Imperial College London

Temporally adaptive classification in consumer banking applications

Niall M. Adams[1], Christoforos Anagnostopoulos[2], Dimitris K. Tasoulis[1], Nicos Pavlidis[2], David J. Hand[1,2]

[1]Department of Mathematics, [2]Institute for Mathematical Sciences
Imperial College London

August 2009

Contents

- Classification, population drift and consumer finance applications (credit scoring for classifying UPLs, transaction fraud)
- Conventional solution to population drift
- Extreme case: streaming data
- Temporally adaptive linear classifiers for streaming data
- UPL example

Work supported by EPSRC/BAE (ALADDIN project) and EPSRC (ThinkCrime project)

Classification methodology usually deployed on the assumption that future test data is drawn from the same distribution as training data.

In many real problems, especially those involving humans and money, this assumption may be untenable. Modes of violation collectively referred to as population drift (and sometimes concept drift).

Classification examples from consumer finance subject to population drift

- loan application classification: changes in underlying economic conditions
- plastic card transaction fraud: arms race

Decision theory

The optimal decision theory of classification builds on Bayes formula

$$Pr(C|X) = \frac{p(X|C)Pr(C)}{p(X)}$$

where X is a feature vector, and $C \in C_1, \ldots, C_K$ is a class indicator. Diagnostic paradigm goes for $Pr(C|X)$, sampling paradigm for $p(X|C)$.

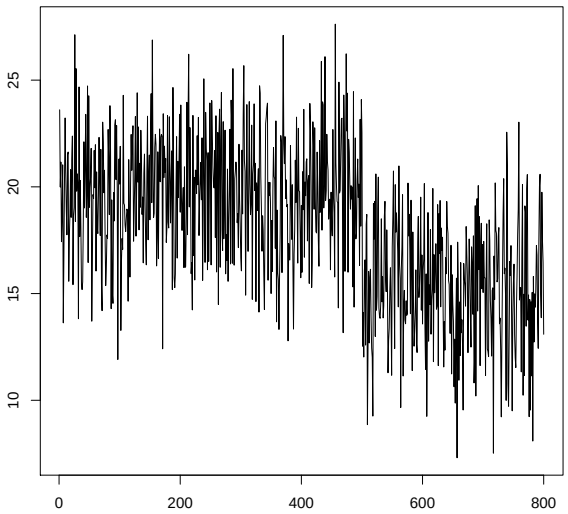
Population drift can happen in the class prior probability, $Pr(C)$, or the class conditional density, $p(X|C)$, or both.
Examples:

- prior: economic conditions reduce size of good risk population of loan applicants
- class conditional: fraud population features move closer to legitimate population

Modes

Often useful to distinguish two simple modes of drift:

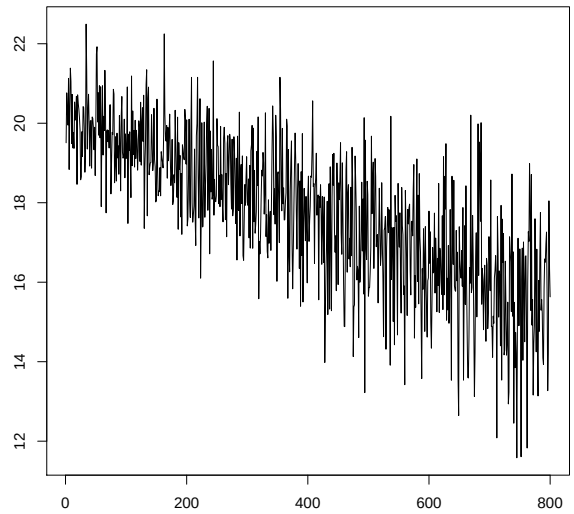
1. Jump



(in mean)

Will give UPL examples later.

2. Gradual change

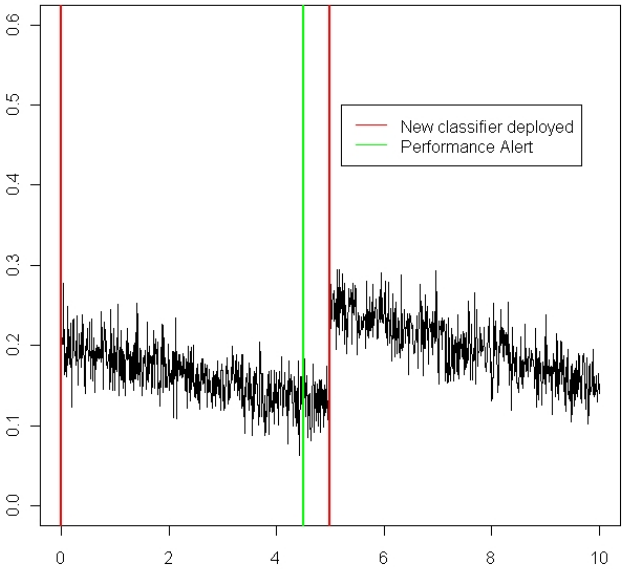


(in mean and variance) Trend, seasonality etc.

Window solution

Population drift usually handled with a window-oriented approach based on a reasonable assumption: recent data is more representative than older data. How to formulate - contiguous windows, sliding windows? fixed or variable size?

For classifying loan applications and plastic card transaction fraud detection: training set fixed size, classifier deployed until appropriate performance metric suggests degradation. Example:



Other aspects

Aspects related to monitoring classifiers

- What to monitor
  - classification/scoring performance
  - model parameters (eg. Whittaker et al (2007), JORS)
- What metric (issue as above)
- defining change (formal approaches)
  - defining alert threshold - eased by large data sets?
  - incremental control chart construction (aside: Axel's talk)
  - change in the presence of drift - EWMA?

Streaming data classification

We are interested in an extreme type of classification problem, associated with streaming data.

Streaming data is characterised by

- continual sequence of data items
- dynamics of sequence unknown and subject to unpredictable change
- computational efficiency critical - update/decide continually

Finance applications, with the character of streaming data

- plastic card fraud detection (adversarial scenario)
- automatic FOREX trading

We have developed methodology for streaming classification. Want to deploy this methodology in the context of loan and fraud classification to see if performance degradation can be reduced between UPL classifier rebuilds.

Note however that we make some simplifying assumptions about order and synchronisation of data.

Will begin with a description of temporally adaptive parametric density estimation – most illustrative – then sketch logistic regression.

Forgetting factor approach

An approach to equip parametric density estimation with adaptive forgetting (Anagnostopoulos et al. 2009). Borrowing ideas from adaptive filtering (eg. Haykin, 1997)

Consider computing the mean vector and covariance matrix of a sequence of n multivariate vectors. Standard recursion

$$m_t = m_{t-1} + x_t,\ \hat{\mu}_t = m_t/t,\ m_0 = 0$$
$$S_t = S_{t-1} + (x_t - \hat{\mu}_t)(x_t - \hat{\mu}_t)^T,\ \hat{\Sigma}_t = S_t/t,\ S_0 = \mathbf{0}$$

After n steps equivalent to offline result. For vectors coming from a non-stationary system, simple averaging of this type is biased.

Knowing precise dynamics of the system gives chance to construct optimal filter. However, not possible with streaming data (and financial data).

Simple solution: estimate the quantities of interest in a window – but potentially difficult to optimise.

Alternatively, use ideas from *adaptive filter* theory, and incorporate a *forgetting factor*, $\lambda \in (0, 1]$, in the previous recursion

$$
\begin{aligned}
n_t &= \lambda n_{t-1} + 1, \; n_0 = 0 \\
m_t &= \lambda m_{t-1} + x_t, \; \hat{\mu}_t = m_t / n_t \\
S_t &= \lambda S_{t-1} + (x_t - \hat{\mu}_t)(x_t - \hat{\mu}_t)^T, \; \hat{\Sigma}_t = S_t / n_t
\end{aligned}
$$

λ down-weights old information more smoothly than a window.

n_t is the *effective sample size* or *memory*. $\lambda = 1$ gives offline solutions, and $n_t = t$. For fixed $\lambda < 1$ memory size tends to $1/(1 - \lambda)$ from below.

Setting λ

Two choices for λ, fixed value, or *variable forgetting*, λ_t. Fixed forgetting: set by trial and error (cf. window).

Variable forgetting: ideas from from Haykin (1997) suggest tuning λ_t according to a local stochastic gradient descent rule

$$\lambda_t = \lambda_{t-1} - \alpha \frac{\partial \xi_t^2}{\partial \lambda}, \quad \xi_t\text{: residual error at time } t, \ \alpha \text{ small}$$

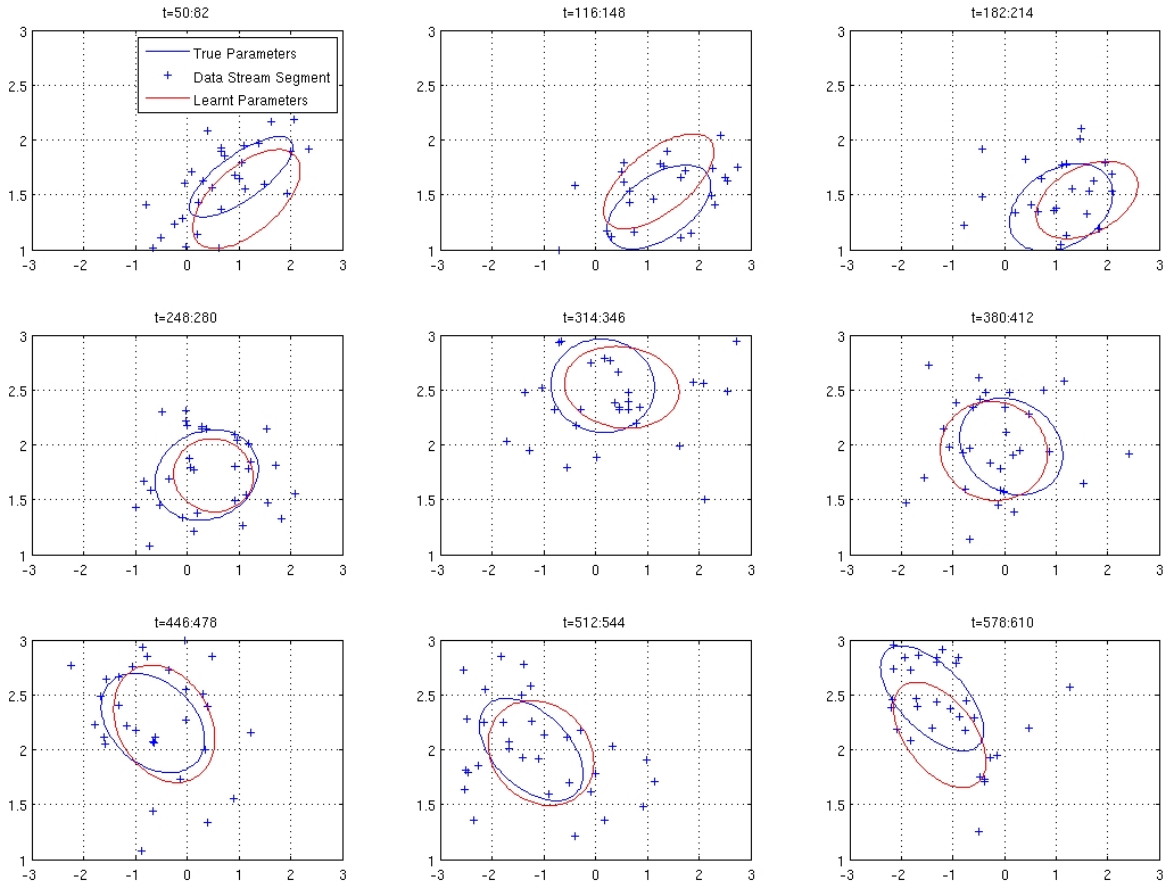
Efficient updating rules implemented via results from numerical linear algebra ($O(d^2)$).

Performance very sensitive to α. Very careful implementation required.

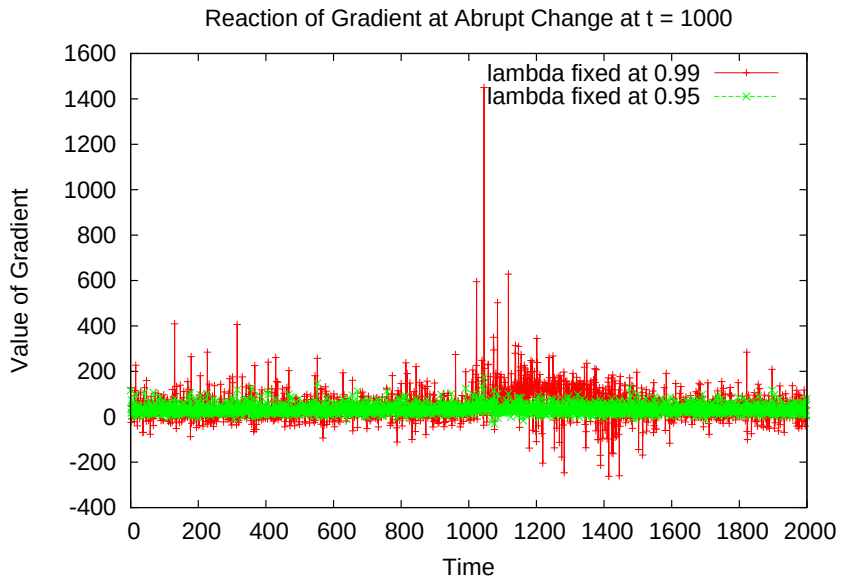
Framework provides a principle for balancing old and new data, which extends to multiple point updates.

Illustrations

Tracking mean and covariance in 2d



Fixed values of λ, 5D stream, abrupt change



Streaming linear Gaussian classifier

Recall

$$Pr(C|X) = \frac{p(X|C)Pr(C)}{p(X)}$$

Linear/quadratic discriminant analysis (LDA/QDA) can be motivated by reasoning that $p(X|C) \sim N(\mu, \Sigma)$:

- LDA: assume covariance matrix common across classes - two options for implementation
- QDA: different covariances

Idea: Use proposed method to adaptively and incrementally estimate mean vectors and covariances matrices.

As noted already, some formulation issues

- regular-spaced data
- timing of label

Replace static estimates of μ and Σ *for each class* with the adaptive estimates. Update class parameters separately, then combine.

Similarly, handle class priors adaptively – proposed framework provides adaptive updating mechanism for prior (multinomial).

Full development of QDA in Anagnostopoulos et al (2009). Good results.

LDA can be motivated by assuming common covariance across K classes: $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_K = \Sigma$. Then, estimate *pooled covariance* as weighted sum.

Various formulations of streaming LDA, including

- *K+1 forgetting factor model*. One FF for each class parameters, one for multinomial prior. Compute pooled covariance, weighted using streaming prior estimate.
- *K+2 forgetting factor model* : one FF for each mean vector (fixed identity covariance) , one for pooled streaming covariance matrix (fixed zero mean vector - add residual), one for prior.

Behaves sensibly with RPROP like implementation of stochastic gradient descent scheme, but problems with prior for K large.

Logistic regression

Deploy similar idea to make implement adaptive logistic regression, by adding forgetting factor to objective function

$$\mathcal{E}(\beta, t) = \sum_{i=1}^{t} \lambda^{t-i} E(\beta, i) = E(\beta, t) + \lambda \mathcal{E}(\beta, t-1)$$

where β are regression coefficients.

Again, can select fixed or variable forgetting.

For the latter, as before, tune in the direction of the derivative of the objective function. This calculation can again be implemented efficiently (though care required).

Opens door for MLP. Aside: Relation between non-linear model and temporal adaption?

Classification for UPLs

Two class UPL problem: *good* or *bad* risk.

For UPLs, legislative requirement to justify decision to reject. Simple models used, usually logistic regression. LDA known to be competitive.

Population drift known to happen - changes in economy etc.

Frequent deployment: build classifier on historic labeled data – predict – rebuild classifier when some measure of classification performance degrades.

Data

Using real UPL data from UK bank, 1993-1998. See Kelly et al. (1999).

20 features, typical of application: age, income, job category, bank account indicators, county court judgments etc. Mixture of continuous and categorical data. Categories converted to indicators.

Class label: bad defined as customer ever ≥ 3 months in arrears.

10 % of data labelled as bad risk

Note: label arrives a long time after application.

Evidence for population drift?

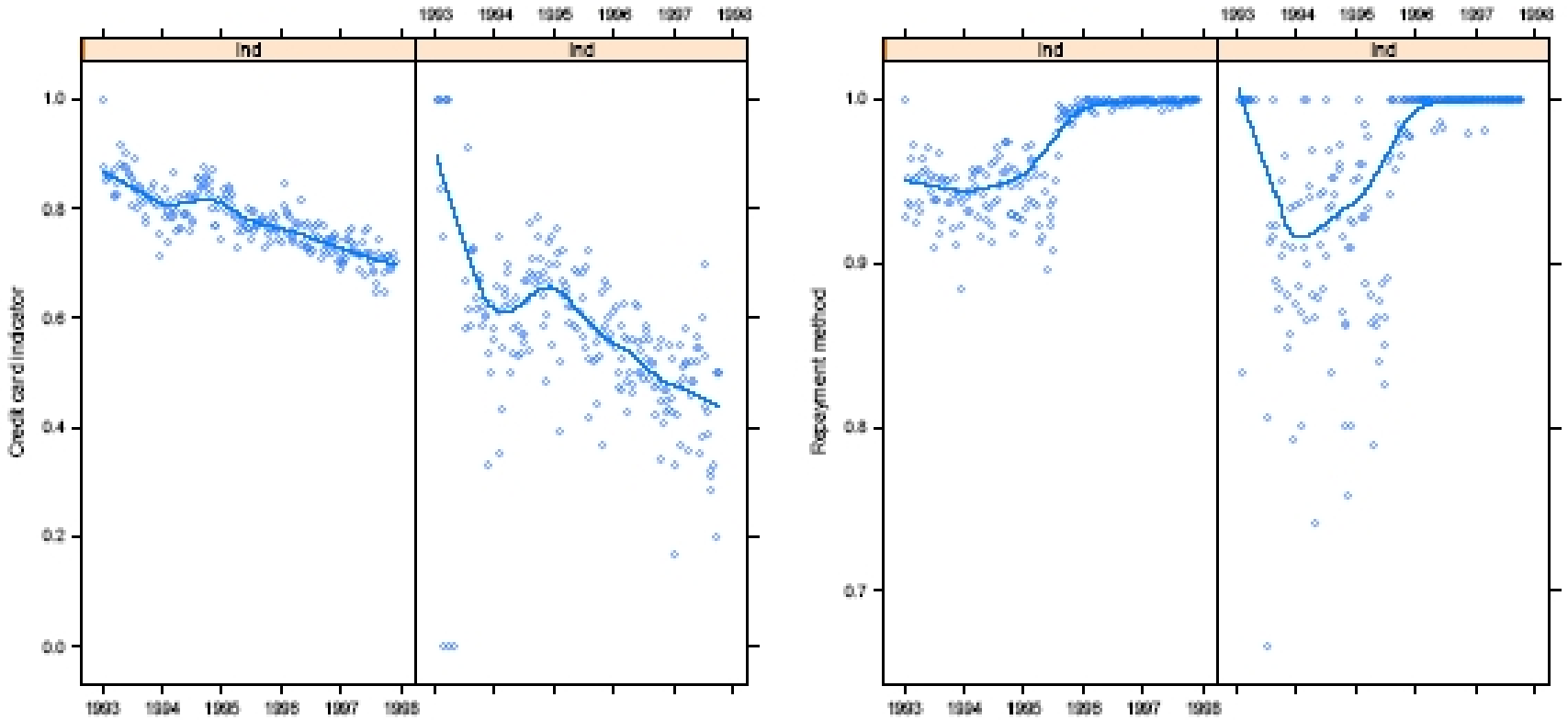


Figure 1: (a): Weekly averages for credit card indicator. (b): Weekly averages for repayment method indicator. Each plot includes good risk (left) and bad risk (right).

Evidence for population drift?

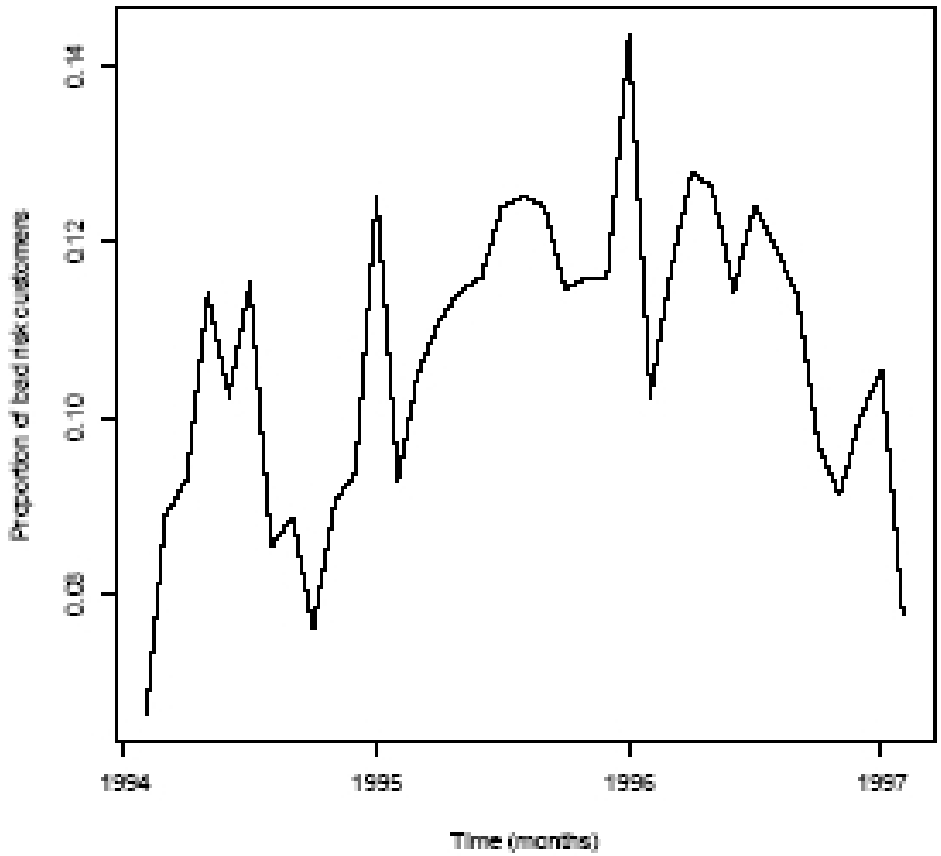


Figure 2: Proportion of bad risk accounts, by month, over the entire observation period.

Implementation of streaming classifier for UPLs

Frequency: arrival rate of data. Two simple approaches

- *Immediate updating* - classify/recompute in order of arrival- treating as regularly spaced. Simple heuristic.
- *Periodic updating* - working day, say. More difficult:
  1. multiple obs per period: incorporate mean vector for period
  2. no obs per period: no update.

Defer delayed label to future work.

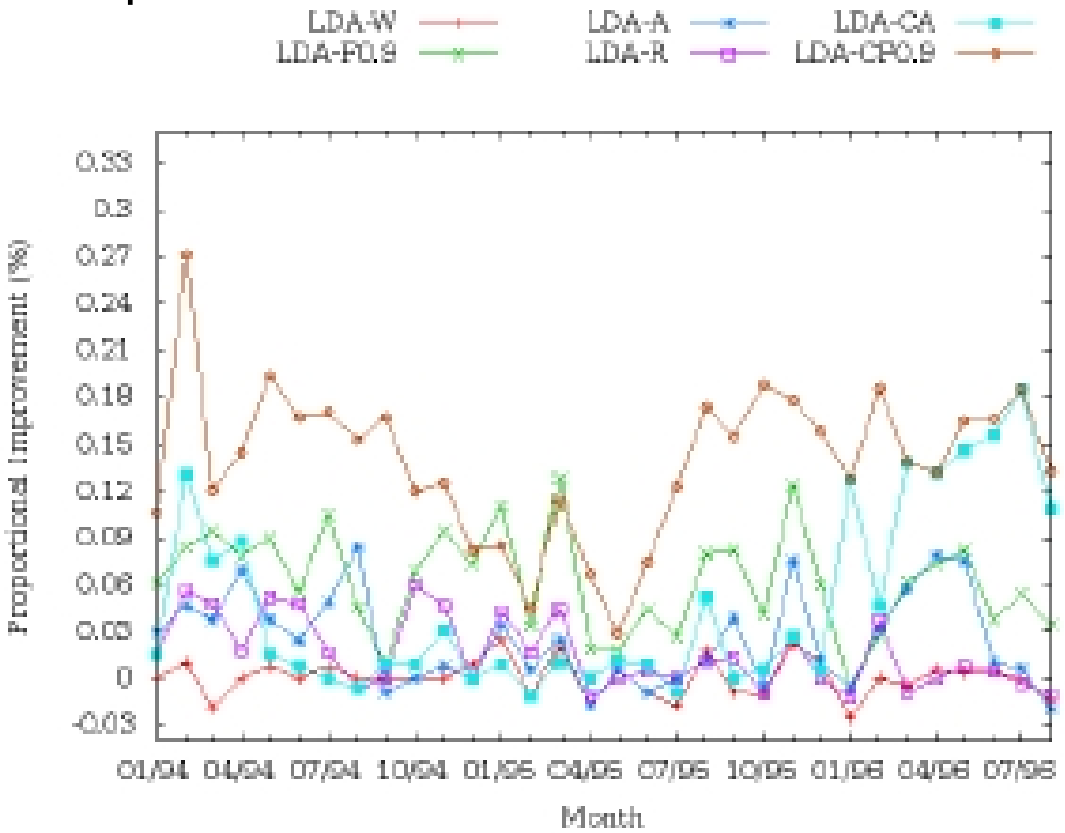
Methods in comparison

1. LDA-S: Static LDA - no rebuild
2. LDA-W: sliding Window LDA classifier - fixed width
3. LDA-R: Rebuilt LDA - yearly
4. LDA-Fλ: Fixed λ, $K + 1$ streaming LDA
5. LDA-CFλ: Centered fixed λ, $K + 2$ streaming LDA
6. LDA-A: Adaptive λ, $K + 1$ streaming LDA
7. LDA-CA: Centered Adaptive λ, $K + 2$ streaming LDA

Results

Base measure: (monthly) bad rate among accepts. Comparative measure: proportional improvement over LDA computed on first year's data.

Immediate update



Relative performance of various methods using immediate updating.

Results

Periodic update results similar - differences not as marked.

Main points:

- all streaming methods outperform static. Improvement small.
- sliding window similar performance to static.
- centered streaming LDA ($K + 2$ parameters) better performance.
- Fixed forgetting better than adaptive (when selected well!)
- Adaptive forgetting approaches exhibit some anomaly detection behaviour - large jumps in forgetting factors associated with calendar events.

Logistic

Comparing streaming with static logistic regression yields similar conclusions, namely

- Fully adaptive can yield improvement, but best fixed preferred.
- No substantive difference between immediate and daily updating.
- careful optimisation required, how does fisher scoring compare, and what is the relation to the flat maximum effect

Conclusions

- handling temporal aspects in consumer banking very difficult
- Some merit in deploying adaptive classifiers to try to retain classification performance between rebuilds
- Merit also for plastic card fraud detection.
- Some hacks: sampling frequency, updating, delay - handle in future work.

References

Adams, N.M., Tasoulis, D.K., Anagnostopoulos, C. and Hand, D.J., "Temporally-adaptive linear classification for population drift in credit scoring". Technical report (2009).

Anagnostopoulos, C., Tasoulis, D.K, Adams, N.M. and Hand, D.J., "Streaming Gaussian classification using recursive maximum likelihood with adaptive forgetting". *Machine Learning*, submitted.

Haykin, S., "Adaptive Filter Theory", Prentice Hall (1996).

Kelly, M.G., Hand, D.J. and Adams, N.M. "The impact of changing populations on classifier performance". In S.Chaudhuri & D.Madigan (eds.), Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (1999), 367-371.

Pavlidis, N.G., Tasoulis, D.K., Adams, N.M. and Hand. D.J.,"Adaptive online logistic regression", *Pattern Recognition*, submitted.

J. Whittaker, et al. "A dynamic scorecard for monitoring baseline performance with application to tracking a mortgage portfolio". *Journal of the Operational Research Society* 58(11), (2007), 911–921.