vs. Batch Learning

- Online learning techniques are used to handle data stream classification
- Online vs Batch learning :
  1. Data is not stored in the memory
  2. The system sees the data once
  3. No epochs on data during the learning process
  4. Algorithm should be much faster than batch learning

Data Stream Classification Evaluation

- No separate train and test sets
- We can use interleaved-test-then-train method
- Simple idea : we use each instance first to **test** the model, and **then** to **train** the model



Now More on: Concept Drift

- What is concept drift?
- Concept drift types
- Concept drift detection:
  - Prediction performance decrease
  - Using an Algorithm: supervised vs. unsupervised
- Concept drift handling
  - Active: Detect and retrain
  - Passive: Let the system handle

Lu, A. Liu, F. Dong, F. Gu, J. Gama and G. Zhang. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12), 2346-2363 (2019).

What is Real Concept Drift?

- Concept drift is the change in the relationships between input and output data in the underlying problem over time.
  Decision: P(Y/X), Input: P(X)
- Real Concept Drift:
  A user may be interested in a movie genre and become uninterested later.
  X: Input (genre) remains the same
  Y: Output (class) changes

What is Virtual Concept Drift?

- P(Y/X), Input: P(X)
- Virtual Concept Drift:

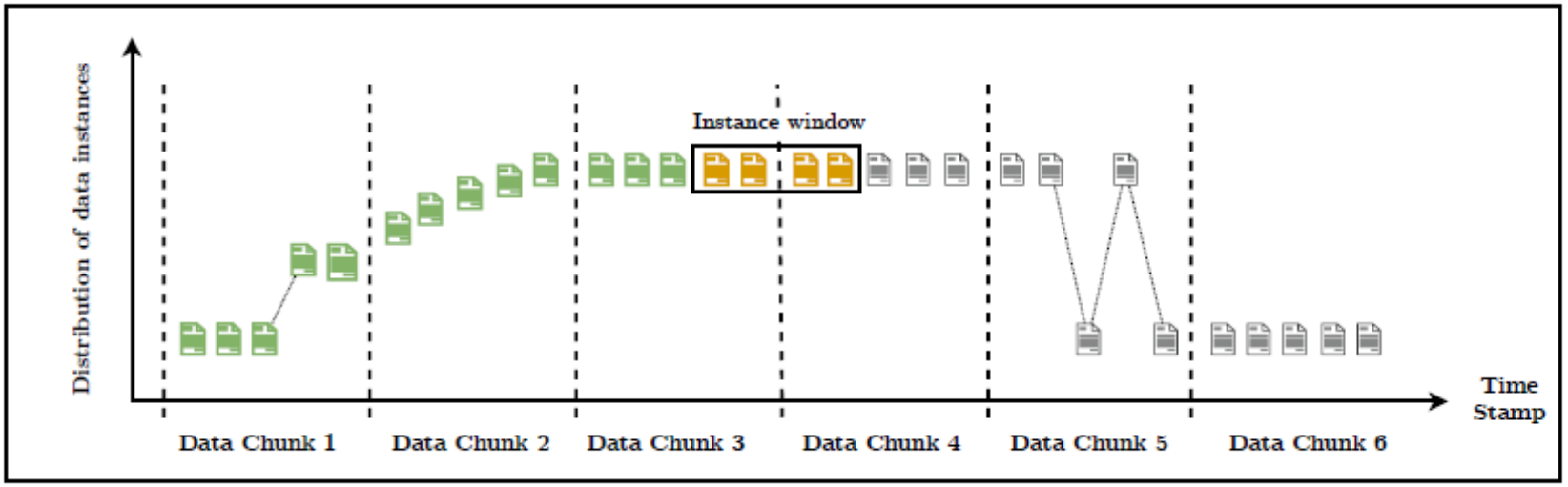
165 cm: definition of tall man in 19th century
175 cm: definition of tall man in 20th century
185 cm: definition of tall man in 21st century

X: Input changes
Y: Output remains the same

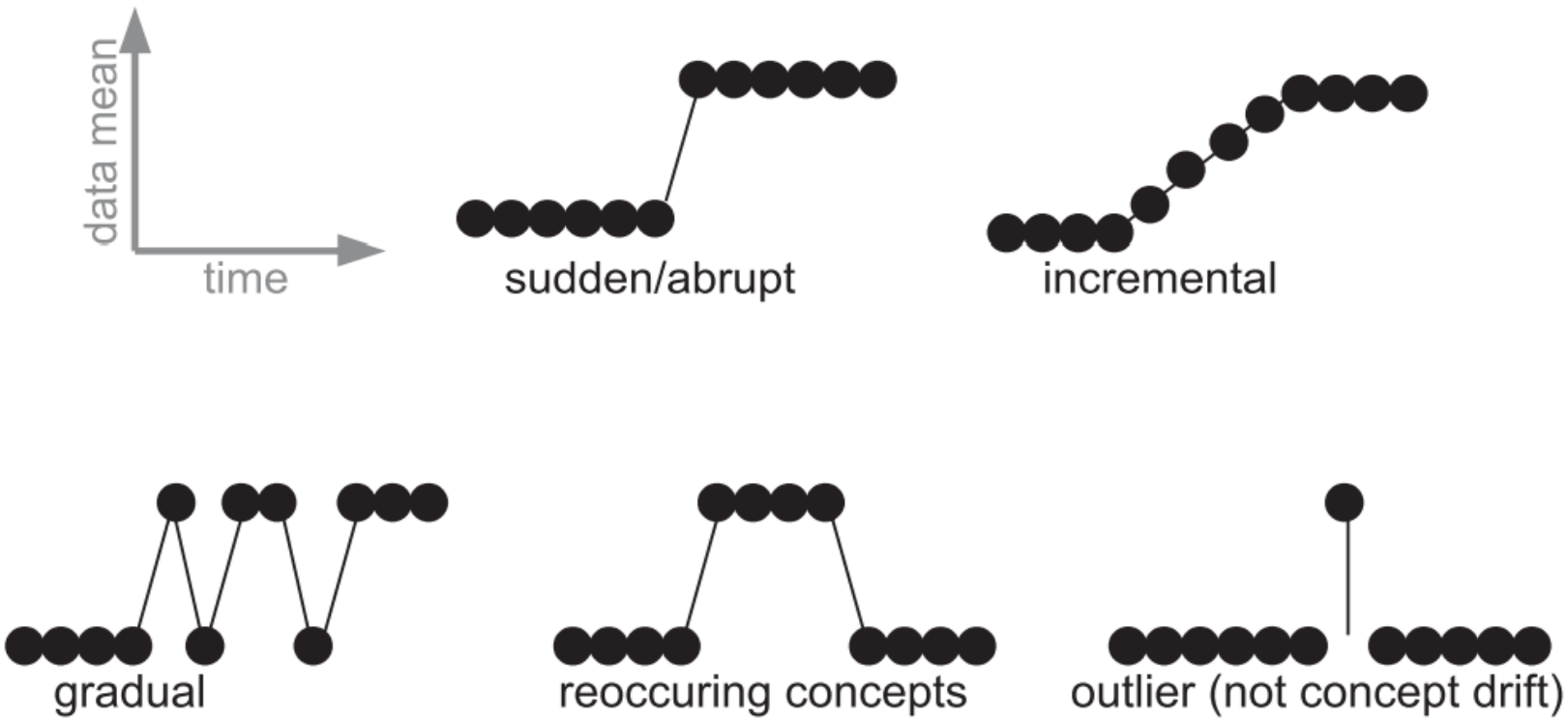
Concept Drift Types

- Abrupt /Sudden
- Gradual
- Incremental
- Reoccurring

Concept Drift Types



Concept Drift Types



Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 44:1–44:37 (2014).

Concept Drift Detection

- Concept drift detection algorithms find the drift point in a data stream.
- They usually detect the change based on the changes in the distribution of data over time.
- Detectors send an alarm in case of any changes and the system adapts to the changes based on a concept drift adaptation algorithm.

Concept Drift Handling

1. Update the model over time in certain time order. For example, update the model every week and learn the new data.
2. Incrementally learn the new data over time.
3. Assign more weights to incoming data right after the drift.
4. Restart learning from the drift point
5. Use ensemble of weak classifiers to adjust to the changes

Multi-class Single Label Data Stream Classification

- Using a single classifier
- Using ensemble methods

Multi-label Classification

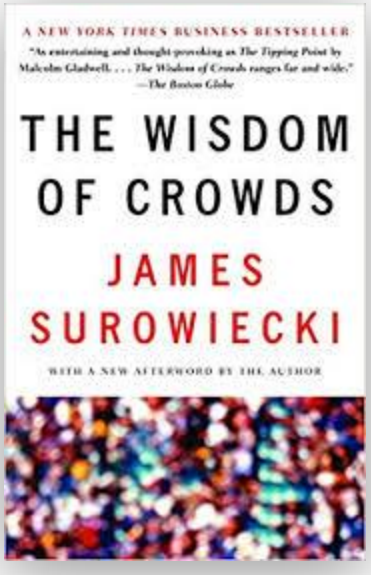
- In multi-label classification each data instance may belong to several classes.
- For instance, a movie may belong to several genres.
- For multi-label classification we can use single classifier or ensemble method.

Online Multi-label Classification

- Most of the Online multi-label classification algorithms are based on ensemble methods :
  1. GOOWE-ML-based methods:
  2. A classifier for each class
  3. Online Bagging (OzaBagging)-based methods
  4. ADWIN Bagging-based methods

A. Büyükçakir, H. Bonab, F. Can : A Novel Online Stacked Ensemble for Multi-Label Stream Classification. CIKM 2018: 1063-1072.

Ensemble Approach:
The Wisdom Of Crowds



Using Ensemble Methods

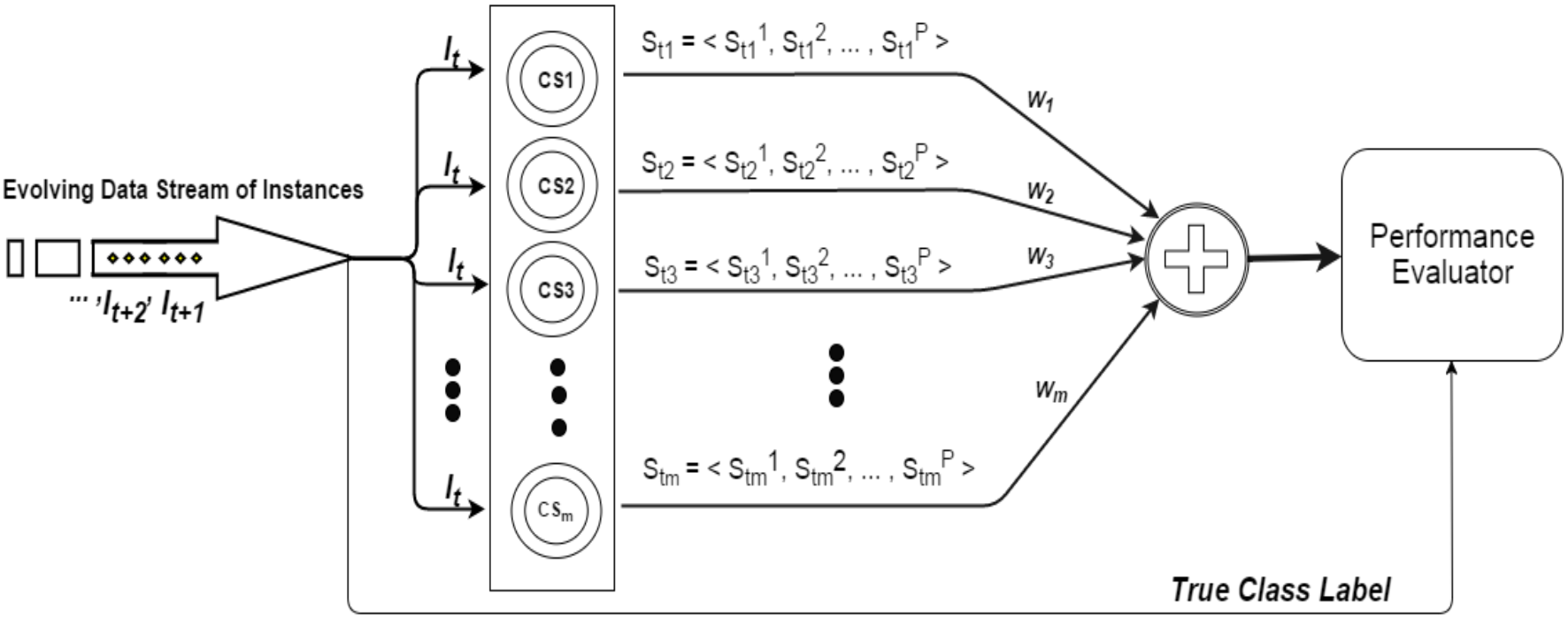
- Ensemble learning is a paradigm where multiple machine learning algorithms or better to say, weak learners are combined to get better results.
- Some Ensemble methods:
  1. Bagging: sampling with replacement and averaging
  2. Boosting: combine weak learners to obtain a strong learner
  3. Stacking: a learner that learns classifiers
- Using these methods and voting systems like majority voting, the system can adapt to new changes.

GOOWE

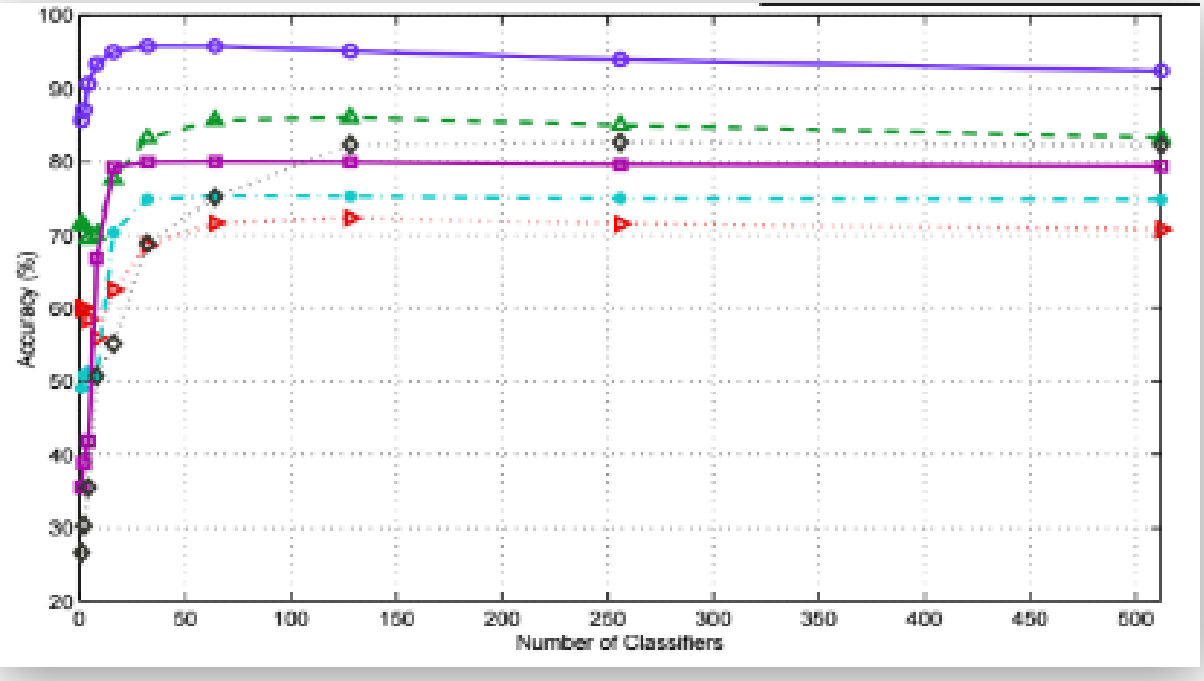
- Notations and Assumptions:
  - **p class labels** as C = (C1, C2… Cp)
    multi-class problem
  - **m classifier** systems as CS = (CS1, CS2… CSm)
  - For each of these Instances (I_i), every classifier system CS_j returns a set of scores as $S_{ij} = (S_{ij}^1, S_{ij}^2 \ldots S_{ij}^p)$

Bonab, H., Can, F. GOOWE: Geometrically optimum and online-weighted ensemble classifier for evolving data streams, ACM Transactions on Knowledge Discovery from Data, 12(2), 1-25:33 (2018).

GOOWE (Cont.)

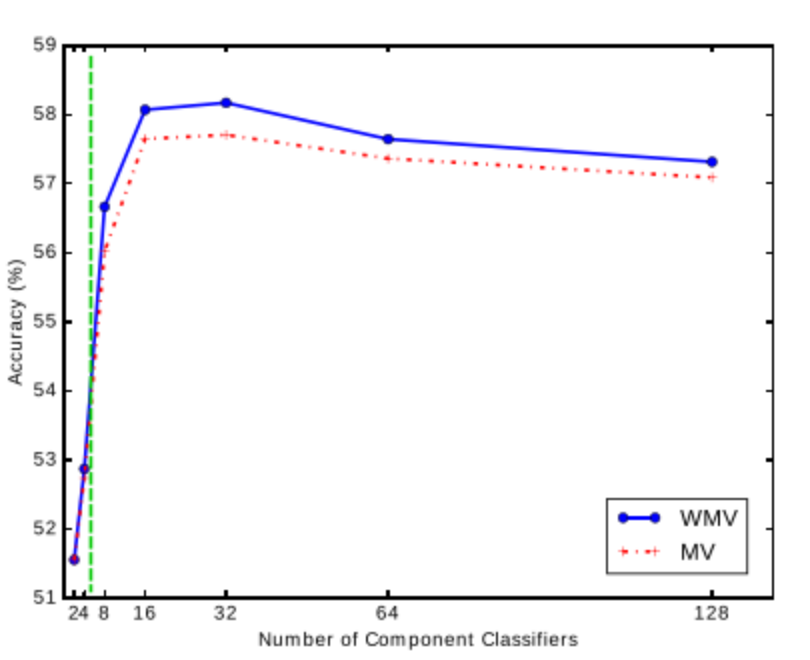

Evolving Data Stream of Instances
... , I_{t+2}, I_{t+1}
I_t
I_t
I_t
I_t
CS1
CS2
CS3
CS_m
$S_{t1} = < S_{t1}^1, S_{t1}^2, \ldots, S_{t1}^P >$
$S_{t2} = < S_{t2}^1, S_{t2}^2, \ldots, S_{t2}^P >$
$S_{t3} = < S_{t3}^1, S_{t3}^2, \ldots, S_{t3}^P >$
$S_{tm} = < S_{tm}^1, S_{tm}^2, \ldots, S_{tm}^P >$
w_1
w_2
w_3
w_m
Performance Evaluator
True Class Label

Law Of Diminishing Returns



Bonab, H., Can, F. GOOWE: Geometrically optimum and online-weighted ensemble classifier for evolving data streams. ACM Transactions on Knowledge Discovery from Data, 12(2), 1-25:33 (2018).

Components must be Independent and Diverse (m = p)



The highest effectiveness is observed much closer to the theoretically ideal m=p green line rather than the maximum number of components.

H. Bonab, F. Can: Less Is More: A Comprehensive Framework for the Number of Components of Ensemble Classifiers. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9), 2735-2745 (2019).

.

What Is Diversity Measure in Ensemble Learning?

- The key point in defining an ensemble of classifiers is to choose diverse classifiers.
- There is no universal definition for diversity in ensemble methods
- This make it hard to measure diversity

Kuncheva, Ludmila I., and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* 51.2: 181-207 (2013).

Key Questions in Diversity Measurement

1. How do we define and measure diversity?
2. How are the various measures of diversity related to each other?
3. How are the measures related to the accuracy of the team?
4. Is there a measure that is best for the purposes of developing committees that minimize error?
5. How can we use the measures in designing the classifier ensemble?

What is Q-statistics?

- Yul's Q statistic :

For two classifiers D_i and D_k we have:

- $Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}$

	D_k correct (1)	D_k wrong (0)
D_i correct (1)	N^{11}	N^{10}
D_i wrong (0)	N^{01}	N^{00}

Total, $N = N^{00} + N^{01} + N^{10} + N^{11}$.

What is Q-statistics?

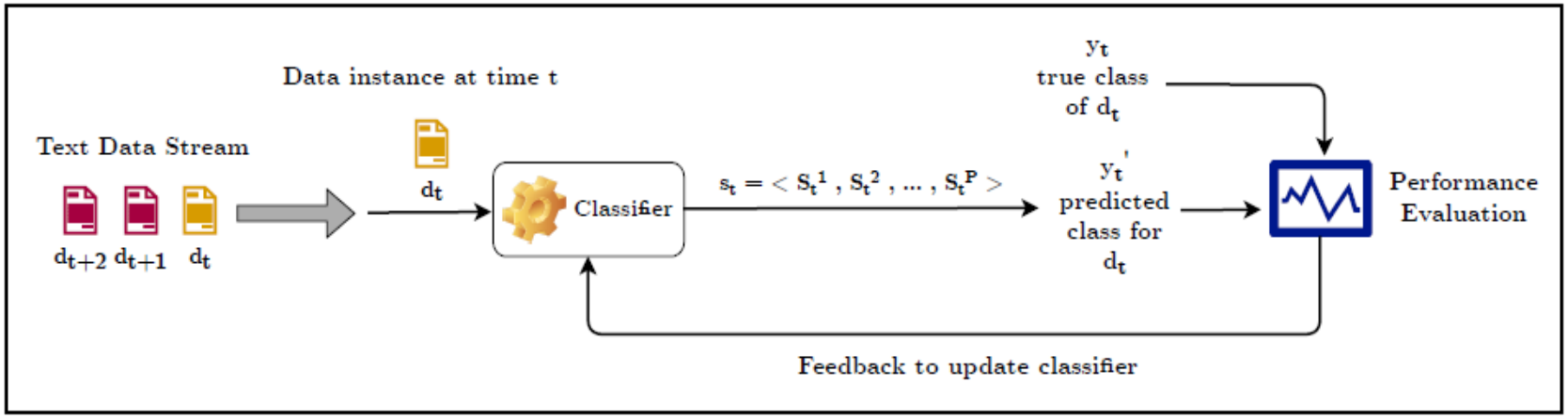
- For instance, suppose that for D_i and D_k we have the following predictions for 10 instances :
- $Predictions\ of\ D_i = \ < 0,\ 1, 0, 1, 0, 0, 1, 1, 1, 1 >$
- $Predictions\ of\ D_k = < 0,\ 0, 0, 0, 0, 1, 1, 0, 1, 1 >$
- $Ground\ Truth = \quad < 1,\ 0, 0, 1,\ 0, 0, 1, 1, 1, 1 >$
- $N^{11} = 5\ , N^{10} = 3\ , N^{01} = 1\ , N^{00} = 1$
- $Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} = \frac{5*1\ -1*3}{5*1+1*3} = \frac{2}{8}$

	D_k correct (1)	D_k wrong (0)
D_i correct (1)	N^{11}	N^{10}
D_i wrong (0)	N^{01}	N^{00}

Total, $N = N^{00} + N^{01} + N^{10} + N^{11}$.

Prediction Effectiveness Measures

- No separate train and test sets
- We can use interleaved-test-then-train method
- Simple idea : we use each instance first to **test** the model, and **then** to **train** the model



Prequential Evaluation

- While calculating prequential accuracy, each data instance is used for two purpose. First we use it for testing then for training.
- Overall accuracy at time t, is the accumulated prequential accuracy calculated for all the stream data until time t.

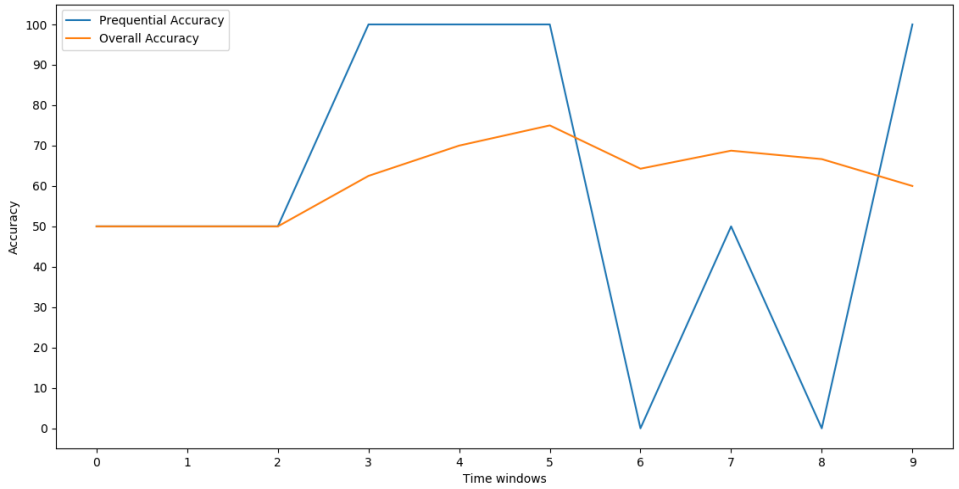
Window Size is Important

- Narrow window: Minor swings can labeled as concept drift
  - Observing a student during a semester
- Wide window: Major swings can be missed
  - Just looking at student overall GPA

Prequential Evaluation Example 1

There are 2 data instances in each time windows

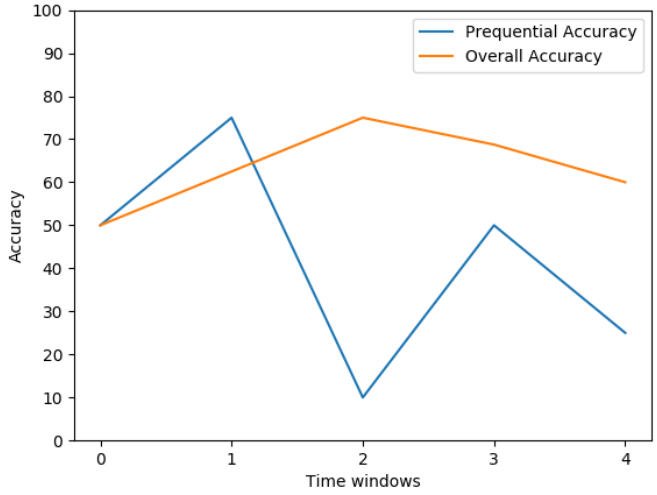
Pediction	1,0	0,1	1,0	1,1	1,1	1,1	0,0	1,1	0,1	0,0
Prequential acc	50%	50%	50%	100%	100%	100%	0%	50%	0%	100%
Overall acc	50%	50%	50%	62.5%	70%	75%	64.28%	68.75%	66.66%	60%



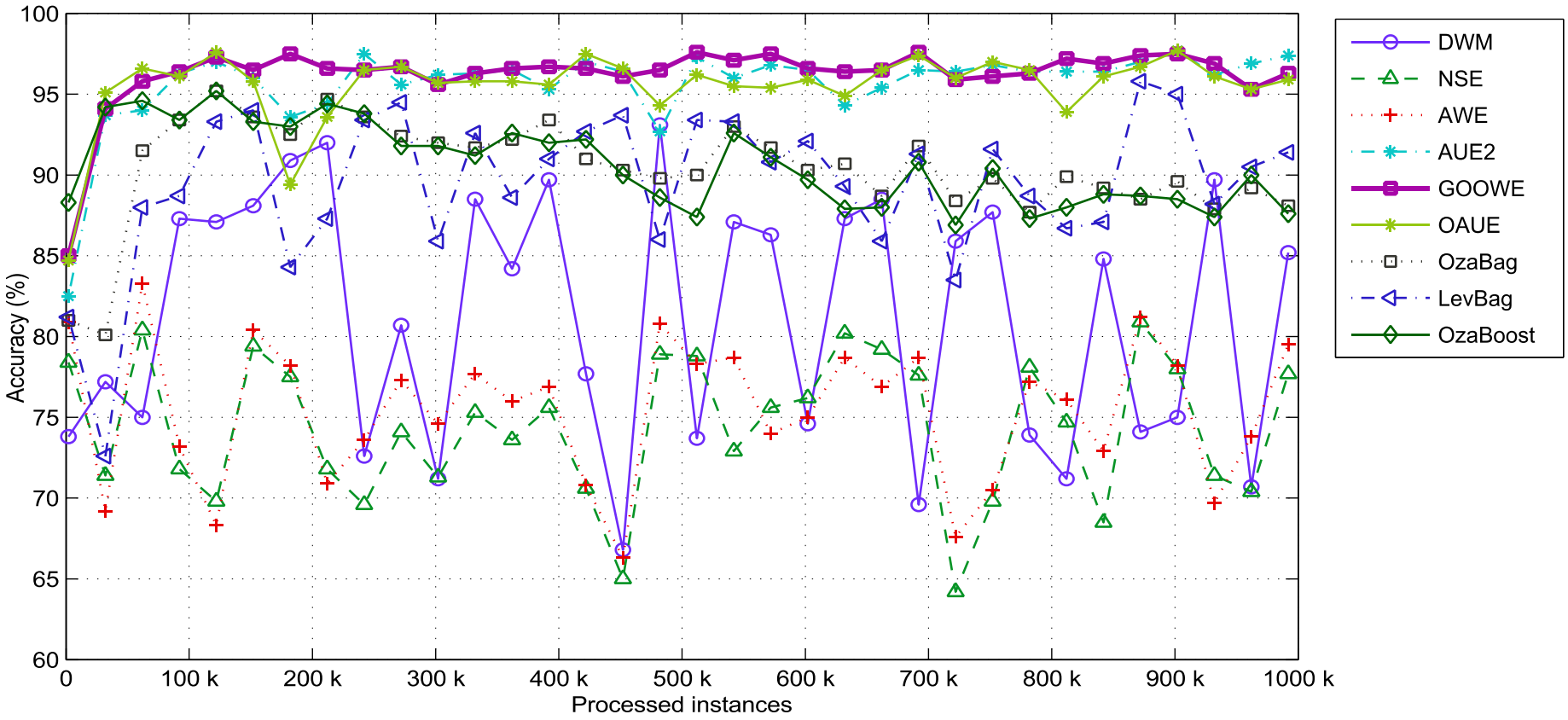
Prequential Evaluation Example 2

There are 4 data instances in each time windows

Pediction	1,0,0,1	1,0,1,1	1,1,1,1	0,0,1,1	0,1,0,0
Prequential acc	50%	75%	100%	50%	25%
Overall acc	50%	62.50%	75%	68.75%	60%



Behavior of Different Classifiers
Rigorous Concept Drift: Gradual Change RBF-G-4-F



(a) accuracy

H. Bonab, F. Can: GOOWE: Geometrically Optimum and Online-Weighted Ensemble Classifier for Evolving Data Streams. *ACM Transactions on Data Discovery from Data* 12(2): 25:1-25:33 (2018)

Multi-label Classification Evaluation

- Evaluation metrics :

1. Instance-Based Metrics
2. Label-Based Metrics
3. Efficiency Metrics: Time and Space

Instance-based Metrics

Instance-based metrics are evaluated for every instance and averaged over the whole dataset

1. Exact Match
2. Hamming Score,
3. Instance-Based (Accuracy, Precision, Recall, F1-Score)

Instance-based Metrics

$$Exact\ match = \frac{1}{N}\sum_{i=1}^{N}[\![y^i = \hat{y}^i]\!]$$

$$Hamming\ score = \frac{1}{LN}\sum_{i=1}^{N}\sum_{j=1}^{L}[\![y_j^i = \hat{y}_j^i]\!]$$

Instance-based Metrics

Example based accuracy, precision, recall and F1 scores

$$Acc_{ex} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y^i \cap \hat{y}^i|}{|y^i \cup \hat{y}^i|} \qquad Pr_{ex} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y^i \cap \hat{y}^i|}{|\hat{y}^i|}$$

$$Re_{ex} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y^i \cap \hat{y}^i|}{|y^i|} \qquad F1_{ex} = \frac{2 * Pr_{ex} * Re_{ex}}{Pr_{ex} + Re_{ex}}$$

Instance-based Metrics Example

- Assume a labelset with 5 labels for a given instance we have :
- $Prediction: \hat{y} = \quad < 1,1,0,0,0 >$
- $Ground\ Truth\ y = < 0,1,1,1,0 >$
- Exact match = 0 , because two vectors are not completely the same
- Hamming score = $^2/_5$, because there are two matching bits between the vectors
- $Acc_{ex} = {}^1/_4$, because 1 relevant bit is mutual among the relevant bits of two vectors
- $Pr_{ex} = {}^1/_2$, because 1 relevant bit is the same among 2 predicted relevant bits
- $Re_{ex} = {}^1/_3$, because 1 relevant bit is the same among 3 correct bits from the ground truth.

Alican Büyükçakır, GOOWE-ML: a novel online stacked ensemble for multi-label classification in data streams, Master Thesis, Computer Engineering Department Bilkent University, 2019.

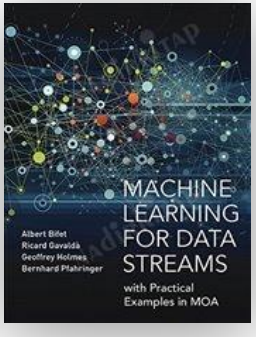
Statistical Tests

For selling your ideas statistical tests are essential in the comparison of methods

1. Show statistical significance of your results: To show that which method is significant better than the baselines
2. Use strong baselines (avoid strawman)
3. Use accessible datasets
4. Make sure that other people can repeat your experiments
5. Remember important keywords of CS research: Scalability, Dynamism, Robustness

MOA

- MOA (Massive Online Analysis) is one of the popular frameworks for data stream analysis in JAVA.
- MOA contains some of the most famous machine learning algorithms e.g. Hoeffding Tree, Naïve Bayes,…
- There are generators in MOA which simulate the stream environment by generating the data over time.
- https://moa.cms.waikato.ac.nz/



Scikit-multiflow

- Scikit-learn is a famous machine learning framework in python
- However, scikit-learn doesn’t contain any module for handling data streams
- Scikit-multiflow is a framework which is the equivalent of MOA in python
- It also has the capabilities of scikit-learn
- https://scikit-multiflow.github.io/

Current Research Opportunities

- Ensemble pruning: which components to delete
- Concept drift detection: supervised vs. unsupervised
- Multi stream environment: fusing different streams
- Construction of ensemble of ensembles: a possibility

REFERENCES

- Bifet, A., et al. MOA: Massive online analysis. *Journal of Machine Learning Research* 11.May (2010): 1601-1604.
- Big data what is big data 3 Vs of big data volume, velocity and variety day 2 of 21. https://blog.sqlauthority.com/2013/10/02/big-data-what-isbig-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/
- H. Bonab, F. Can: Less Is More: A Comprehensive Framework for the Number of Components of Ensemble Classifiers. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9), 2735-2745 (2019).
- H. Bonab, F. Can: GOOWE: Geometrically Optimum and Online-Weighted Ensemble Classifier for Evolving Data Streams. *ACM Transactions on Knowledge Discovery from Data* 12(2), 25:1-25:33 (2018).
- Büyükçakır, A. GOOWE-ML: A novel online stacked ensemble for multi-label classification in data streams, Master Thesis, Computer Engineering Department Bilkent University, 2019.
- A. Büyükçakir, H. Bonab, F. Can : A Novel Online Stacked Ensemble for Multi-Label Stream Classification. CIKM 2018: 1063-1072.
- Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A. A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 44:1–44:37 (2014).
- G. I. Webb, R. Hyde, H. Cao, H.-L. Nguyen, F. Petitjean. Characterizing Concept Drift. CoRR abs/1511.03816 (2015)

REFERENCES

- J. L. Hennessy, D. A. Patterson: A new golden age for computer architecture. *Commun. ACM* 62(2): 48-60 (2019).
- Kuncheva, L. I., and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* 51.2: 181-207 ((2003).
- Lewis, M. *A Wall Street Revolt: FlashBoys*, Norton paperback (2014).
- Lu, A. Liu, F. Dong, F. Gu, J. Gama and G. Zhang, 2019, Learning under Concept Drift: A Review, *IEEE Transactions on Knowledge and Data Engineering*. 31(12), 2346-2363 (2019).
- MOA: https://moa.cms.waikato.ac.nz/
- Scikit multiflow: https://scikit-multiflow.github.io/