Modeling concept drift: A probabilistic graphical model based approach

Hanen Borchani[1], Ana M. Martínez[1], Andrés R. Masegosa[2], Helge Langseth[2], Thomas D. Nielsen[1], Antonio Salmerón[3], Antonio Fernández[4], Anders L. Madsen[1,5] & Ramón Sáez[4]

[1] Department of Computer Science, Aalborg University, Denmark

[2] Department of Computer and Information Science, NTNU, Norway

[3] Department of Mathematics, University of Almería, Spain

[4] Banco de Crédito Cooperativo, Spain

[5] Hugin Expert A/S, Aalborg, Denmark

Introduction

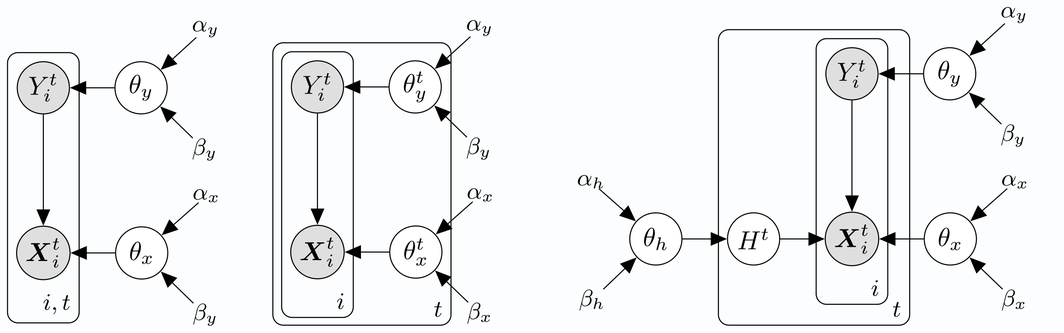
- Classification in a streaming context amounts to observing objects at different points in time $t = t_1, t_2, \ldots$, and at each time-point t classifying the object based on the information collected up to and including time t, $\bigcup_{j:t_j \leq t} \boldsymbol{x}_{t_j}$.
- Main challenges of data stream classification:
  - the streaming data may not be independent and identically distributed.
  - concept drift: the underlying distribution generating the data changes over time.
  - computational problems due to the unbounded high velocity data.

The financial data set

- Provided by Banco de Crédito Cooperativo (BCC).
- Consists of monthly aggregated information captured by 11 quantitative attributes, denoted X_i^t, for $50\,000$ BCC clients for the period from April 2007 to March 2014.
- Each client i has an associated class variable Y_i^t for each time step t, which indicates if that particular client will default during the following 12 months.

Modeling concept drift using latent variables

- We propose a new framework, based on probabilistic graphical models [1], that explicitly represents and detects concept drift using latent variables.
- The modeling technique addresses the general situation, where we, at each time point t, have a collection $(\boldsymbol{x}_i^t, y_i^t)$, for $i = 1 : N_t$, of instances.



(a) No concept drift (b) Concept drift (c) Modeling concept drift with a latent variable

Figure 1: Modeling concept drift through parameter duplication. In all figures, $\alpha_{(\cdot)}$ and $\beta_{(\cdot)}$ are hyper-parameters for the distributions over the parameters θ_x and θ_y

- The concept drift can also be captured across time through the dependence relations among the latent H^t variables.

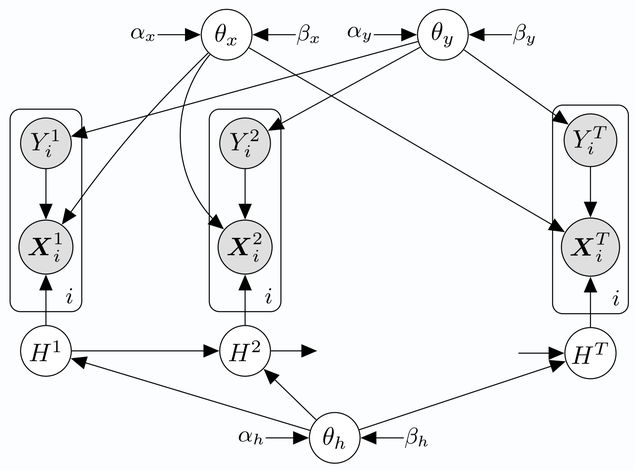


Figure 2: Concept drift is preserved over time

Bayesian inference with streaming data

- Bayesian paradigm: model learning can be considered as an inference task.
- In a streaming context, a new data sample $(\boldsymbol{x}_{t+1}, y_{t+1})$ in the stream is included by simply updating the posterior distribution using Bayes' rule.
- For the models considered in this work, we rely on the variational Bayes framework for doing approximate inference and learning.
- The calculations have been structured efficiently in conjugate exponential models using a variational message passing scheme [3].

Results

- All the experiments have been performed using MOA [2], where the AMIDST model (in Fig. 2), over $\mathbf{X}_i^t$ and Y_i^t, has been integrated as a new naive Bayes streaming classifier, named *bayes.amidstModels*. The Java code to reproduce the experiments can be downloaded from `http://amidst.github.io/toolbox/`.

Synthetic data sets

- For the SEA data set, the expected value of latent H^t variable detects the drift points and clearly identifies the occurrences of the four different phases in the data.
- For the hyperplane data sets, the different drift magnitudes (i.e., 0.1, 0.5 and 1) are directly reflected in the development trends of the latent H^t variables.

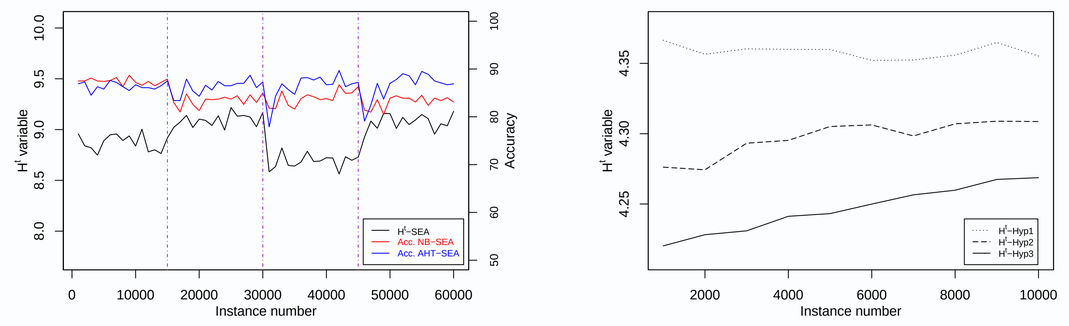


Figure 3: Left: Results for the SEA data set. Right: Results for the hyperplane data sets

Financial data set

- The evolution of the H^t variable over time captures the concept drift and reflects the seasonal effect in the financial data set.

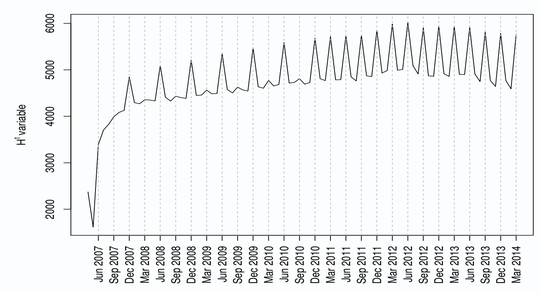


Figure 4: Evolution of the latent H^t variable for the financial data set

- The global trend of the latent H^t variable also illustrates the economic climate in Almería during the studied period, and a close correlation is identified between the unemployment rate and the expected value of the H^t variable.

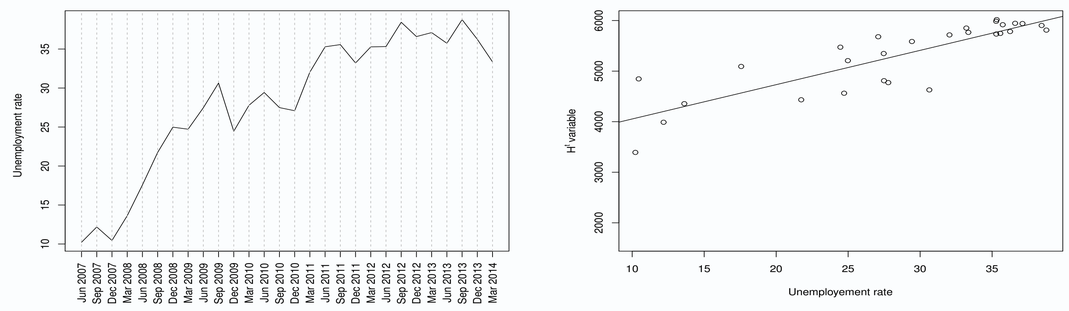


Figure 5: Left: Unemployment rate in Almería. Right: Scatter plot of the unemployment rate and the expected value of the latent H^t variable (Spearman's rank correlation coefficient is 0.85)

Conclusion

- Our proposed approach distinguishes itself from traditional alternatives by explicitly including the effect of the concept drift in the model using latent variables.
- In the future, we intend to consider more sophisticated base classifiers and extend the concept drift modelling itself, e.g., by using more than one latent variable.

References

[1] Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs. Springer Verlag, Berlin, Germany (2007).
[2] Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B.: MOA: Massive Online Analysis. Journal of Machine Learning Research 11 (2010) 1601–1604.
[3] Winn, J.M., Bishop, C.M.: Variational message passing. Journal of Machine Learning Research 6 (2005) 661–694.

AMIDST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 619209.