Online learning / concept drift

Filip Paulů

Czech Technical University in Prague

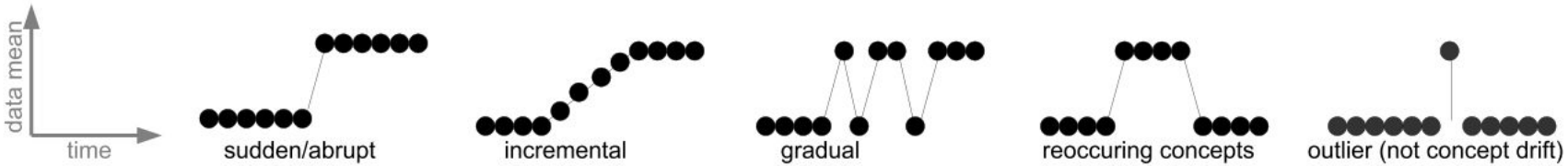
Motivation

- Online news
- Customers’ buying preferences
- Whether prediction
- Electrical sensors
- Hardware or software faults
- 2.8 zetabytes data generated in 2012
- NetFlix
- ...

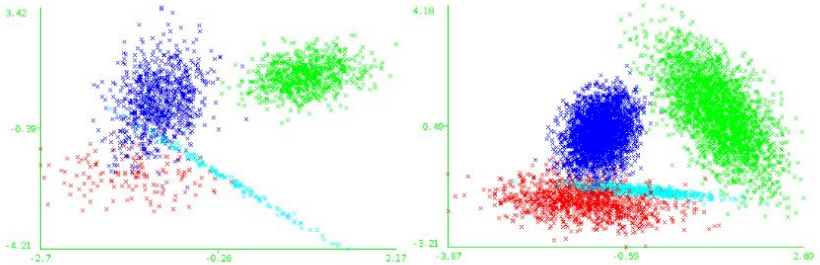
Concept Drift

- Changing data distribution.
- Hidden context.

Types of concept drift

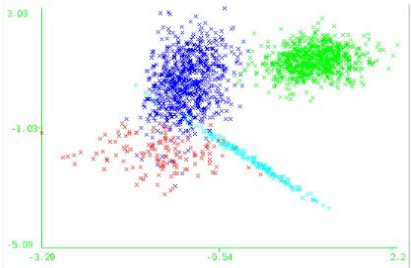


Initial and final concepts for two data streams

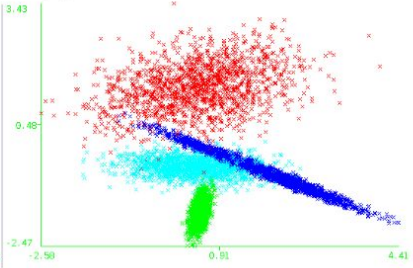


(a) Stream 1: Initial Concept

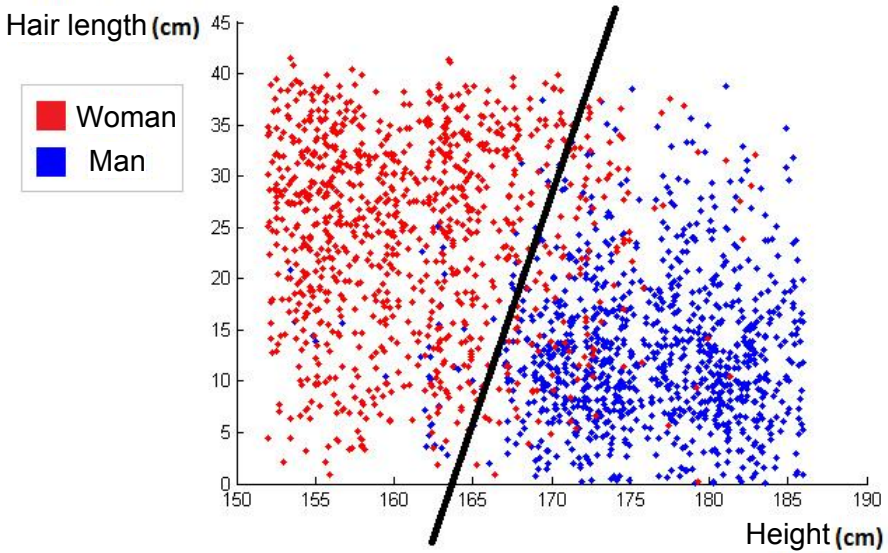
(b) Stream 1: Final Concept



(c) Stream 2: Initial Concept



(d) Stream 2: Final Concept



Popularity

- Spans across different research fields.
- Scattered among various communities.
- Goes beyond the areas of machine learning, data mining.

Adaptation algorithms

Passive approach

- continuously updates the model over time

Active approach

- detects change of data distribution

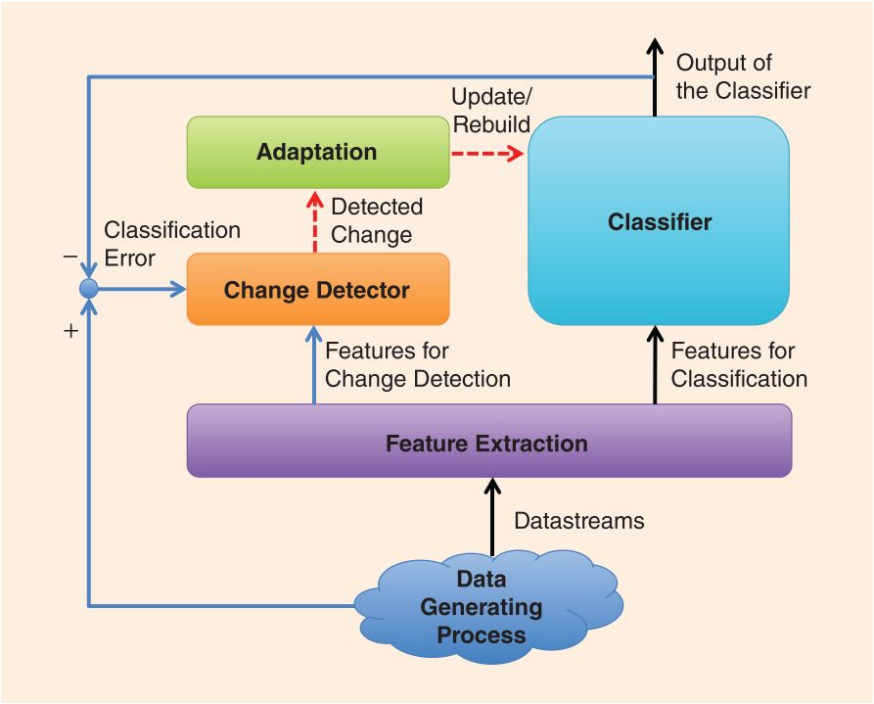
Passive Approaches

- Simply accepts that the underlying data distributions may (or may not) change at any time with any rate of change.

- Avoiding the potential pitfall associated.

Models: Single Classifier Models, Ensemble Classifier Models, Streaming Ensemble Algorithm

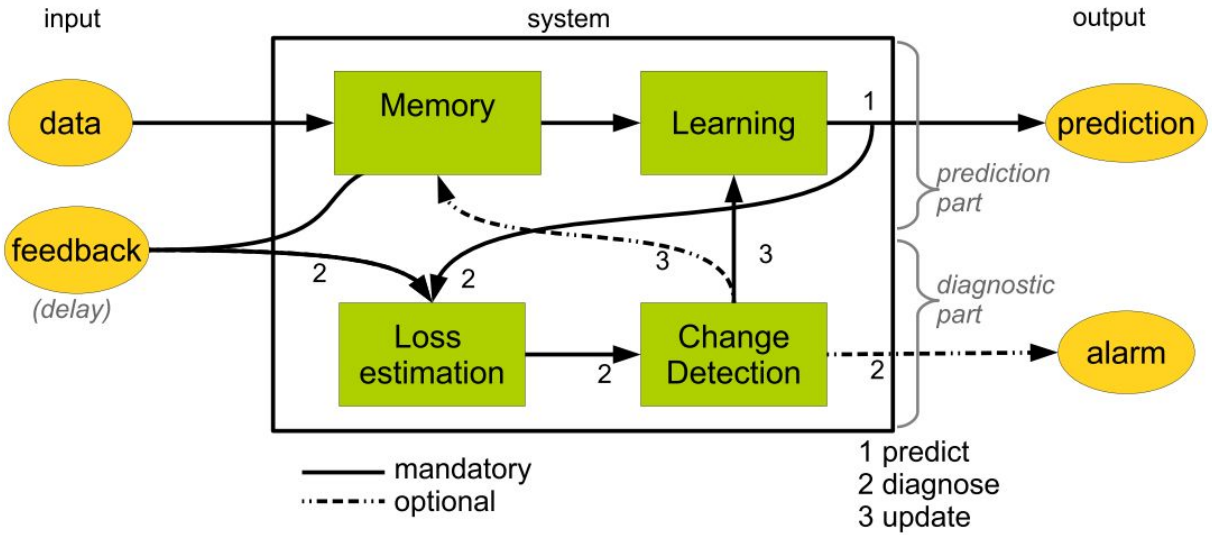
Active Approaches



- Detector of data distribution changes.
- May fail to detect a change or falsely detect a non-existent change (false alarm).

Online adaptive learning procedure

- Predict
- Diagnose
- Update



$$y = \mathcal{L}(X)$$

$$\mathcal{L}_{t+1} = \texttt{train}\left((X_i, y_i), \ldots, (X_t, y_t), \mathcal{L}_t\right)$$

Online adaptive learning procedure

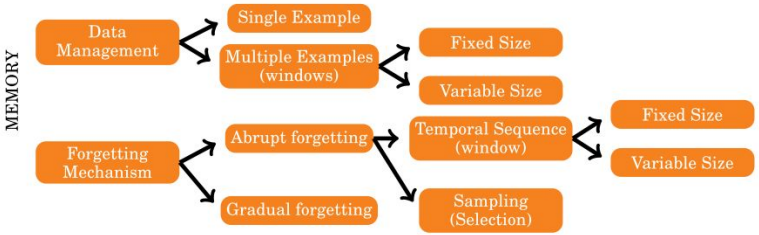


Fig. 5. Taxonomy of memory properties of methods.

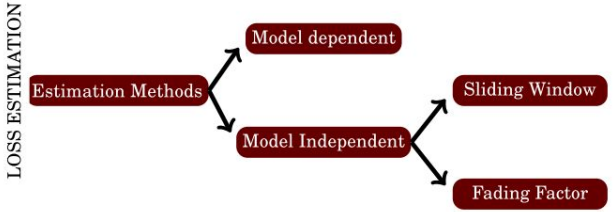


Fig. 8. Taxonomy of loss estimation properties of methods.

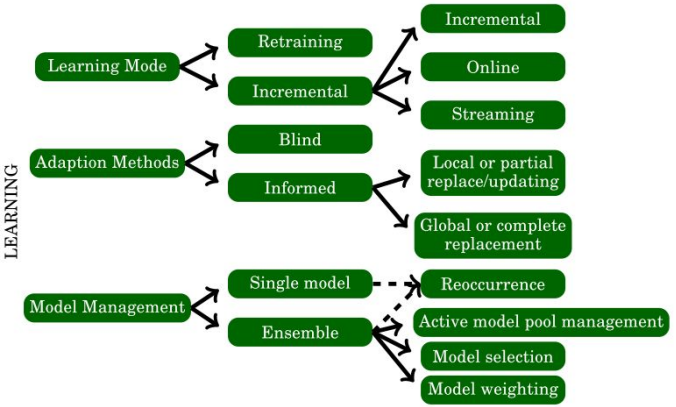


Fig. 7. Taxonomy of learning properties of methods.

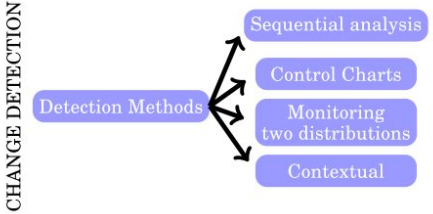


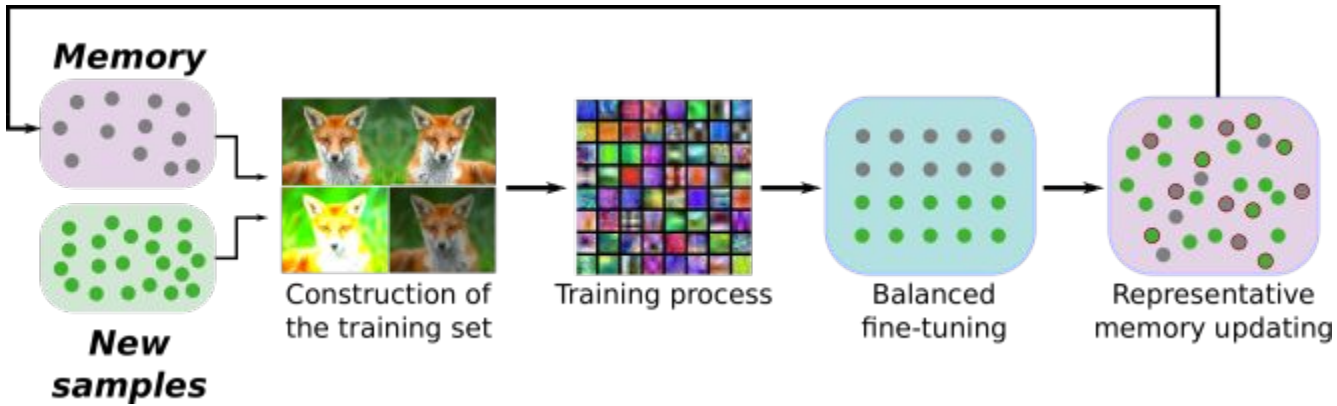
Fig. 6. Taxonomy of control properties of methods.

Categorization of Learning Techniques

Learning Mode		
Retraining		[Street and Kim 2001], [Zeira et al. 2004], [Klinkenberg and Joachims 2000]
Incremental		[Schlimmer and Granger 1986],[Littlestone 1987], [Bifet et al. 2009],[Hulten et al. 2001],[Polikar et al. 2001]
Streaming		[Gama et al. 2006],[Ikonomovska et al. 2011]
Adaptation Methods		
Blind		[Littlestone 1987], [Maloof and Michalski 2000], [Klinkenberg and Renz 1998] [Chu and Zaniolo 2004], [Bessa et al. 2009]
Informed		[Hulten et al. 2001], [Gama et al. 2006],[Ikonomovska et al. 2011]
Model Adaptation		
Model Specific		[Hulten et al. 2001], [Gama et al. 2006], [Harries et al. 1998]
Model Independent		[Wald 1947], [Gama et al. 2004], [Wang et al. 2003] [Bifet and Gavalda 2006], [Kuncheva and Zliobaite 2009]
Model Management		
Single Model		[Hulten et al. 2001], [Gama et al. 2006], [Ikonomovska et al. 2011]
	Recurrent	[Widmer 1997], [Gama and Kosina 2011]
Ensemble		[Polikar et al. 2001], [Street and Kim 2001],[Kolter and Maloof 2005], [Gao et al. 2007], [Minku and Yao 2011], [Elwell and Polikar 2011]
	Recurrent	[Yang et al. 2006], [Katakis et al. 2010], [Gomes et al. 2011]

Incremental learning

- Model is continuously extended.



Incremental learning

- Change of the underlying distribution

$$p_t(\mathbf{x}, y) \neq p_{t-1}(\mathbf{x}, y)$$

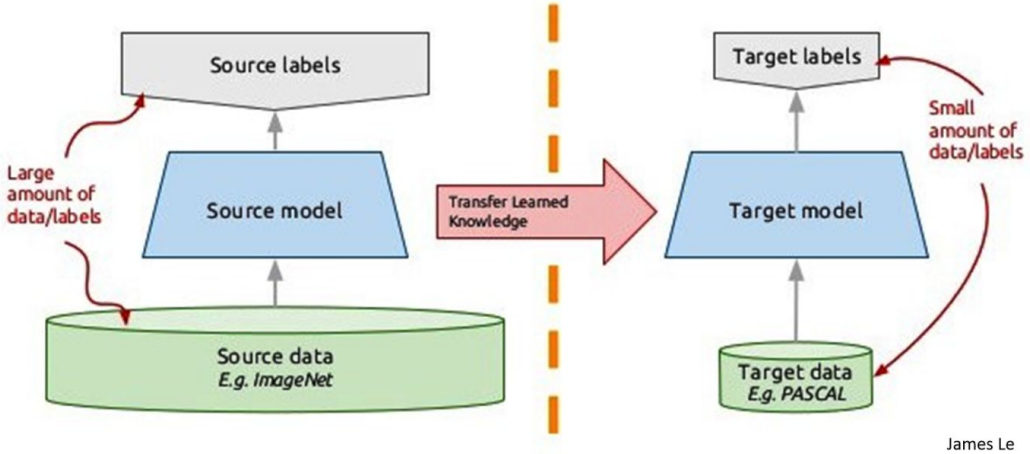
- Model $$F_t = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{(\mathbf{x},y)\in p_t(\mathbf{x},y)}[\ell(f(\mathbf{x}), y)]$$

- Whole learning process

$$\min_{F_1, F_2, \ldots, F_t, \ldots} \sum_t \mathbb{E}_{(\mathbf{x},y)\in p_t(\mathbf{x},y)}[\ell(F_t(\mathbf{x}), y)]$$

Transfer learning

- Storing gained knowledge.
- Applying it to a different but related problem.



Real examples

Online learning

DTEL (2018)

Diversity and Transfer-based Ensemble Learning (DTEL).

- Which historical models should be preserved?
- How to utilize them?

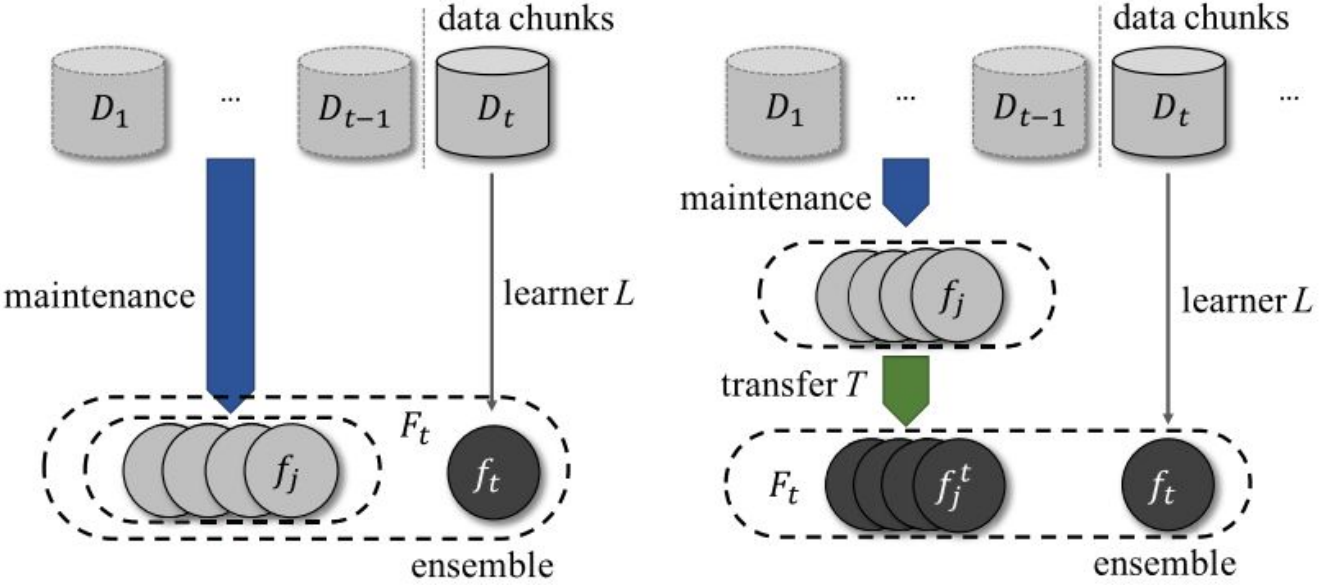
- transfer learning
  - decision tree

DTEL - proposed approach

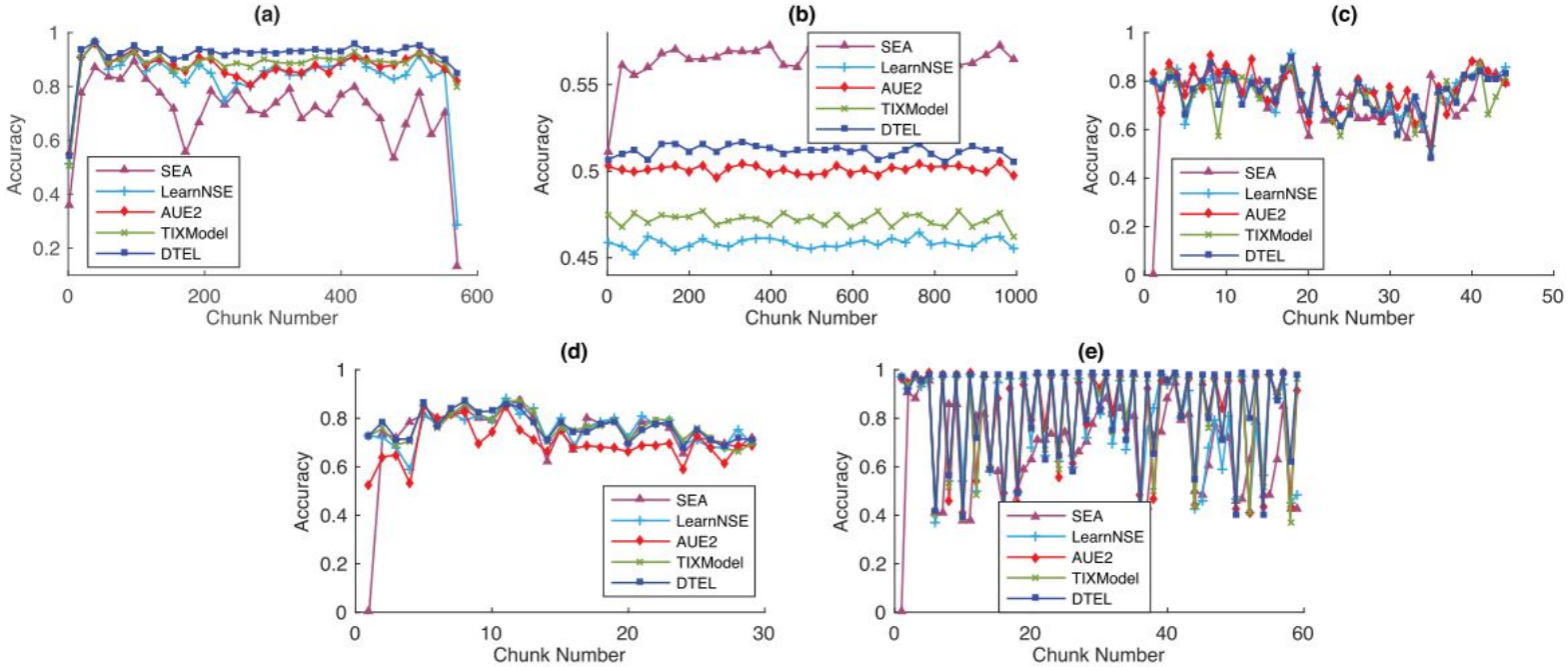
When a new data chunk arrives, all the approaches utilize the preserved historical models without adapting them to the new training data.

In the extreme case the adaptation weight is zero (only transfer learning).

DTEL - learning flow



DTEL - Accuracy results on real-world data streams.



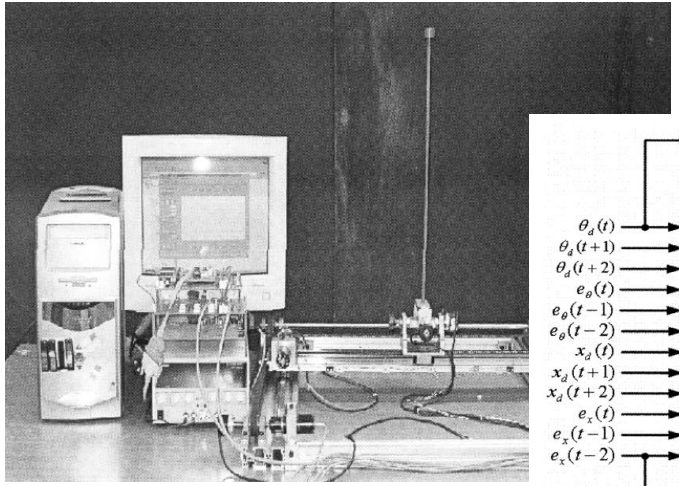
Accuracy results on real-world data streams. (a) Covertype. (b) Pokerhand. (c) Electricity. (d) Christmas. (e) CTR prediction.

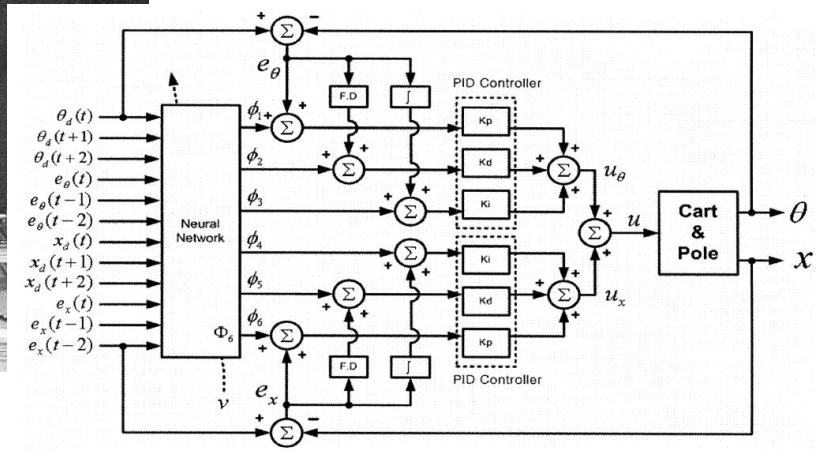
DTEL

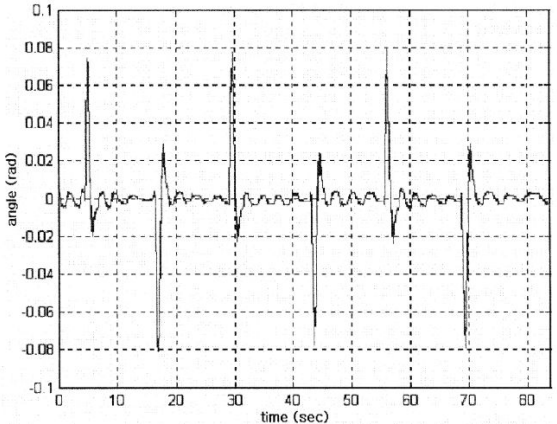
- costly computation

+ can be parallelized

Real-Time Neural Network Controller







Challenges

- Theoretical frameworks for learning.
- Robustness and reliability.
- Moving from so called black box adaptation to more interpretable and explainable adaptation.
- Reducing the dependence on timely and accurate feed-back (true labels).

References

- Ditzler, G., Roveri, M., Alippi, C., & Polikar, R. (2015). Learning in Nonstationary Environments: A Survey. IEEE Computational Intelligence Magazine, 10(4), 12–25.

- Sun, Y., Tang, K., Zhu, Z., & Yao, X. (2018). Concept Drift Adaptation by Exploiting Historical Knowledge. IEEE Transactions on Neural Networks and Learning Systems, 29(10), 4822–4832.

- Wang, F. S., & Lin, C. W. (2013). A Survey on Concept Drift Adaptation.

- Tsymbal, A. (2004). The problem of concept drift: definitions and related work.

- Moulton, R. H., Viktor, H. L., Japkowicz, N., & Gama, J. (2018). Clustering in the Presence of Concept Drift. To Appear in Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.

Thank you for your attention

Filip Paulů