Classification in Presence of Drift and Latency

HaCDAIS Workshop at ICDM2011 / Vancouver, Canada / 2011-12-11

Georg Krempl
Knowledge Management & Discovery
Otto-von-Guericke University Magdeburg
g.krempl@iti.cs.uni-magdeburg.de

Vera Hofer
Statistics & Operations Research
University of Graz
vera.hofer@uni-graz.at

Outline

- Introduction
  - Motivation
  - Population Drift
  - Verification Latency
- Mining Drift: Generative Models
  - Drifting Decision Boundary
  - Drifting Subpopulations
  - Global Prior Drift
- Experiments
- Conclusion

Motivation

Why study & categorize drift?

- All adaptive classification models make *assumptions on the type of drift*
- Alignment of adaptive strategy to drift in reality requires *identification* and *categorization* of drift

Population Drift

Population Drift

Changes in distributions over time: Kelly et al., 1999
Here used synonymously to *concept drift* (Schlimmer and Granger, 1986)

- Static feature space
- Drift can affect:
  - Posterior distribution $P(Y|X)$
  - Feature distribution $P(X)$
  - Class prior distribution $P(Y)$
- Notation:
  - X Explanatory variable(s) (feature(s))
  - Y Binary response (label)

Verification Latency

Example

- Classifier for credit scoring:
  Predict default of loan
- Maturity of loan: 3 years
- Most recent available labelled data:
  November 2008

  Representative for today's applications?

Verification Latency

Time interval between *classification* and *verification of the prediction* (Marrs et al., 2010)

Also denoted as:

- Time lag (Lucas, 2004)
- Label delay (Kuncheva, 2008)

Concurrence of Drift & Latency

Data Availability Problem:

Whenever predicting outcomes far in the future:

- Available labelled data is outdated
- No actual and labelled data is available

- labelled (old) data
- new (unlabelled) data

Idea of Drift Mining:

- Analyse change in historic, labelled data:
  Is drift *systematic* ?
- Identify invariances in change:
  Do *drift patterns* exist, that relate posterior change to
  - the course of time,
  - changes in the feature distribution ?
- Predict current joint and posterior distributions
- Update classifier accordingly

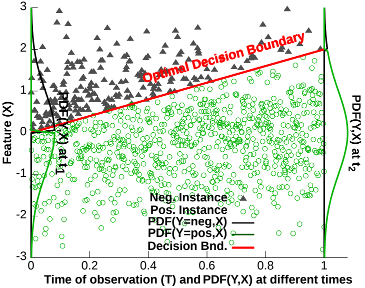
- Drift Models

Drifting Decision Boundary

- Strong relation between X and Y, threshold τ determines class
- Threshold changes over time (cmp. moving hyperplane, Hulten et al.(2001))

- **Drift pattern:**
  Direct relation between posterior drift and *course of time*

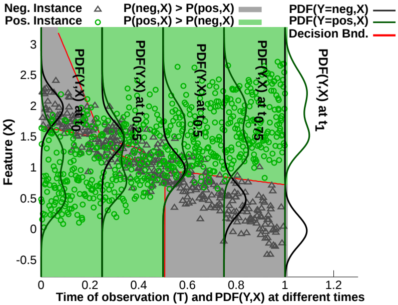
  Approach:
  Learn movement of dec. boundary



Drifting Sub-Populations

- Differently evolving subpopulations
- Clusters evolve gradually
- **Drift Pattern:** Relation between change of $P(Y|X)$ and $P(X)$

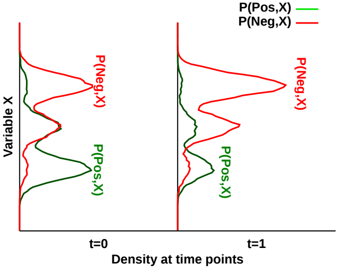
  Approach: Identify & track sub-populations in unlabelled data over time



Global Prior Drift

- Change of class prior
  (over the whole feature space)
- Multiplicative model,
  growth factors δ_p, δ_n
- **Drift pattern:**
  Relation between change
  of $P(Y|X)$ and $P(X)$

  Approach: Estimate δs
  from unlabelled data



Global Prior Change: True vs. Predicted

Can $P(Y)$ be estimated by analysing $P(X)$ **in reality**?

Results on a real-world credit scoring data set:[1]

True Prior Changes δ_p

To / From	2007	2008	2009
2006	1.87	2.81	2.54
2007	—	1.50	1.36
2008	—	—	0.90

Predicted Prior Changes $\hat{\delta}_p$

To / From	2007	2008	2009
2006	1.84	2.82	2.52
2007	—	1.57	1.40
2008	—	—	0.88

Results obtained using a SSE-minimizing estimate of the feature distribution.

[1] In upcoming publication *Mining Drift in Data* (contact me for details).

Conclusion & Outlook

Drift Mining:

- Aim: Identification of *drift patterns* (invariances in the change of distributions)
- Use knowledge of drift as substitute for new, labelled data
- Applicable on *systematic drift*
- Advantageous in presence of *verification latency*
  Always up-to-date classifier

Current & Future Work:

- Application to more real-world data sets (in different application domains)
- Extension of drift models

Thank you for your attention!

Special thanks to Gary Marrs,
Myra Spiliopoulou, Bernhard Nessler
and to the referees!

Contact: Georg Krempl
KMD Workinggroup Magdeburg
g.krempl@iti.cs.uni-magdeburg.de

Bibliography

G. Hulten, L. Spencer, and P. Domingos.
Mining time-changing data streams.
In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 97–106, New York, NY, USA, 2001. ACM.

M. G. Kelly, D. J. Hand, and N. M. Adams.
The impact of changing populations on classifier performance.
In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 367–371, 1999.

G. Krempl.
The algorithm apt to classify in concurrence of latency and drift.
In J. Gama, E. Bradley, and J. Hollmén, editors, *Advances in Intelligent Data Analysis X*, volume 7014 of *Lecture Notes in Computer Science*, pages 222–233. Springer Berlin / Heidelberg, 2011.

G. M. Krempl.
Adaptive Prediction Models and their Application to Credit Scoring.
PhD thesis, University of Graz, 2011.

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

A. Lucas.
Updating scorecards: Removing the mystique.
In L. C. Thomas, D. B. Edelman, and J. N. Crook, editors, *Readings in Credit Scoring*, pages 93–110. Oxford University Press, 2004.

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

J. C. Schlimmer and R. H. Granger.
Beyond incremental processing: Tracking concept drift.
In *AAAI*, pages 502–507, 1986.

Application of Drift Mining

	No Latency	**Latency**	
Systematic Drift	✓	✓	**Drift Mining**
Unsystematic Drift	✓	?	
	Incremental Learning		