UF UNIVERSITY of FLORIDA

Concept Drift Detection – the State-of-the-Art

Shujian Yu,
Ph.D. Candidate
Department of Electrical and Computer Engineering
yusjlcy9011@ufl.edu

UF UNIVERSITY of FLORIDA

Acknowledgements

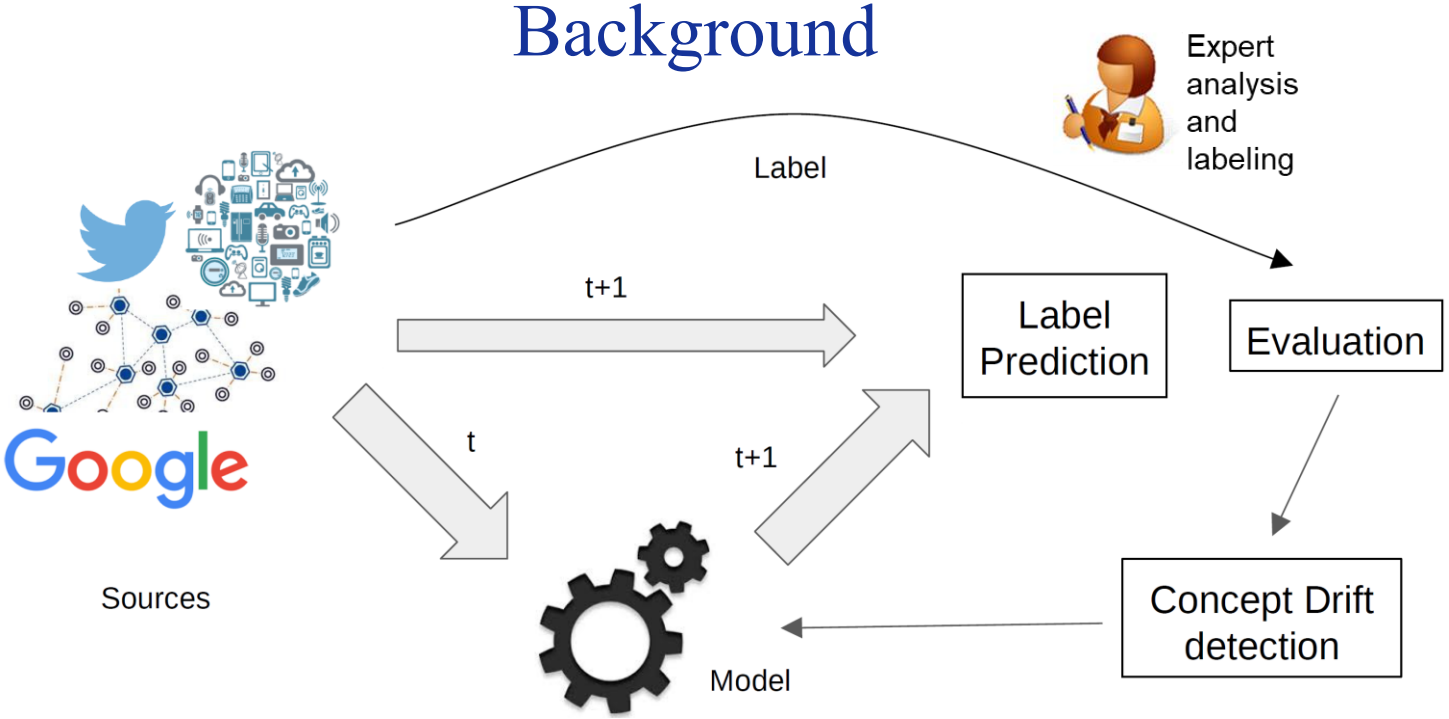
- Joint work with my supervisors/mentors:
  - Dr. Jose C. Principe
    Distinguished Professor at Department of ECE
  - Dr. Zubin Abraham
    Senior Data Mining Research Scientist at Robert Bosch Research Center, CA
  - Dr. Xiaoyang Wang
    Machine Learning Research Scientist at Nokia Bell Labs, NJ
- Some contents were/will be presented in:
  - Bay Area Machine Learning Symposium (2016. 10)
  - SIAM International Conference on Data Mining (2017. 4)
  - Nokia Bell Labs (2017. 9)
  - International Joint Conference on Artificial Intelligence (2018.7)
  - …

UF UNIVERSITY of FLORIDA

Acknowledgements

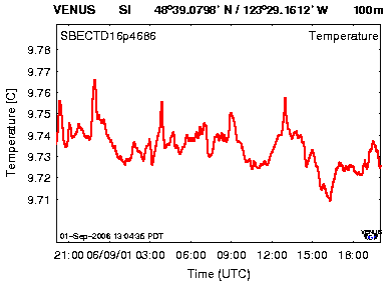
- Related publications
  - Yu, Shujian, and Zubin Abraham. “Concept drift detection with hierarchical hypothesis testing.” In Proceedings of the 2017 SIAM International Conference on Data Mining, pp. 768-776. Society for Industrial and Applied Mathematics, 2017.
  - Yu, Shujian, Xiaoyang Wang, and José C. Prıncipe. “Request-and-Reverify: Hierarchical Hypothesis Testing for Concept Drift Detection with Expensive Labels.” In Proceedings of the 2018 International Joint Conference on Artificial Intelligence, pp. 3033-3039.
  - Yu, Shujian, etc. “Concept drift detection and adaptation with hierarchical hypothesis testing.” To appear in *Journal of The Franklin Institute* (under minor revision).
  - …

Background



Examples of sources

Network traffic



Sensor data

Call center records

UF UNIVERSITY of FLORIDA

Background

- What are the applications?
  - Network monitoring and traffic engineering
  - Business: credit card transaction flows
  - Telecommunication call records
- Challenges?
  - Infinite length
  - Concept drift

Let $\{(\mathbf{X}_t, y_t)\}_{t=1}^{\infty}$ be a labeled time series where $\mathbf{X}_t \in \mathbb{R}^d$ and $y_t \in \mathbb{Z}^1$ (classification) or $y_t \in \mathbb{R}^1$ (prediction). As time moves on, the **Concept Drift** is defined as the scenario when the underlying distribution that generates labeled instances changes over time. Formally, the **Concept Drift** occurs when the joint distribution $p(\mathbf{X}_t, y_t)$ changes.

several years later

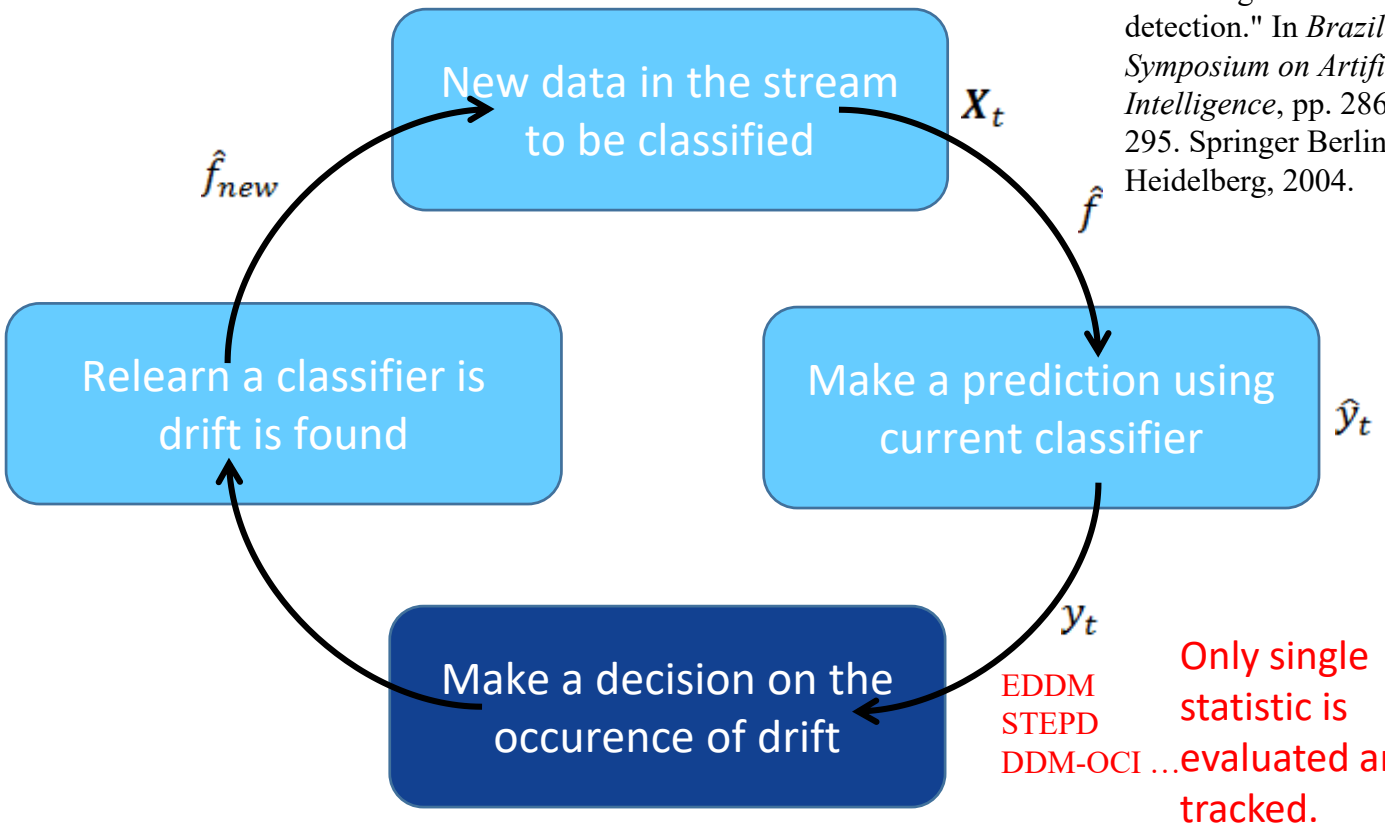
several years later

$\mathbf{X}_t = \begin{pmatrix} \text{Color} \\ \text{Price} \\ \text{Size} \end{pmatrix}$ $y_t = f_1(\mathbf{X}_t)$ $y_t = f_2(\mathbf{X}_t)$ $y_t = f_3(\mathbf{X}_t)$

$y_t = \begin{cases} 1, & \text{like} \\ 0, & \text{dislike} \end{cases}$

Previous works and general framework

- Drift Detection Method (DDM)
  - error monitor + hypothesis testing



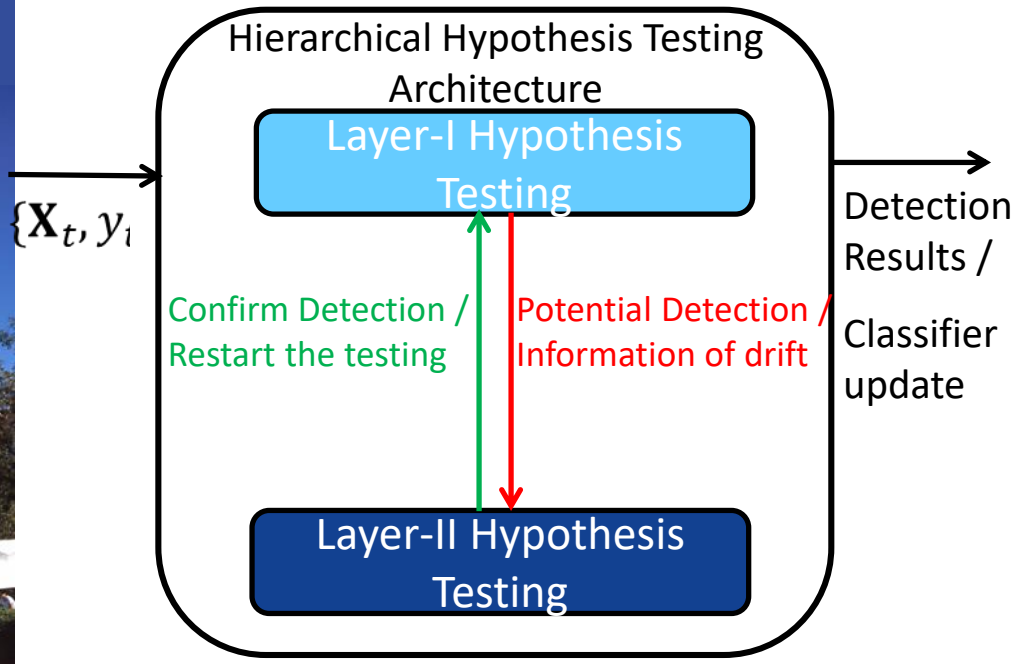
Gama, Joao, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. "Learning with drift detection." In *Brazilian Symposium on Artificial Intelligence*, pp. 286-295. Springer Berlin Heidelberg, 2004.

EDDM
STEPD
DDM-OCI …

Only single statistic is evaluated and tracked.

Hierarchical Hypothesis Testing (HLFR) Framework

- Hierarchical Hypothesis Testing (HHT) framework
  - HHT features two layers of hypothesis test: Layer-I outputs potential drift points, Layer-II reduce false alarms
  - Hierarchical Linear Four Rates (HLFR) is developed under HHT framework



Given
Layer-I test Type-I error α_1,
Layer-I test Type-II error β_1,
Layer-II test Type-I error α_2,
Layer-II test Type-II error β_2,

then
Type-I error of HHT:

$$\alpha = \alpha_1 \alpha_2$$

Type-II error of HHT:

$$\beta = \beta_1 + (1 - \beta_1)\beta_2 \approx \beta_1 + \beta_2$$

UF UNIVERSITY of FLORIDA

Hierarchical Linear Four Rates (HLFR) Algorithm

- Layer-I test: Linear Four Rates (LFR) test

Predict / True	0	1	
0	TN	FN	NPV= TN/(TN+FV)
1	FP	TP	PPV= TP/(FV+TP)
	TNR= TN/(TN+FP)	TPR= TP/(FN+TP)	

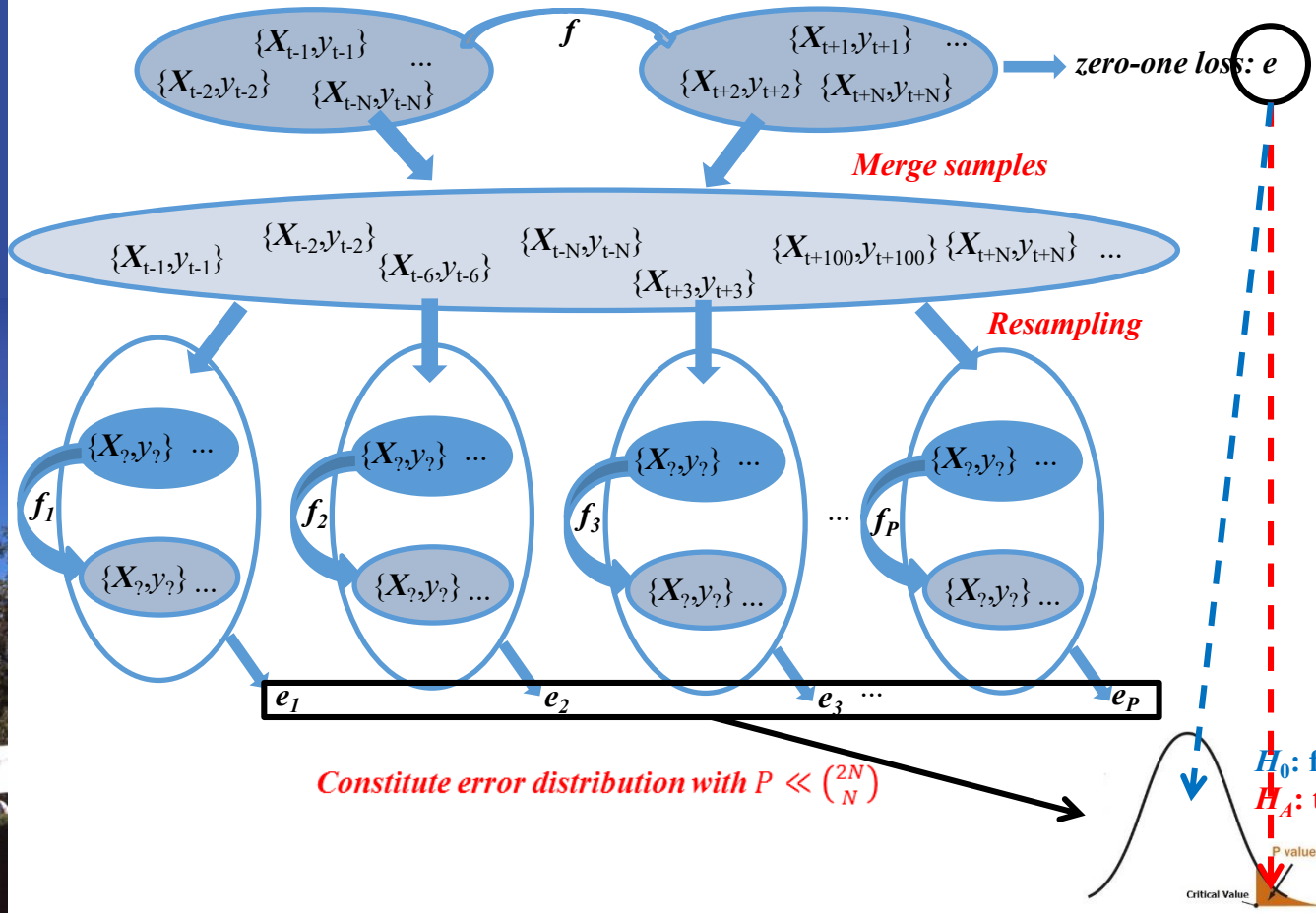
$$H_0: \forall_\star, P\left(\hat{P}_\star^{(t-1)}\right) = P\left(\hat{P}_\star^{(t)}\right)$$
$$H_A: \exists_\star, P\left(\hat{P}_\star^{(t-1)}\right) \neq P\left(\hat{P}_\star^{(t)}\right)$$
$$\star \in \{tpr, tnr, ppv, npv\}$$

geometrically weighted sum of Bernoulli random variables

Monitor four rates (i.e., positive predictive rate, negative predictive rate, true positive rate and true negative rate) associated with the confusion matrix and ALARM loudly if there is any significant change.

Hierarchical Linear Four Rates (HLFR) Algorithm

- Layer-II test: permutation test



UF UNIVERSITY of FLORIDA

Conclusions

- A novel Hierarchical Hypothesis Testing (HHT) framework is developed for concept drift detection.
- Hierarchical Linear Four Rates (HLFR) is designed under HHT framework
- HLFR significantly outperforms benchmark approaches in terms of accuracy, G-mean, recall, delay of detection.

- Perfect? No!
- Let us continue …

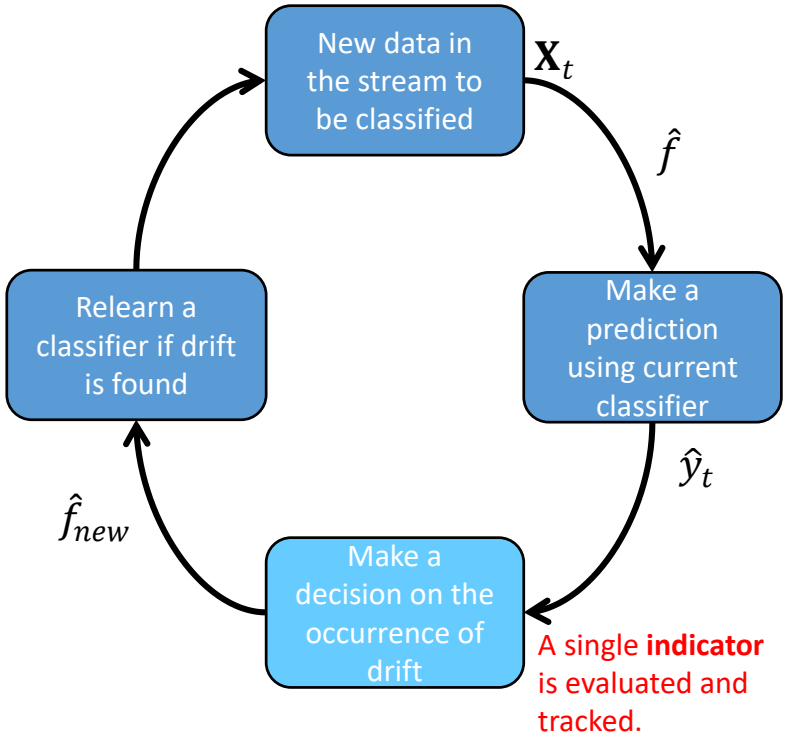
UF UNIVERSITY of FLORIDA

Concept drift detection in the context of expensive labels:

methods and applications

Recall the general framework

- General framework
  - "indicator" monitoring
  + hypothesis test
- State of the art
  - Supervised
  + re-training strategy
    - **HLFR**, STEPD, etc.
  - Unsupervised
  + active training strategy
    - MD3, CDBD, etc.



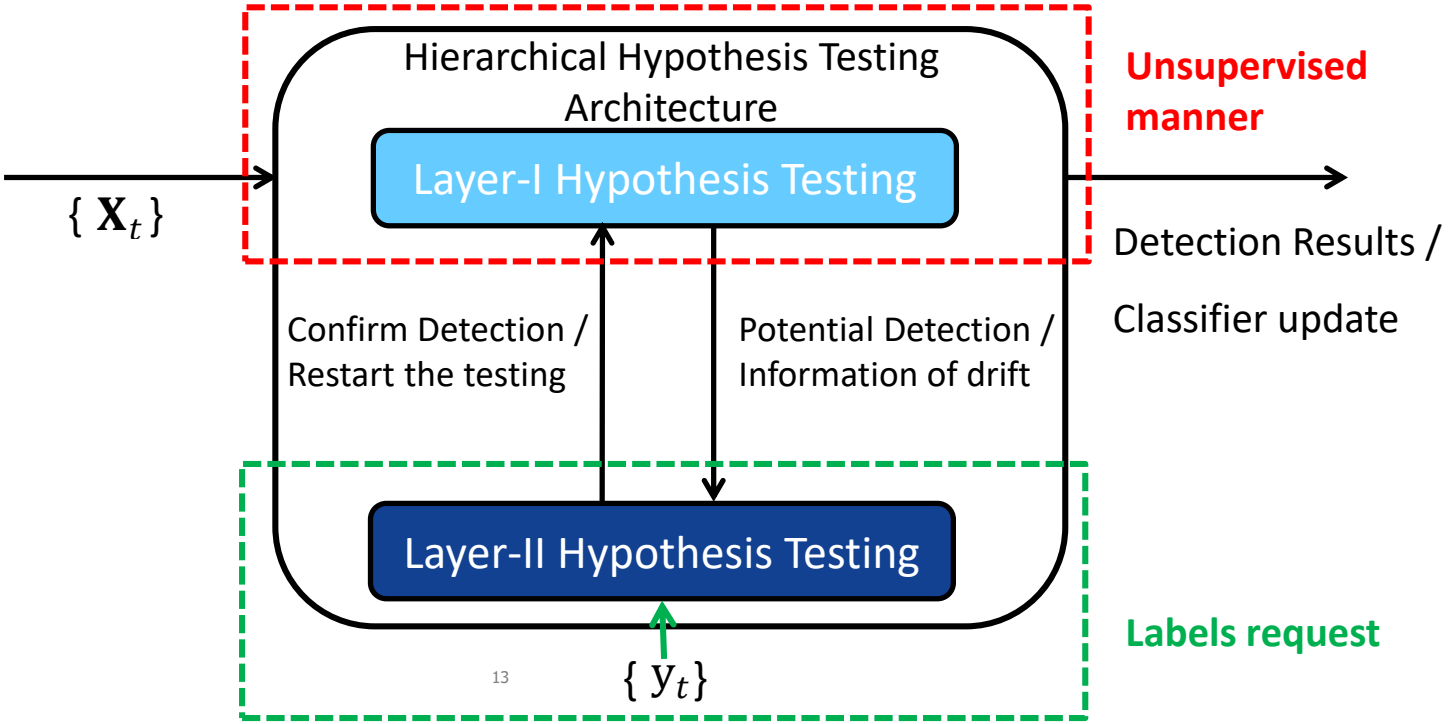
A single **indicator** is evaluated and tracked.

supervised **indicator**: classification error, confusion matrix, etc.
unsupervised **indicator**: margin density, classification score divergence, etc.

- Limitations and motivations
  - Expensive labels --> Accurate detection with minimum labels
  - Multi-class streaming data --> Explicit handle multi-class scenario

Our methods

- A novel Hierarchical Hypothesis Testing (HHT) framework
  - HHT features two layers of hypothesis test: Layer-I outputs potential drift points, Layer-II reduce false alarms



Our methods

- Method I: Hierarchical Hypothesis Testing with classification uncertainty (HHT-CU)
  - Layer-I: Uncertainty measurement + sample mean test
    - Uncertainty u is defined as the $\ell_2$distance between estimated posterior $p(y_t|\mathbf{X}_t)$ probability and estimated class label $\hat{y}_t$, i.e., $u_t = \|\hat{y}_t - p(y_t|\mathbf{X}_t)\|_2$
    - If the classifier is tested in a stationary environments, the estimated sample uncertainty mean will not deviate too much from its previous value given a new sample.

Theorem 1 (Hoeffding's inequality) *Let X_1, X_2,..., X_n be independent random variables such that $X_i \in [a_i, b_i]$, and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then for $\varepsilon \geq 0$:*

$$\mathbb{P}\{\bar{X} - \mathbb{E}(\bar{X}) \geq \varepsilon\} \leq e^{\frac{-2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}. \quad (1)$$

where $\mathbb{E}$ denotes the expectation. Using this theorem, given a specific significance level α, the error ε_α can be computed with:

$$\varepsilon_\alpha = \sqrt{\frac{1}{2n}\ln\frac{1}{\alpha}}. \quad (2)$$

Corollary 1.1 (Layer-I test of HHT-CU) *If X_1, X_2, ..., X_n, X_{n+1}, ..., X_{n+m} be independent random variables with values in the interval $[a, b]$, and if $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Z} = \frac{1}{n+m}\sum_{i=1}^{n+m} X_i$, then for $\varepsilon \geq 0$:*

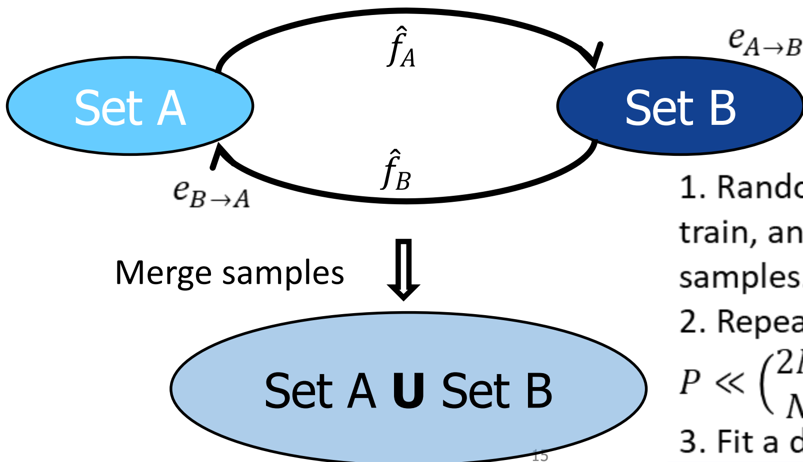
$$\mathbb{P}\{\bar{X} - \bar{Z} - (\mathbb{E}(\bar{X}) - \mathbb{E}(\bar{Z})) \geq \varepsilon\} \leq e^{\frac{-2n(n+m)\varepsilon^2}{m(b-a)^2}}. \quad (3)$$

By definition, $u_t \in [0, \sqrt{\frac{K-1}{K}}]$, where K is the number of classes. $\bar{X}$ denotes the *classification uncertainty* moving average before a cutoff point, and $\bar{Z}$ denotes the moving average over the whole sequence. The rule to reject the null hypothesis $H_0 : \mathbb{E}(\bar{X}) > \mathbb{E}(\bar{Z})$ against the alternative one $H_1 : \mathbb{E}(\bar{X}) \leq \mathbb{E}(\bar{Z})$ at the significance level α will be $\bar{Z} - \bar{X} \geq \varepsilon_\alpha$, where

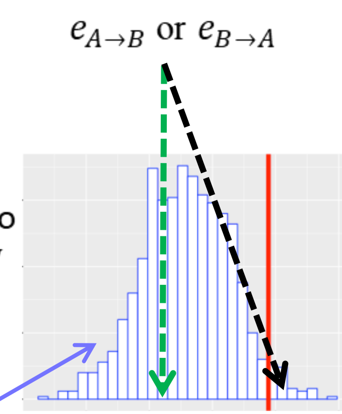
$$\varepsilon_\alpha = \sqrt{\frac{K-1}{K}} \times \sqrt{\frac{m}{2n(n+m)}\ln\frac{1}{\alpha}}. \quad (4)$$

Our methods

- Method I: Hierarchical Hypothesis Testing with classification uncertainty (HHT-CU)
  - Layer-II: Permutation test for potential point T
    - Set A consists of N samples prior to T:
      $A = [(\boldsymbol{X}_{T-N+1}, y_{T-N+1}), (\boldsymbol{X}_{T-N+2}, y_{T-N+2}), \ldots, (\boldsymbol{X}_T, y_T)]$
    - Set B consists of N samples after T:
      $B = [(\boldsymbol{X}_{T+1}, y_{T+1}), (\boldsymbol{X}_{T+2}, y_{T+2}), \ldots, (\boldsymbol{X}_{T+N}, y_{T+N})]$
    - Reject the null hypothesis if $e_{A\to B}$ or $e_{B\to A}$ is above the significant level of estimated error distribution.



1. Random select N samples to train, and test on remaining N samples.
2. Repeat for P times with $P \ll \binom{2N}{N}$
3. Fit a distribution for P errors $e_1, e_2, e_3, \ldots, e_P$.



H0: false decision
HA: true decision

Our methods

- Method II: Hierarchical Hypothesis Testing with Attribute-wise "Goodness of fit" (HHT-AG)
  - Layer-I: Attribute-wise Kolmogorov–Smirnov (KS) test for $p(\mathbf{X}_t^i)|_{i=1}^d$
    - The KS test can be used to test whether two underlying one-dimensional probability distribution differ. In this case, the KS statistics is given by $D_{n,m} = \sup_x |\boldsymbol{F}_{1,n}(x) - \boldsymbol{F}_{2,m}(x)|$, where $\boldsymbol{F}_{1,n}(x)$ and $\boldsymbol{F}_{2,m}(x)$ are the **empirical distribution functions** (estimation of CDF) of the first and second set of examples respectively, and sup is the supremum function.
    - The null hypothesis is rejected at level α if $D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$, with $c(\alpha) = \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)}$

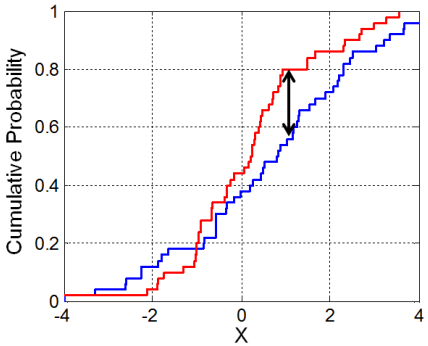


Illustration of the one-dimensional Kolmogorov–Smirnov (KS) statistic. **Red** and **blue** lines each correspond to an empirical distribution function, and the **black** arrow is the two-sample KS statistic.

.

Our methods

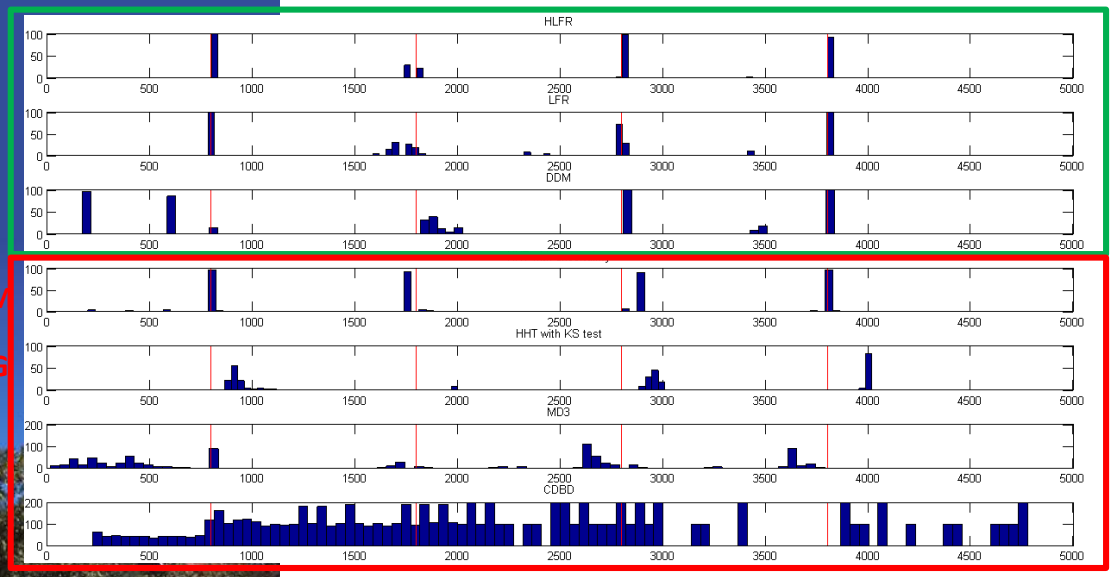
- Method II: Hierarchical Hypothesis Testing with Attribute-wise "Goodness of fit" (HHT-AG)
  - Layer-II: Two-dimensional Attribute-wise Kolmogorov-Smirnov (KS) test for $p(\mathbf{X}_t^i, y_t)|_{i=1}^d$
    - Peacock, 1983 [1] proposed 2D KS test for Astronomy applications.
    - We apply it on our Layer-II test.
    - Peacock's test demands partitioning the n points in 4n2 quadrants and then computing the maximum absolute difference between cumulative distribution functions in all quadrants.

[1] Peacock, J. A. "Two-dimensional goodness-of-fit testing in astronomy."

Monthly Notices of the Royal Astronomical Society, vol. 202, no. 3, pp: 615-627, 1983.

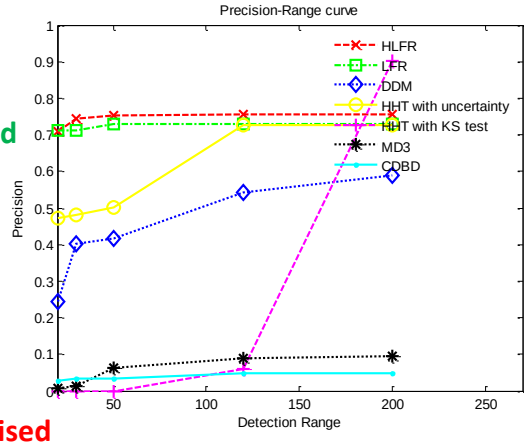
Results

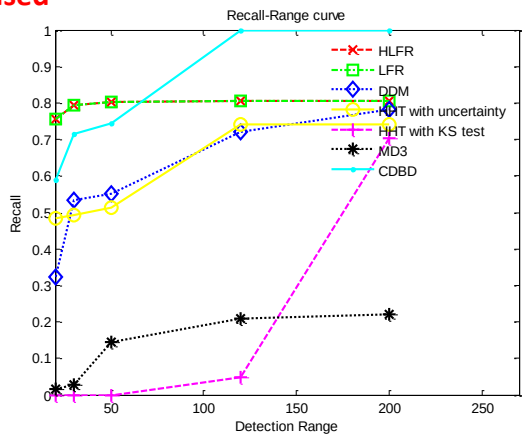
- Public available data
  - UG-2C-2D: Two Bi-dimensional unimodal Gaussian Classes



supervised

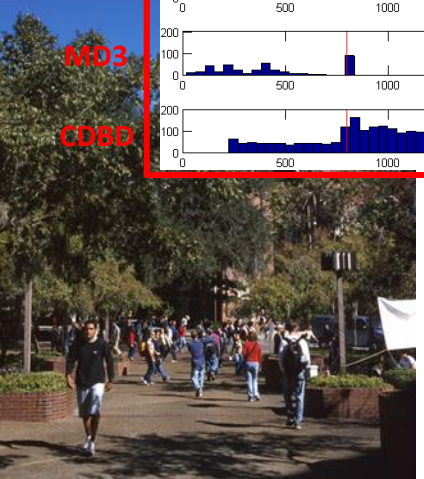
unsupervised





The **red** columns denote the ground truth of drift points, the **blue** columns represent the histogram of detected drift points generated from 100 Monte-Carlo simulations.

Our HHT methods (4th and 5th row) provide consistently superior performance than state-of-the-art unsupervised methods. Besides, it is interesting to find that HHT-UM is even better than the benchmark supervised method.



UF UNIVERSITY of FLORIDA

Real applications

- Analysis of encrypted wireless video stream
  - Mobile video will account for 75% of total mobile data traffic by 2020.
  - Popular video providers such as YouTube and Netflix now encrypt a large part of video content. Trend indicates most video traffic will be encrypted soon.
  - Wireless carriers (e.g. Verizon, AT&T) want to monitor from encrypted video data the Quality of Experience (QoE) on video delivery.
    - Is the video in HD or not?
    - Has the video play ever been frozen (stall) or not? If so, when?
    - How is the client buffer status? (e.g., empty, full, seconds of video in buffer)

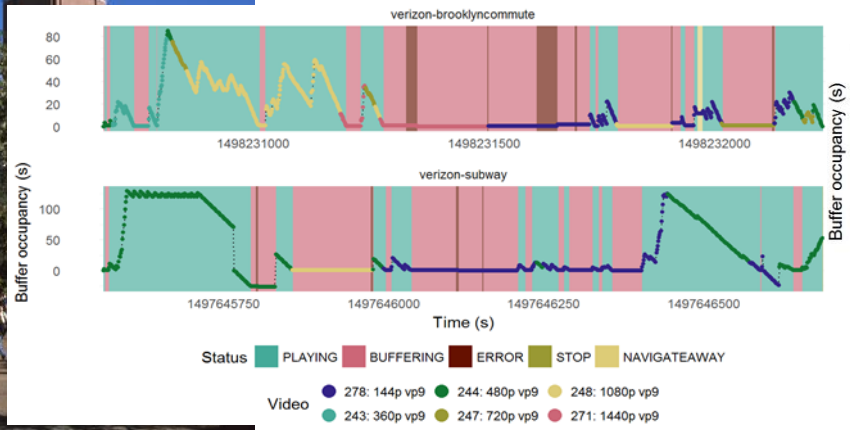
UF UNIVERSITY of FLORIDA

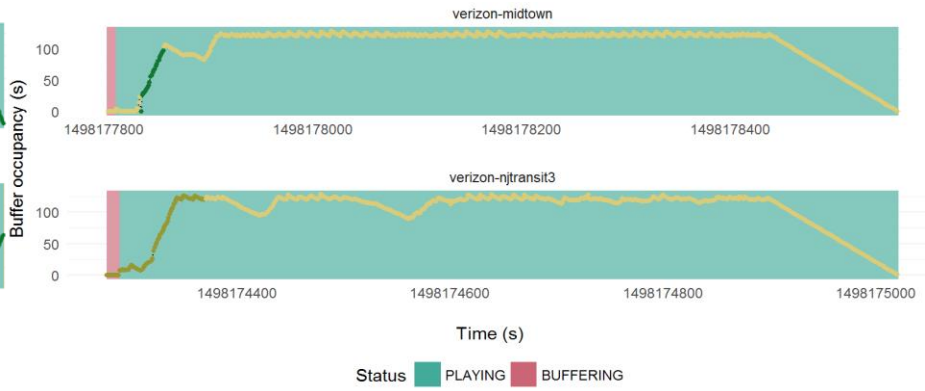
Real applications

- Analysis of encrypted wireless video stream
  - Mobile video will account for 75% of total mobile data traffic by 2020.
  - Popular video providers such as YouTube and Netflix now encrypt a large part of video content. Trend indicates most video traffic will be encrypted soon.
  - Wireless carriers (e.g. Verizon, AT&T) want to monitor from encrypted video data the Quality of Experience (QoE) on video delivery.
    - Is the video in HD or not?
    - Has the video play ever been frozen (stall) or not? If so, when?
    - How is the client buffer status? (e.g., empty, full, seconds of video in buffer)

Real applications

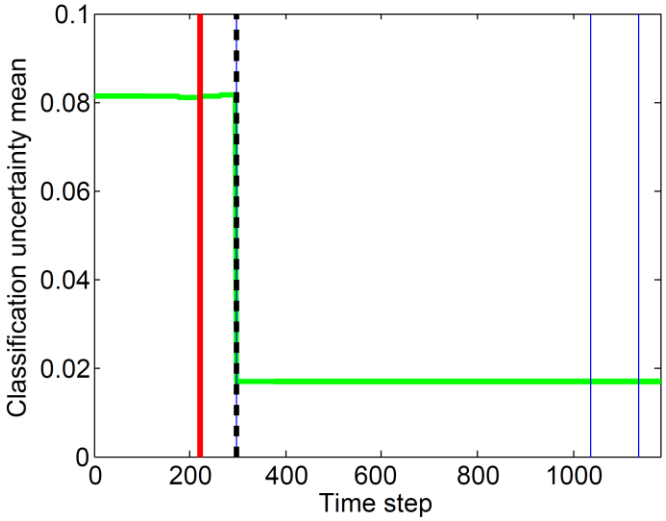
- Analysis of encrypted wireless video stream
  - In collaboration with New York University, Columbia University and Nokia Bell Labs.
  - As the initial step, NYU identified the three buffer status to classify: Filling the Buffer (F) vs. Steady (S) vs. Draining the Buffer (D).
  - However, when the network conditions is compromised, the buffer status could become “ugly”. It brings down the performance of classifiers.

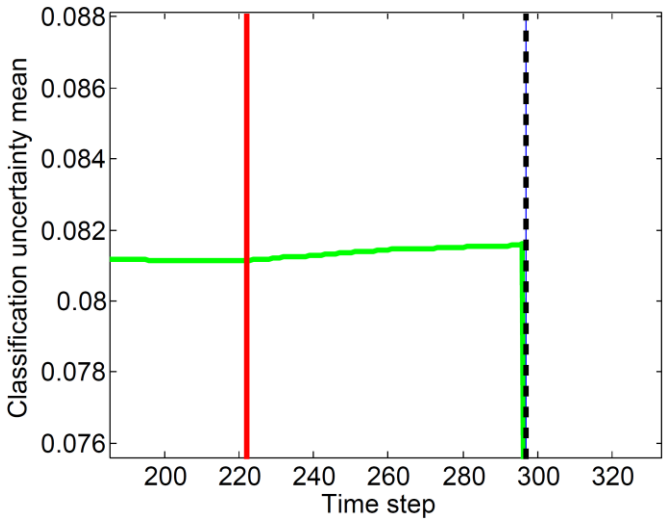




Real applications

- Analysis of encrypted wireless video stream
  - **Concept Drift**: detect the "good" to "congested" drift of network condition, and apply a different classifier for a different network condition.





	Accuracy (%)		
Model	Overall	Steady Stage	Buffering Stage
Unified Model	56.36	70.62	38.87
PCM	59.02	71.36	43.31
Relative Improvement	4.72%	1.04%	11.42%

UF UNIVERSITY of FLORIDA

Future work

- Open toolbox to support various state-of-the-art concept drift detection methods
  - 13 methods in total.
  - Matlab and R
  - 2019 Spring
- Improve Hoeffding's inequality
  - Relax i.i.d. assumption

UF UNIVERSITY of FLORIDA

Thank you!