FACULTY OF
COMPUTER SCIENCE

Knowledge
Management &
Discovery

Predicting and Monitoring Changes in Scoring Data

Edinburgh, 27th of August 2015

Vera Hofer

Dep. Statistics & Operations Res.
University Graz, Austria

Georg Krempl*

Business Information Systems Group
University Magdeburg, Germany
georg.krempl@ovgu.de

Outline

Outline

1. Motivating challenges

Outline

1. Motivating challenges
2. Our approach:
   - Unsupervised adaptation to class prior changes
   - Anticipative classification by density drift extrapolation
   - Monitoring of goodness-of-fit of current & suggested models
   - Controlled adaptation process

Outline

1. Motivating challenges
2. Our approach:
   - Unsupervised adaptation to class prior changes
   - Anticipative classification by density drift extrapolation
   - Monitoring of goodness-of-fit of current & suggested models
   - Controlled adaptation process
3. Preliminary results & lessons learnt

Motivation: Challenges

Stream Classification

- Data arrives continuously
- Classification model is used over long periods of time

Motivation: Challenges

Stream Classification

- Data arrives continuously
- Classification model is used over long periods of time

Non-Stationarity

If distributions change (**drift**) over time:

- Can be sudden/gradual, singular/recurring, systematic/unsystematic
- Static model's performance deteriorates
- Adaptation is required
- Ex-post adaptation to concept drift

Motivation: Challenges

Stream Classification

- Data arrives continuously
- Classification model is used over long periods of time

Non-Stationarity

If distributions change (**drift**) over time:

- Can be sudden/gradual, singular/recurring, systematic/unsystematic
- Static model's performance deteriorates
- Adaptation is required
- Ex-post adaptation to concept drift

Label Paucity of Delay

- Labels arrive with delay (**verification latency**, [Kuncheva, 2008, Marrs et al., 2010])
- Is waiting for sufficient labelled information affordable?

Change Detection

- Well established [Sebastião and Gama, 2009], e.g.
  - Statistical Process Control [Gama et al., 2004]
  - ADaptive WINdowing [Bifet and Gavaldá, 2007]
  - Nonparametric Monitoring [Ross et al., 2011]
- Focus on performance or on location/scale of parameters
- Mostly assume available recent labelled data
  (some ideas on usage of unlabelled data: [Zliobaitė, 2010])
- Aim is detecting change points: focus on sudden shift

Change Detection

- Well established [Sebastião and Gama, 2009], e.g.
  - Statistical Process Control [Gama et al., 2004]
  - ADaptive WINdowing [Bifet and Gavaldá, 2007]
  - Nonparametric Monitoring [Ross et al., 2011]
- Focus on performance or on location/scale of parameters
- Mostly assume available recent labelled data
  (some ideas on usage of unlabelled data: [Zliobaitė, 2010])
- Aim is detecting change points: focus on sudden shift

Change Diagnosis

- Spatio-temporal density estimation
- Velocity Density Estimation [Aggarwal, 2005]
- Ex-post analysis of changes in distributions

Drift Mining

- Understanding how distributions change,
  identifying patterns (drift models [Hofer and Krempl, 2013])
- Extrapolation of trends:
  Anticipative decision trees [Böttcher et al., 2009]
- Estimating new prior from unlabelled data:
  Global drift models [Hofer and Krempl, 2013]

Our Approach in a Nutshell

Contributions in a Nutshell

Approach

- Use unlabelled data for detecting class prior changes:
  Unsupervised prior estimation
- Anticipate changes due to gradual, continuous drift:
  Temporal density extrapolation
- Monitor goodness-of-fit of current & suggested models
  Kullback-Leibler-Divergence calculation on (un)labelled data
- Controlled adaptation process
  Suggest best fitting variant according to available data

Contributions in a Nutshell

Approach

- Use unlabelled data for detecting class prior changes:
  Unsupervised prior estimation
- Anticipate changes due to gradual, continuous drift:
  Temporal density extrapolation
- Monitor goodness-of-fit of current & suggested models
  Kullback-Leibler-Divergence calculation on (un)labelled data
- Controlled adaptation process
  Suggest best fitting variant according to available data

Results

Some preliminary experimental results (with NB Classifier):

- Density extrapolation is (sometimes) useful
- Extrapolation/Prior-Update without monitoring is (sometimes) misleading
- Monitoring KL-Divergence on unlabelled data helps identifying best variant
- Each variable might behave differently

The Building Blocks in Detail

Details: Unsupervised prior estimation

Unsupervised prior estimation

- Use unlabelled data for detecting class prior changes
- Estimate the new class prior $p_1(y = 1) = \delta_1 p_0(y = 1)$ with

$$\delta_1 = \frac{\sum_{x \in \Omega} (f_1(x) - f_0(x|y = 0)) (f_0(x|y = 1) - f_0(x|y = 0)) w(x)}{p_0(y = 1) \sum_{x \in \Omega} (f_0(x|y = 1) - f_0(x|y = 0))^2 w(x)}$$

  where

- previous (old) distribution f_0 and class prior p_0
- current (new) distribution f_1 and class prior p_1
- weighted SSQ criterion $w(x) = f_1(x)$

- For more details, see [Hofer and Krempl, 2013]

Details: Temporal density extrapolation

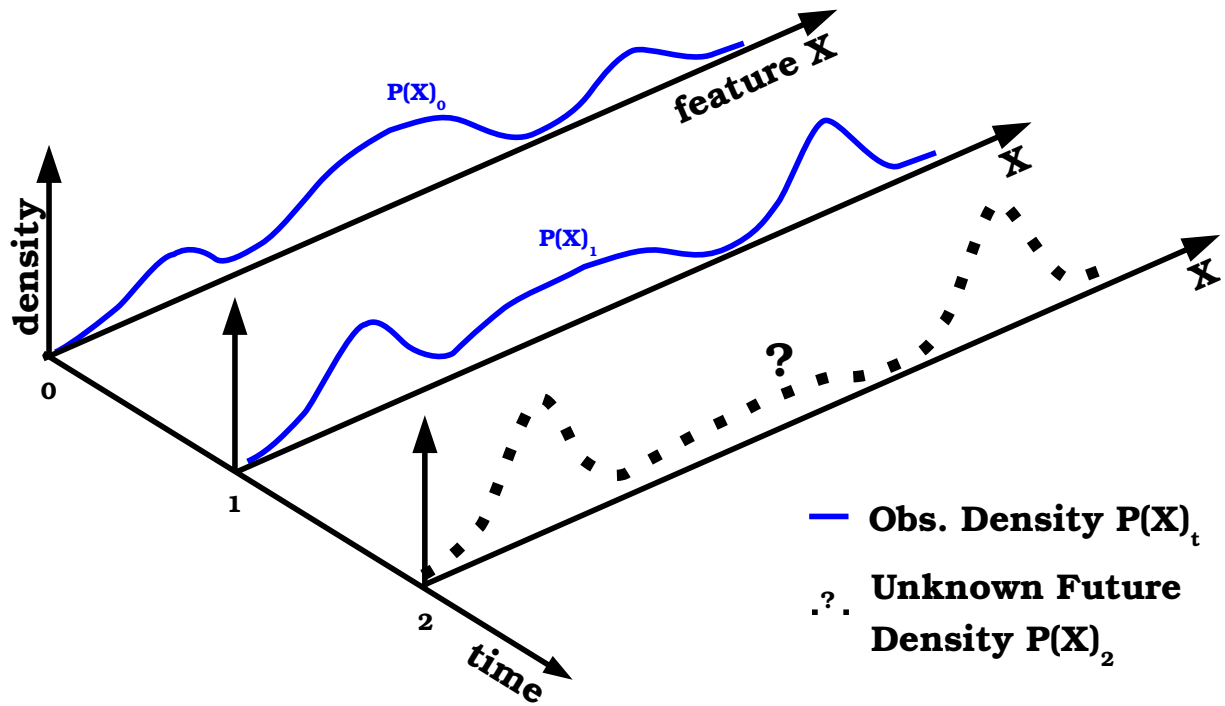
Temporal density extrapolation

- Extrapolate trend patterns in drift
- Kernel-Density-Extrapolation approach:
  - Spatio-temporal kernel density estimation
  - Pseudo-points with extrapolated, time-varying weights
- For more details, see [Krempl, 2015]
- Related: Density forecasting [Tay and Wallis, 2002] (unimodal)

Details: Temporal density extrapolation

Temporal density extrapolation

- Extrapolate trend patterns in drift
- Kernel-Density-Extrapolation approach:
  - Spatio-temporal kernel density estimation
  - Pseudo-points with extrapolated, time-varying weights

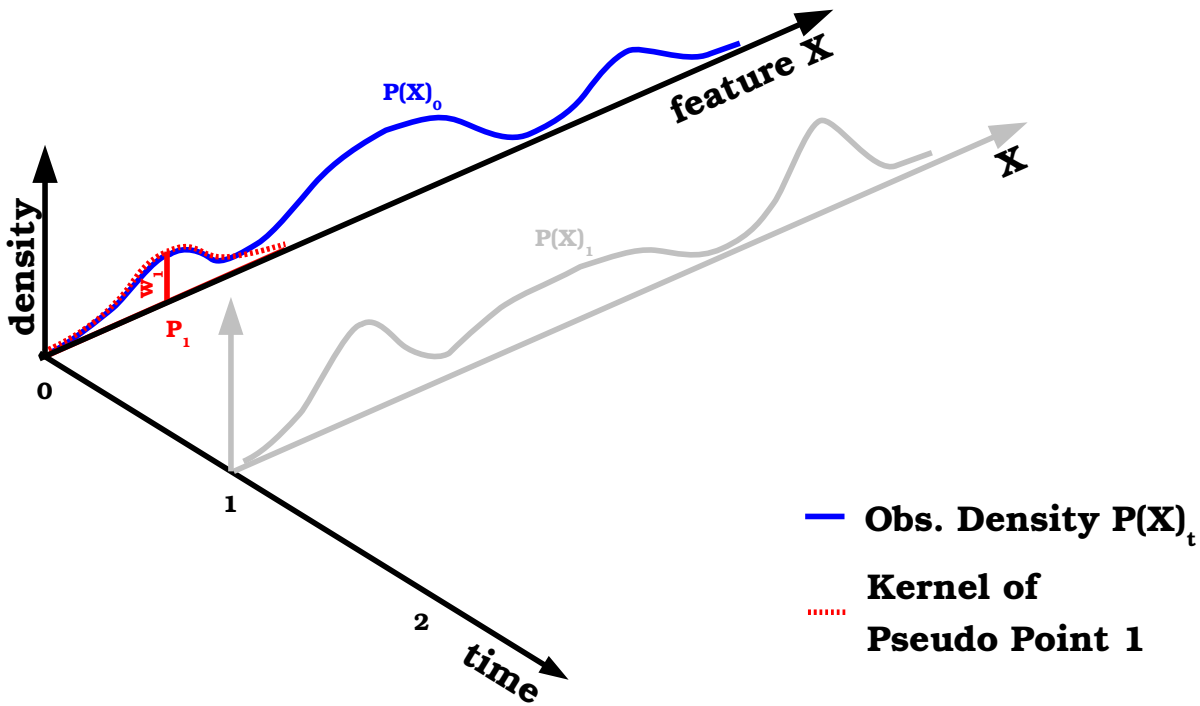


- For more details, see [Krempl, 2015]
- Related: Density forecasting [Tay and Wallis, 2002] (unimodal)

Details: Temporal density extrapolation

Temporal density extrapolation

- Extrapolate trend patterns in drift
- Kernel-Density-Extrapolation approach:
  - Spatio-temporal kernel density estimation
  - Pseudo-points with extrapolated, time-varying weights

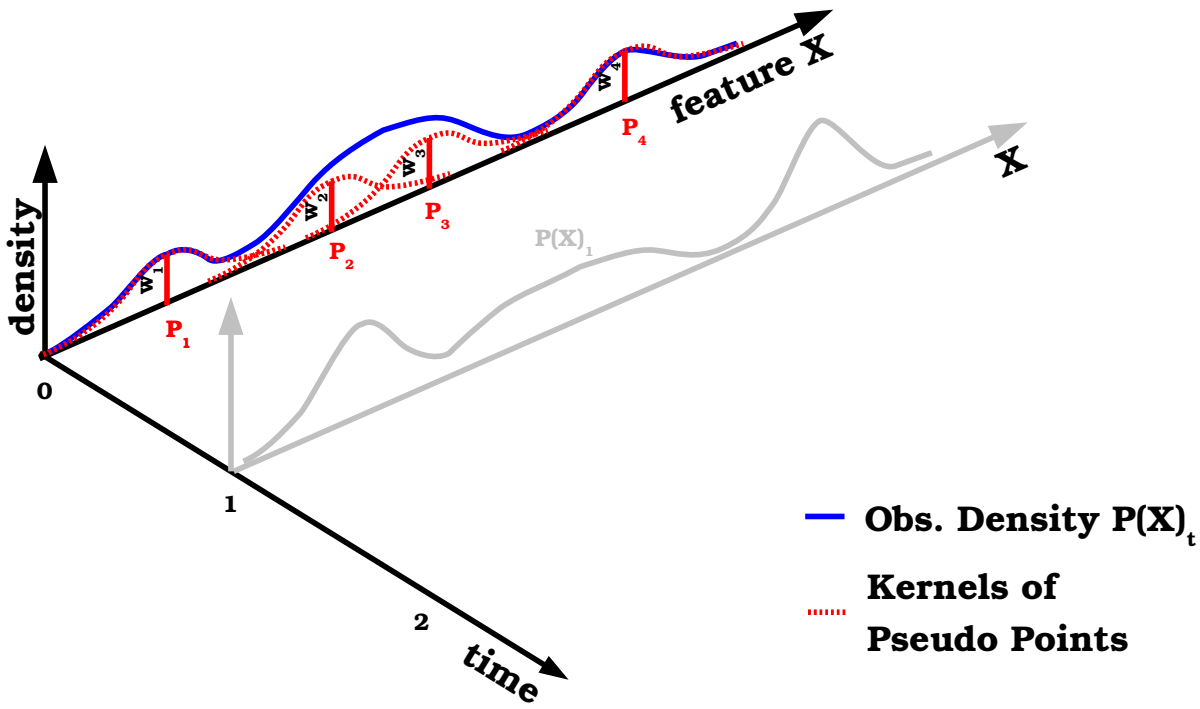


- For more details, see [Krempl, 2015]
- Related: Density forecasting [Tay and Wallis, 2002] (unimodal)

Details: Temporal density extrapolation

Temporal density extrapolation

- Extrapolate trend patterns in drift
- Kernel-Density-Extrapolation approach:
  - Spatio-temporal kernel density estimation
  - Pseudo-points with extrapolated, time-varying weights

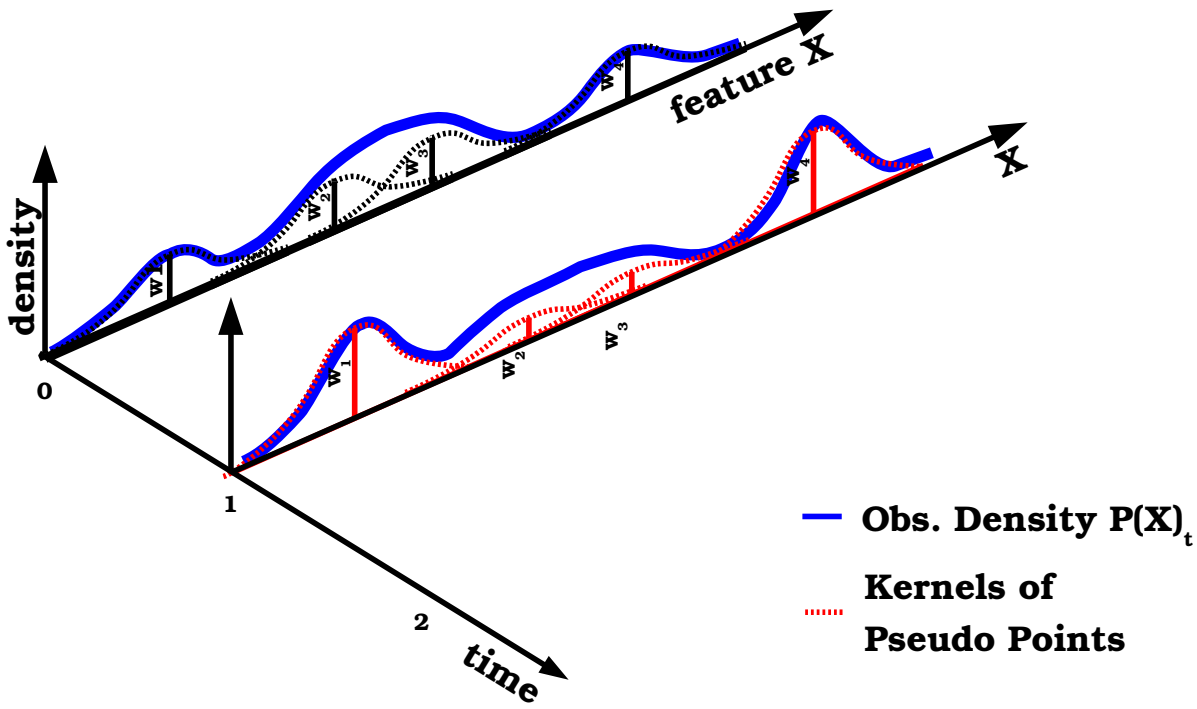


- For more details, see [Krempl, 2015]
- Related: Density forecasting [Tay and Wallis, 2002] (unimodal)

Details: Temporal density extrapolation

Temporal density extrapolation

- Extrapolate trend patterns in drift
- Kernel-Density-Extrapolation approach:
  - Spatio-temporal kernel density estimation
  - Pseudo-points with extrapolated, time-varying weights

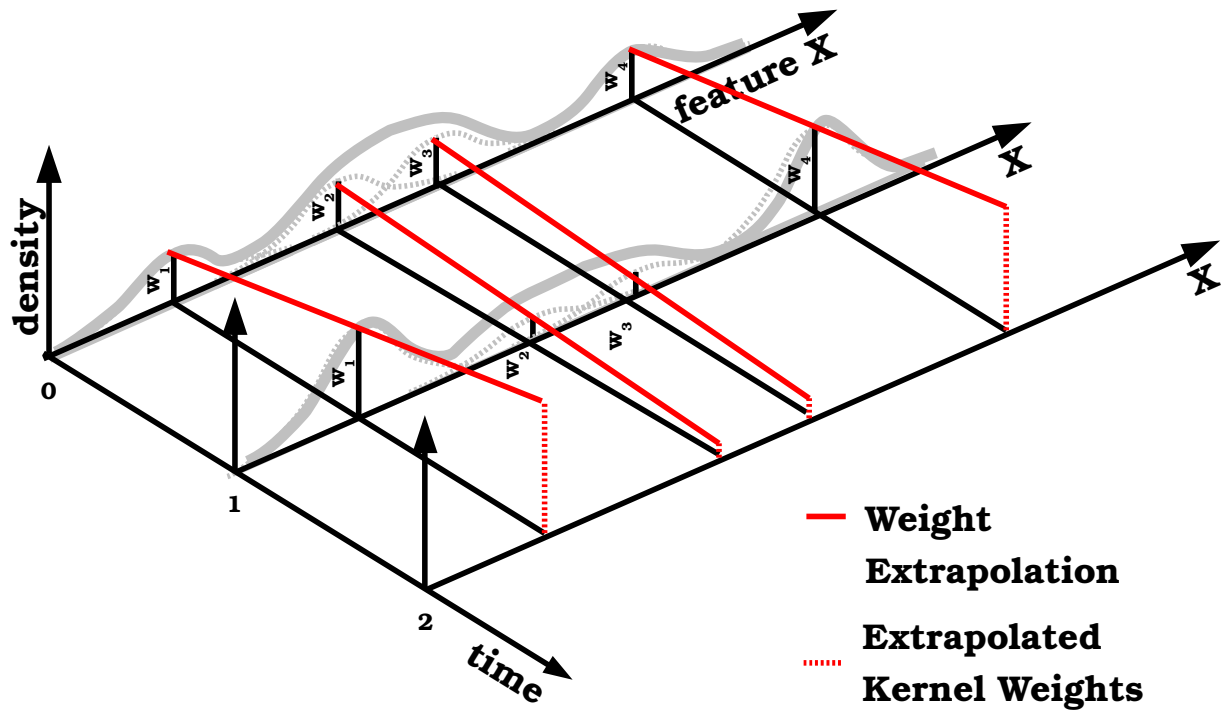


- For more details, see [Krempl, 2015]
- Related: Density forecasting [Tay and Wallis, 2002] (unimodal)

Details: Temporal density extrapolation

Temporal density extrapolation

- Extrapolate trend patterns in drift
- Kernel-Density-Extrapolation approach:
  - Spatio-temporal kernel density estimation
  - Pseudo-points with extrapolated, time-varying weights

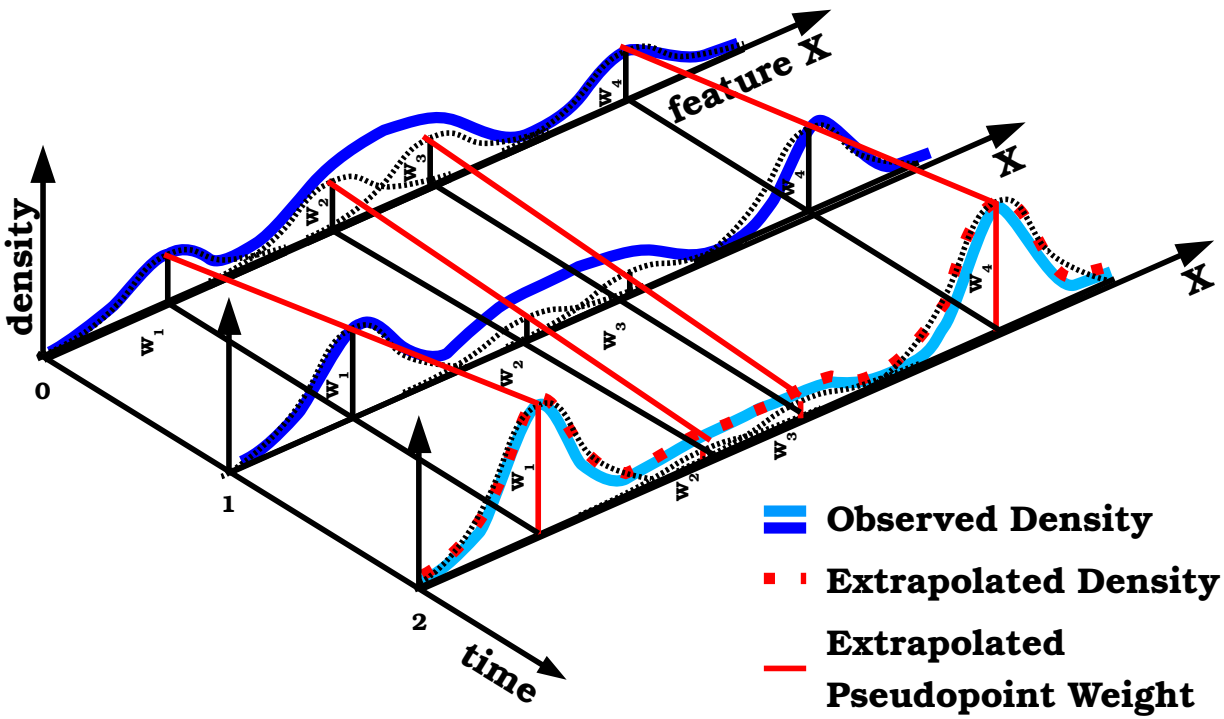


- For more details, see [Krempl, 2015]
- Related: Density forecasting [Tay and Wallis, 2002] (unimodal)

Details: Temporal density extrapolation

Temporal density extrapolation

- Extrapolate trend patterns in drift
- Kernel-Density-Extrapolation approach:
  - Spatio-temporal kernel density estimation
  - Pseudo-points with extrapolated, time-varying weights



- For more details, see [Krempl, 2015]
- Related: Density forecasting [Tay and Wallis, 2002] (unimodal)

Experiment Setup

- Public available dataset: Bank Marketing
  UCI machine learning repository [Asuncion and Newman, 2015]
- Mix of numerical and categorical attributes
- 45211 instances collected over 30 months
- Chunk-based processing (100 chunks, window size 3 chunks)

- Naive Bayes Classifier
- Kernel-Density-Estimation
- Kernel-Density-Extrapolation
- Implemented in Octave/MATLAB

Results (1)

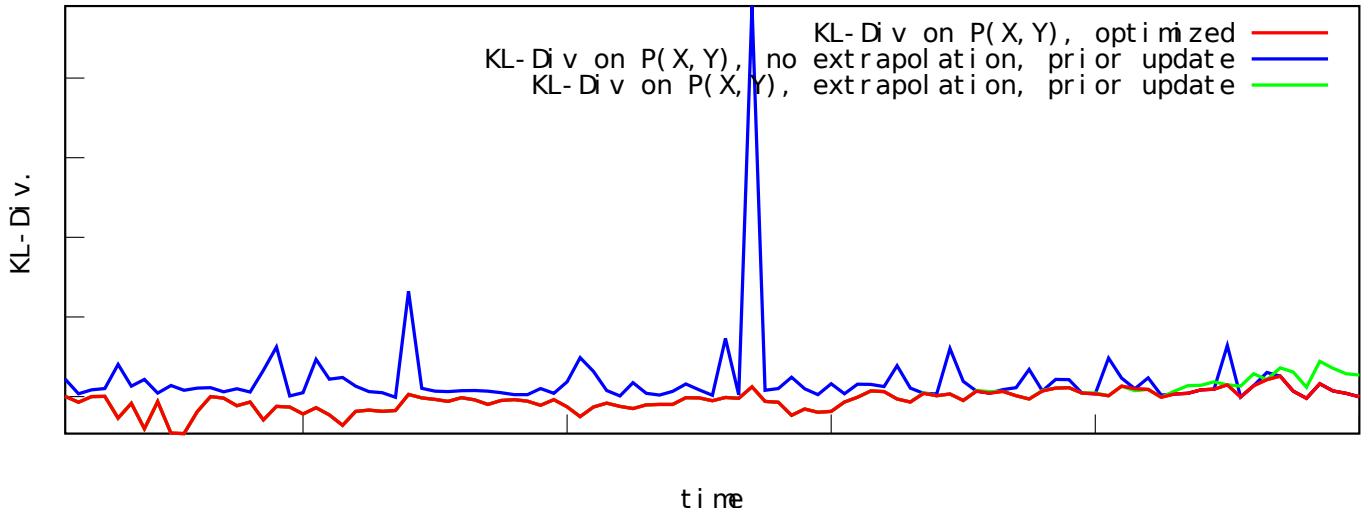
Preliminary Results

- Some improvements over KDE w.r.t. $\hat{f}_1(X, Y)$:

Results (1)

Preliminary Results

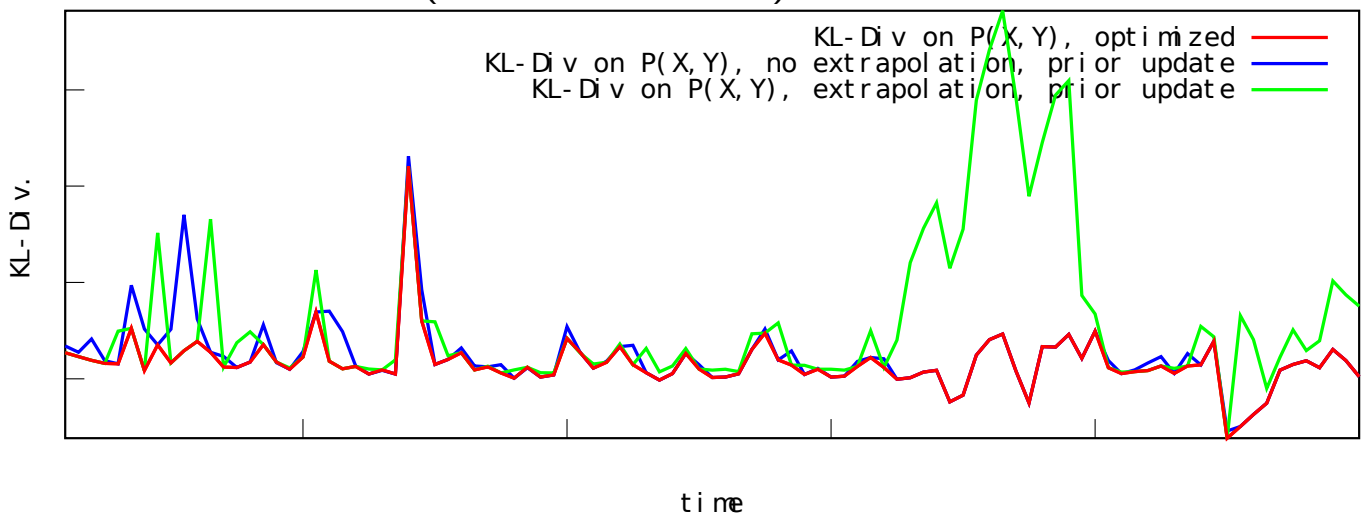
- Some improvements over KDE w.r.t. $\hat{f}_1(X, Y)$:

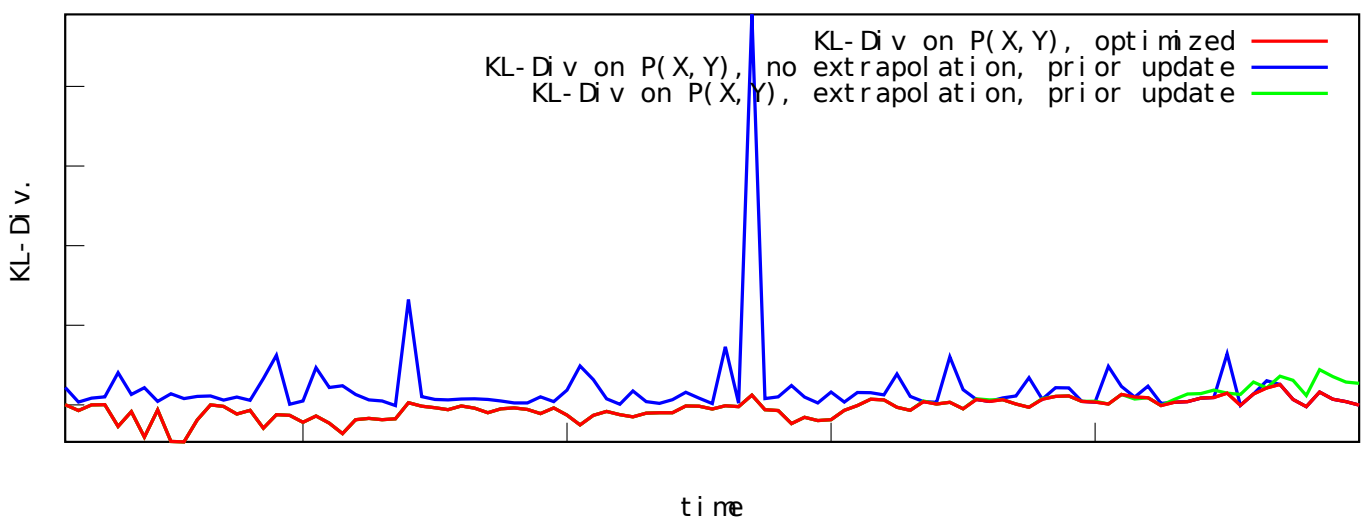


Results (1)

Preliminary Results

- Some improvements over KDE w.r.t. $\hat{f}_1(X, Y)$:
- However, sometimes (on other features) it is not beneficial

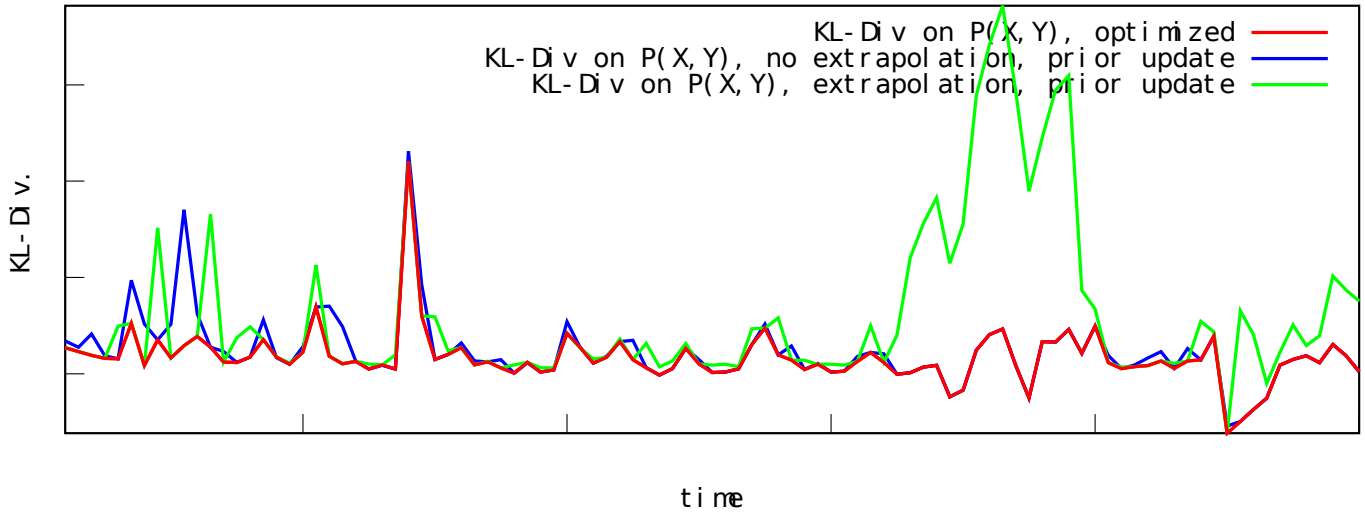


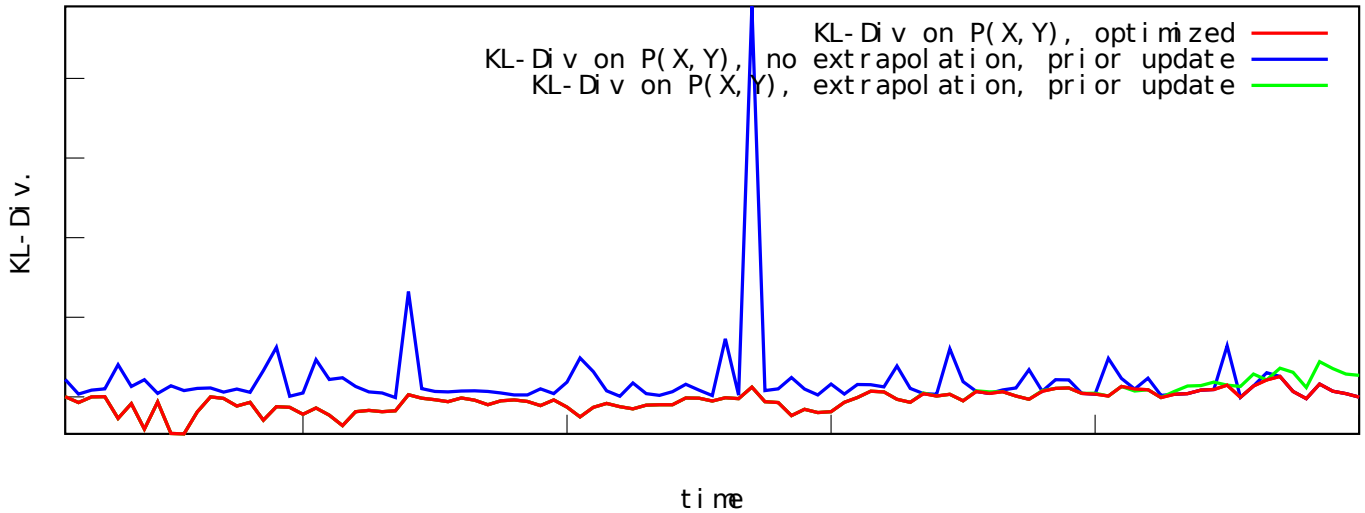


Results (1)

Preliminary Results

- Some improvements over KDE w.r.t. $\hat{f}_1(X, Y)$:
- However, sometimes (on other features) it is not beneficial





- How to avoid misleading extrapolation?

Goodness-of-Fit Monitoring

- Available information:
  - $f_0(X, Y)$ and its derivatives
  - $f_1(X)$
  - estimates for $\hat{f}_1(X, Y)$ and its derivatives

Goodness-of-Fit Monitoring

- Available information:
  - $f_0(X, Y)$ and its derivatives
  - $f_1(X)$
  - estimates for $\hat{f}_1(X, Y)$ and its derivatives
- Estimate $\hat{f}_1(X)$ according to model

Goodness-of-Fit Monitoring

- Available information:
  - $f_0(X, Y)$ and its derivatives
  - $f_1(X)$
  - estimates for $\hat{f}_1(X, Y)$ and its derivatives
- Estimate $\hat{f}_1(X)$ according to model
- Compute its goodness-of-fit compared to $f_1(X)$ using Kullback-Leibler Divergence

Goodness-of-Fit Monitoring

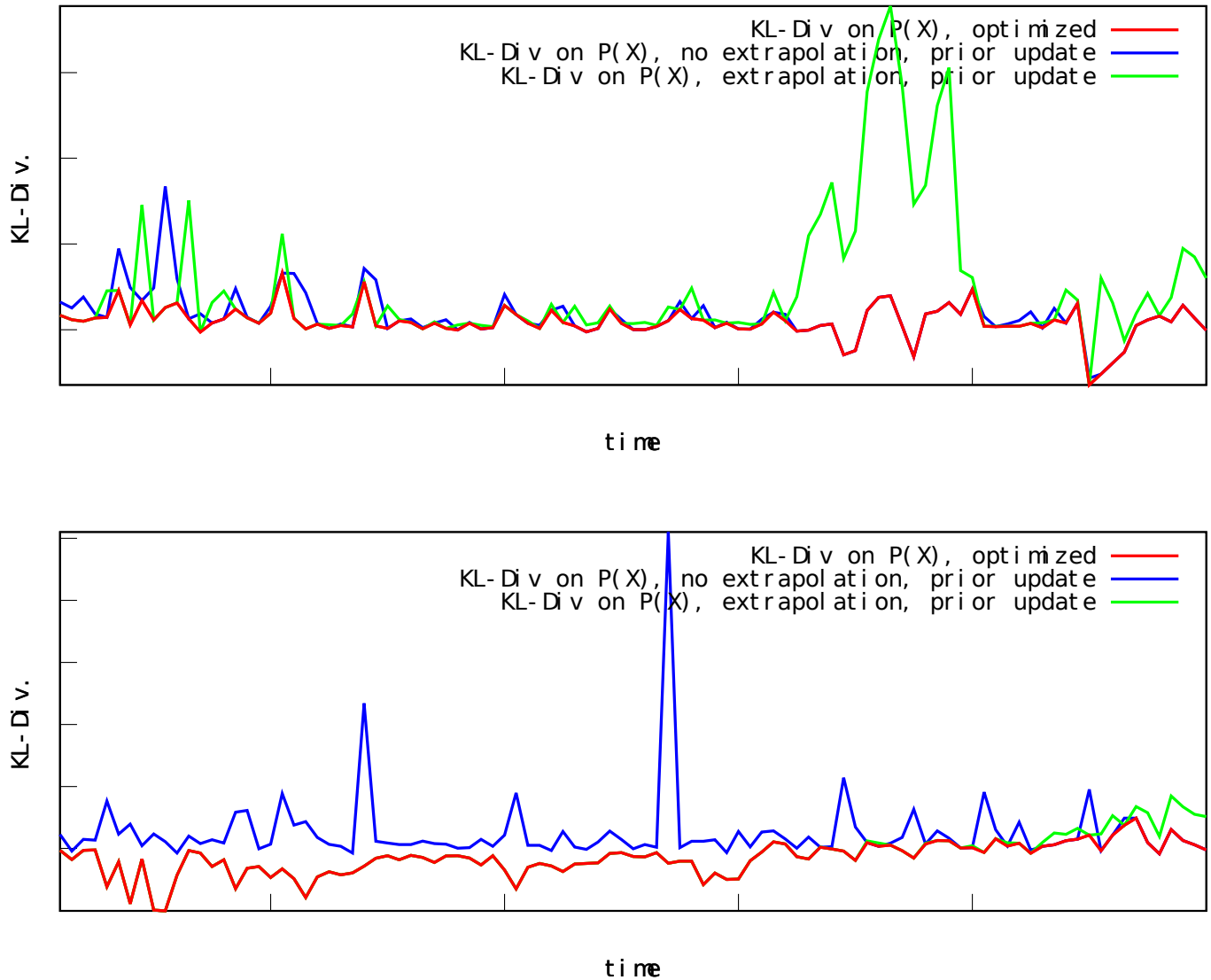
- Available information:
  - $f_0(X, Y)$ and its derivatives
  - $f_1(X)$
  - estimates for $\hat{f}_1(X, Y)$ and its derivatives
- Estimate $\hat{f}_1(X)$ according to model
- Compute its goodness-of-fit compared to $f_1(X)$ using Kullback-Leibler Divergence
- Perform the divergence-minimising update

Goodness-of-Fit Monitoring

- Available information:
  - $f_0(X, Y)$ and its derivatives
  - $f_1(X)$
  - estimates for $\hat{f}_1(X, Y)$ and its derivatives
- Estimate $\hat{f}_1(X)$ according to model
- Compute its goodness-of-fit compared to $f_1(X)$ using Kullback-Leibler Divergence
- Perform the divergence-minimising update
- Is this a suitable indicator?

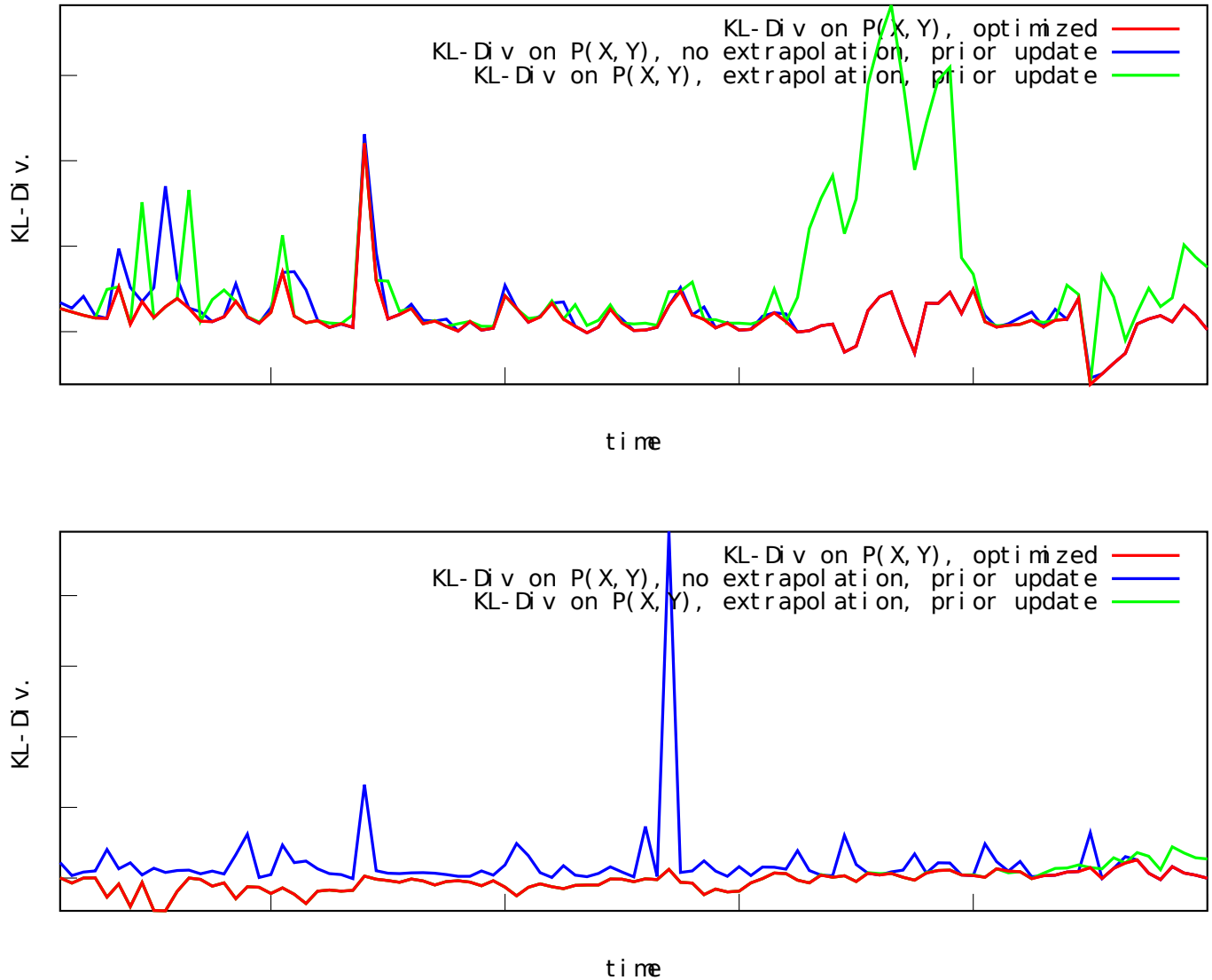
Results (2)

Kullback-Leibler Divergence between $f_1(X)$ and $\hat{f}_1(X)$



Results (3)

Kullback-Leibler Divergence between $f_1(X, Y)$ and $\hat{f}_1(X, Y)$



Summary

Approach Combines:

- Unsupervised prior estimation
- Temporal density extrapolation
- Monitor goodness-of-fit for controlled adaptation

Summary

Approach Combines:

- Unsupervised prior estimation
- Temporal density extrapolation
- Monitor goodness-of-fit for controlled adaptation

Preliminary Results:

- Density extrapolation is (sometimes) useful, but
- Extrapolation/Prior-Update (sometimes) misleading
- Monitoring goodness-of-fit is required
- KL-Divergence on unlabelled data helps identifying best variant
- Each variable might behave differently

Summary

Approach Combines:

- Unsupervised prior estimation
- Temporal density extrapolation
- Monitor goodness-of-fit for controlled adaptation

Preliminary Results:

- Density extrapolation is (sometimes) useful, but
- Extrapolation/Prior-Update (sometimes) misleading
- Monitoring goodness-of-fit is required
- KL-Divergence on unlabelled data helps identifying best variant
- Each variable might behave differently

Outlook / Further Challenges:

- KL-Div considers whole feature space,
  Focus rather on regions around decision boundaries [Vinciotti and Hand, 2003]?
- Modelling seasonal trends (recurring context)
- Prediction-Induced Drift [Krempl et al., 2015]
  Self-Fulfilling / Self-Defeating prophecies

Bibliography I


Aggarwal, C. C. (2005).
On change diagnosis in evolving data streams.
IEEE Transactions on Knowledge and Data Engineering, 17(5):587–600.


Anagnostopoulos, C., Adams, N. M., Hand, D. J., and Leslie, D. (2013).
Handling the risk of obsolete information is there a one-size-fits-all strategy?
In *Credit Scoring and Credit Control XIII*.


Asuncion, A. and Newman, D. J. (2015).
UCI machine learning repository.


Bifet, A. and Gavaldá, R. (2007).
Learning from time-changing data with adaptive windowing.
In *Proceedings of the Seventh SIAM International Conference on Data Mining, SDM 2007, April 26-28, 2007, Minneapolis, Minnesota, USA*. SIAM.


Böttcher, M., Spott, M., and Kruse, R. (2009).
An algorithm for anticipating future decision trees from concept-drifting data.
In Bramer, M., Coenen, F., and Petridis, M., editors, *Research And Development In Intelligent Systems 25*, number 7 in Proceedings of AI-2008, pages 293–306, London. Springer.

Bibliography II


Gama, J., Medas, P., Castillo, G., and Rodríguez, P. (2004).
Learning with drift detection.
In *Advances in Artificial Intelligence, Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA 2004)*, volume 3171 of *Lecture Notes in Computer Science*, pages 286–295.


Hofer, V. and Krempl, G. (2013).
Drift mining in data: A framework for addressing drift in classification.
Computational Statistics and Data Analysis, 57(1):377–391.


Krempl, G. (2015).
Temporal density extrapolation.
In Douzal-Chouakria, A., Vilar, J. A., Marteau, P.-F., Maharaj, A., Alonso, A. M., Otranto, E., and Nicolae, M.-I., editors, *Proc. of the 1st Int. Workshop on Advanced Analytics and Learning on Temporal Data (AALTD) co-located with ECML PKDD 2015*, volume 1425. CEUR Workshop Proceedings.


Krempl, G., Bodnar, D., and Hrubos, A. (2015).
When learning indeed changes the world: Diagnosing prediction-induced drift.
In De Bie, T. and Fromont, E., editors, *Advances in Intelligent Data Analysis XIV - 14th Int. Symposium, IDA 2015, St. Etienne, France*, volume to appear of *Lecture Notes in Computer Science*. Springer.

Bibliography III


Kuncheva, L. I. (2008).
Classifier ensembles for detecting concept change in streaming data: Overview and perspectives.
In Okun, O. and Valentini, G., editors, *Proceedings of the second workshop on supervised and unsupervised ensemble methods and their applications (SUEMA2008)*, volume 245 of *Studies in Computational Intelligence*, pages 5–10. Springer.


Marrs, G., Hickey, R., and Black, M. (2010).
The impact of latency on online classification learning with concept drift.
In Bi, Y. and Williams, M.-A., editors, *Knowledge Science, Engineering and Management*, volume 6291 of *Lecture Notes in Computer Science*, pages 459–469. Springer.


Ross, G. J., Tasoulis, D. K., and Adams, N. M. (2011).
Nonparametric monitoring of data streams for changes in location and scale.
Technometrics, 53(4):379–389.


Sebastião, R. and Gama, J. (2009).
A study on change detection methods.
In *Proceedings of the 4th Portuguese Conf. on Artificial Intelligence, Lisbon*.


Tay, A. S. and Wallis, K. F. (2002).
Density forecasting: A survey.
Companion to Economic Forecasting, pages 45–68.

Bibliography IV


Vinciotti, V. and Hand, D. J. (2003).
Local versus global models for classification problems: Fitting models where it matters.
The American Statistician, 57(2):124–131.


Zliobaitė, I. (2010).
Change with delayed labeling: When is it detectable?
In *IEEE International Conference on Data Mining Workshops (ICDMW 2010)*, pages 843 – 850.