UNIVERSIDAD
DE MÁLAGA

Using GNUsmail to Compare Data Stream Mining Methods for On-line Email Classification

José M. Carmona-Cejudo[1], Manuel Baena-García[1], José del Campo-Ávila[1], João Gama[2], Albert Bifet[3] and Rafael Morales-Bueno[1]

[1]Universidad de Málaga, Spain
[2]University of Porto, Portugal
[3]University of Waikato, New Zealand

Manuel Baena-García
Castro Urdiales, October 2011

Content

Context

Email mining

- Spam detection: a two-class problem usually solved with bayesian classifiers.
- Email classification: a multi-class problem to sort email into folders.

Email classification approaches

- Batch learning. The whole dataset is available before the beginning of the learning process.
- online learning. Data are continually being received and processed over time.

Hypothesis

There is a lack of systems to compare and evaluate different machine learning models for email classification.

Context

Email mining

- Spam detection: a two-class problem usually solved with bayesian classifiers.
- Email classification: a multi-class problem to sort email into folders.

Email classification approaches

- Batch learning. The whole dataset is available before the beginning of the learning process.
- online learning. Data are continually being received and processed over time.

Hypothesis

There is a lack of systems to compare and evaluate different machine learning models for email classification.

Context

Email mining

- Spam detection: a two-class problem usually solved with bayesian classifiers.
- Email classification: a multi-class problem to sort email into folders.

Email classification approaches

- Batch learning. The whole dataset is available before the beginning of the learning process.
- online learning. Data are continually being received and processed over time.

Hypothesis

There is a lack of systems to compare and evaluate different machine learning models for email classification.

Context

GNUsmail

GNUsmail is a framework that allows to compare different email classification algorithms.

Contributions

We introduce next improvements to GNUsmail:

- to carry out *replicable experimentation*.
- to evaluate data stream mining methods by using:
  - sliding windows.
  - fading factors.
- to use recently proposed statistical tests to compare the performance of online algorithms.

Content

GNUsmail: Architecture and Characteristics

http://code.google.com/p/gnusmail/

Characteristics

- Open source framework for online adaptive email classification.
- It contains modules for reading email, preprocessing text and learning.
- The email messages are read as the model is built.

Architecture

- **Reading email module** can obtain email messages from local filesystem or remote IMAP server.
- **Text processing module** based on filters that extract attributes from emails.
- **Learning module** into which new algorithms, methods and libraries can be integrated.

Text Preprocessing Module

Structure

- A pipeline of (linguistic) operators which extract relevant features from every mail.
- Some ready-to-use filters are implemented as part of the GNUsmail core, and new ones can be incorporated.

Filters

- Relevant words based on the ranking provided by the tf-idf function.
- Sender, CC.
- Domain of sender.
- Capital letters proportion.
- Language.
- Number of receivers.

Learning Module

Structure

Based on WEKA and MOA frameworks:

- WEKA methods are used with small datasets in environments without time and memory restrictions.
- MOA methods are used in more demanding problems.

WEKA methods

- Multinomial Naïve Bayes
- IBk, k-nearest neighbours
- NN-ge (Nearest Neighbour with Generalised Exemplars)

MOA methods

Learning Module

structure

Based on WEKA and MOA frameworks:

- WEKA methods are used with small datasets in environments without time and memory restrictions.
- MOA methods are used in more demanding problems.

WEKA methods

MOA methods

GNUsmail uses MOA by including its tools for evaluation, classification, and drift detection.

- HoeffdingTree
- OzaBag, OzaBoost
- DDM

Content

Evaluation of Data Stream Mining Methods

In data stream contexts, neither cross-validation nor other sampling procedures are suitable for evaluation.

Prequential measures

- A prediction is made for each new example.
- Once the real class is known we update a cumulative loss function.

Forgetting mechanisms

- Sliding windows.
- Fading factors (preferred method).

Comparing the performance

Adapted McNemar statistic (M)

$$M_i = sign(a_i - b_i) \times \frac{(a_i - b_i)^2}{a_i + b_i}$$

- $a_i = f_i + \alpha \cdot a_{i-1}$
- $b_i = g_i + \alpha \cdot b_{i-1}$
- f_i: 1 iff the example i is misclassified by the first classifier and not by the second one (0 otherwise).
- g_i: 1 iff the example i is misclassified by the second classifier and not by the first one (0 otherwise).

Content

Experimental Setup

Initial setup

- Based on ENRON email dataset.
- We have selected seven specific users.
- We have used only topic folders with more than two messages.
- GNUsmail checks the availability of data, offering to download it.
- The messages are analyzed in chronological order.

Attributes

- Sender username, sender domain.
- Number of recipients, body length, capital letters proportion, size of email, subject length.
- Most relevant words.

Experimental Setup

Algorithms

- OzaBag over NNge, using DDM for concept drift detection.
- NNge.
- Hoeffding Trees.
- Majority class.

Output

GNUsmail plots...

- the prequential-based metrics
- and the adapted McNemar test

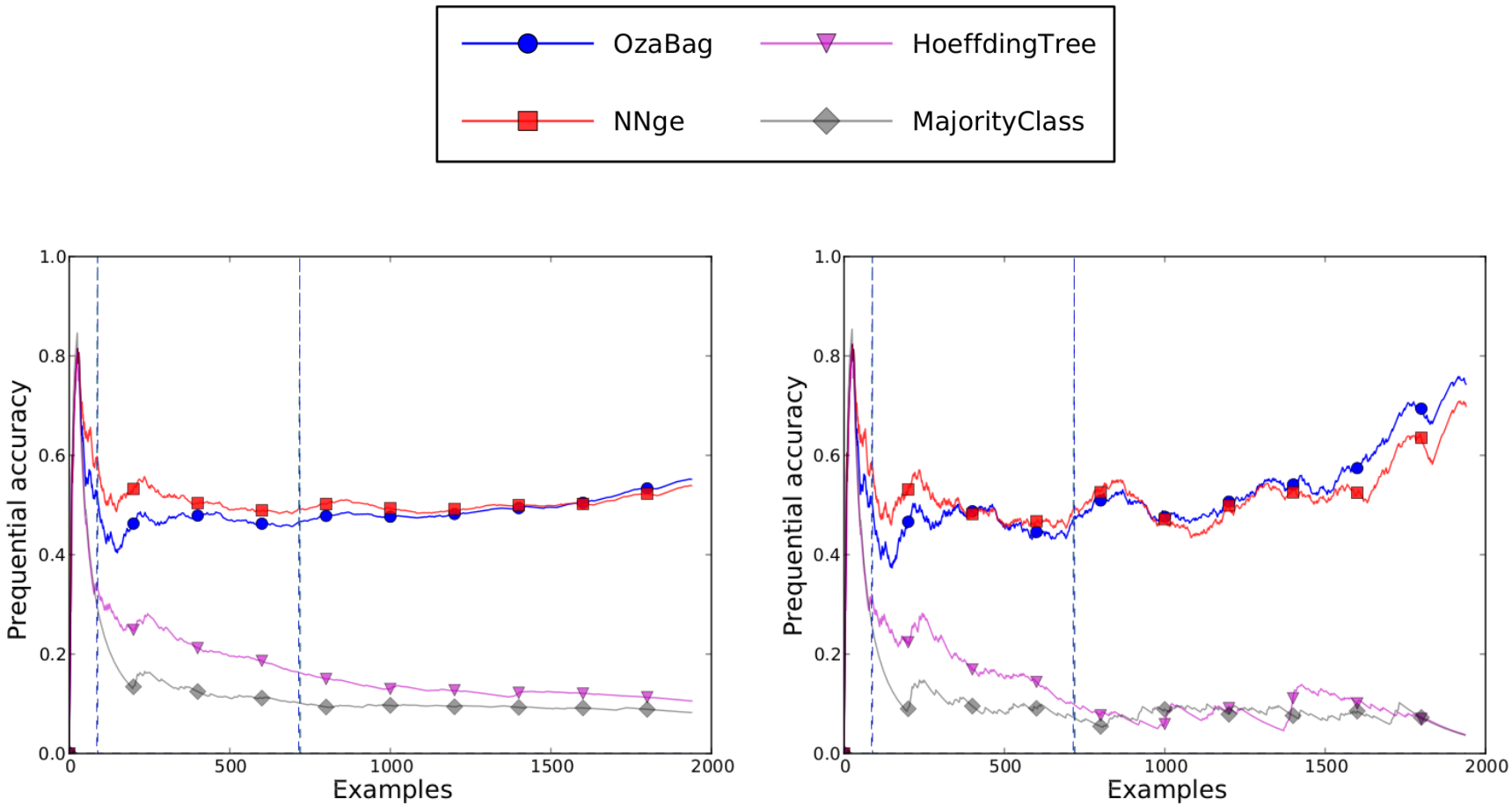
... to visually analyse the differences in performance.

Results

Folder-wise prequential accuracies with bagging of DDM and NN-ge

Folder	Correct/Total	Percentage
beck-s (101 folders)	1071/1941	55.18%
europe	131/162	80.86%
calendar	104/123	84.55%
recruiting	89/114	78.07%
doorstep	49/86	56.97%
kaminsky-v (41 folders)	1798/2699	66.62%
universities	298/365	81.64%
resumes	420/545	77.06%
personal	154/278	55.4%
conferences	163/221	73.76%
lokay-m (11 folders)	1953/2479	78.78%
tw_commercial_group	1095/1156	94.72%
corporate	345/400	86.25%
articles	152/232	65.51%
enron_t_s	86/176	48.86%
williams-w3 (18 folders)	2653/2778	95.5%
schedule_crawler	1397/1398	99.91%
bill_williams_iii	1000/1021	97.94%
hr	74/86	86.05%
symsees	74/81	91.36%

Results for beck-s

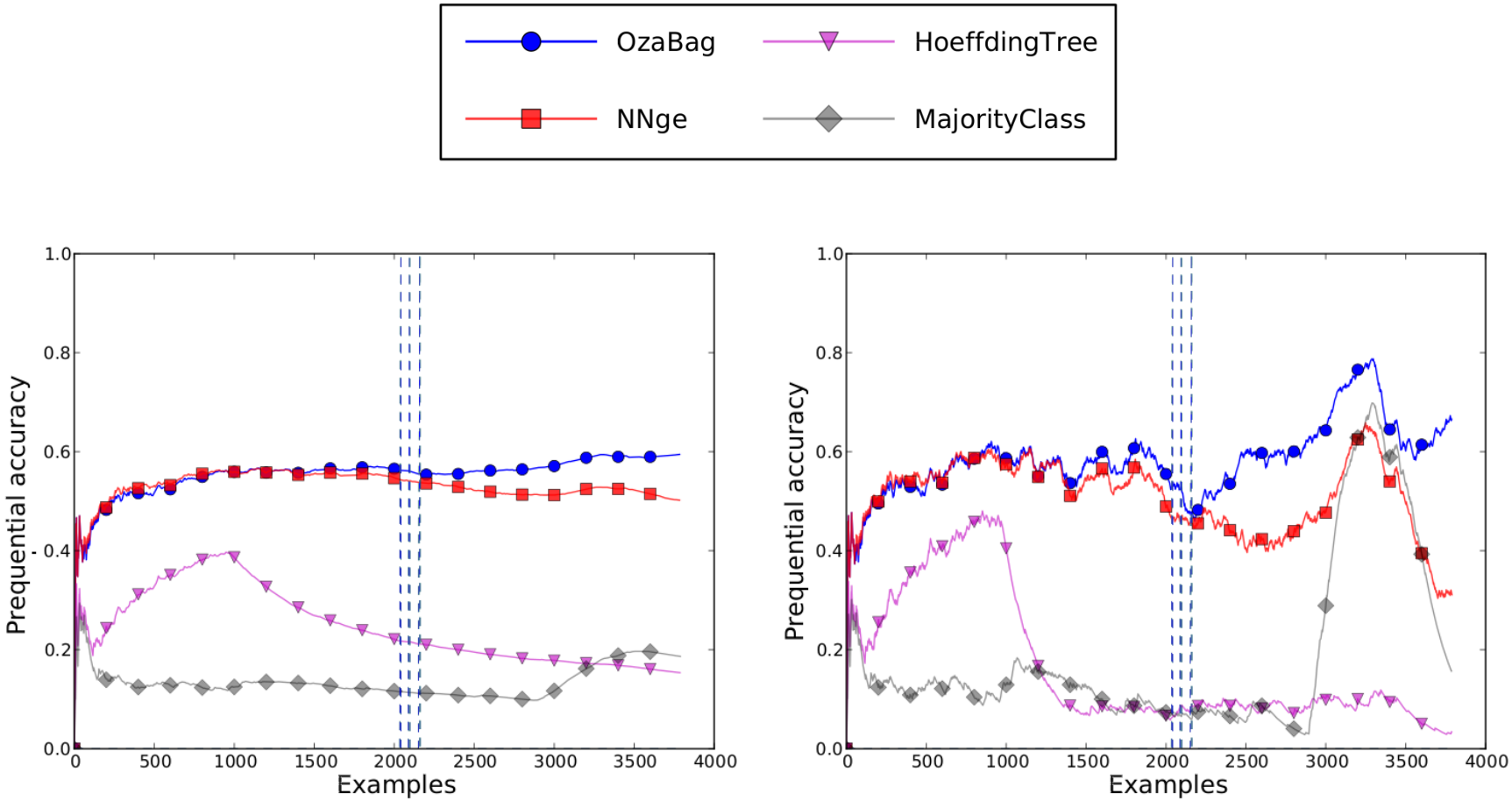


(a) Prequential error

(b) Fading factors preq. (α = 0.995)

Figure: Prequential based results for beck-s

Results for kitchen-l

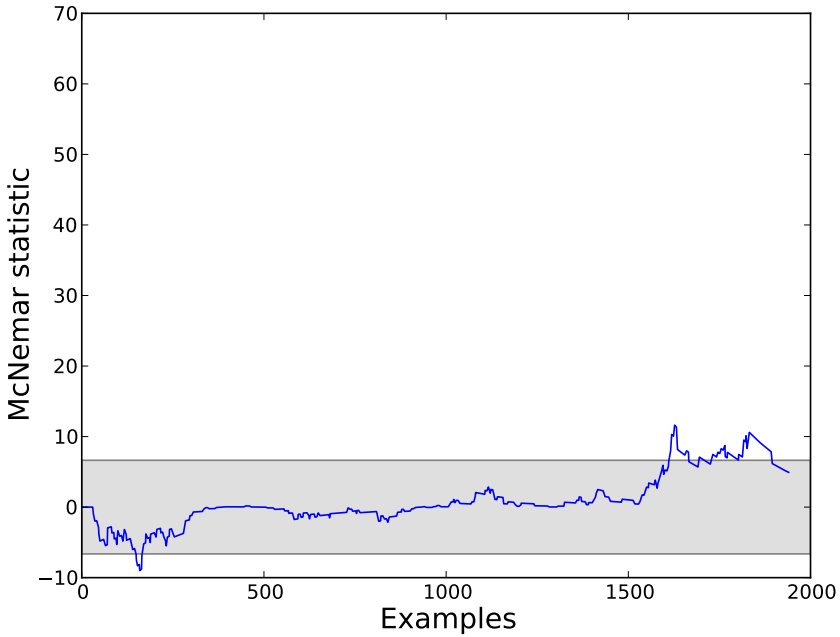


(a) Prequential error

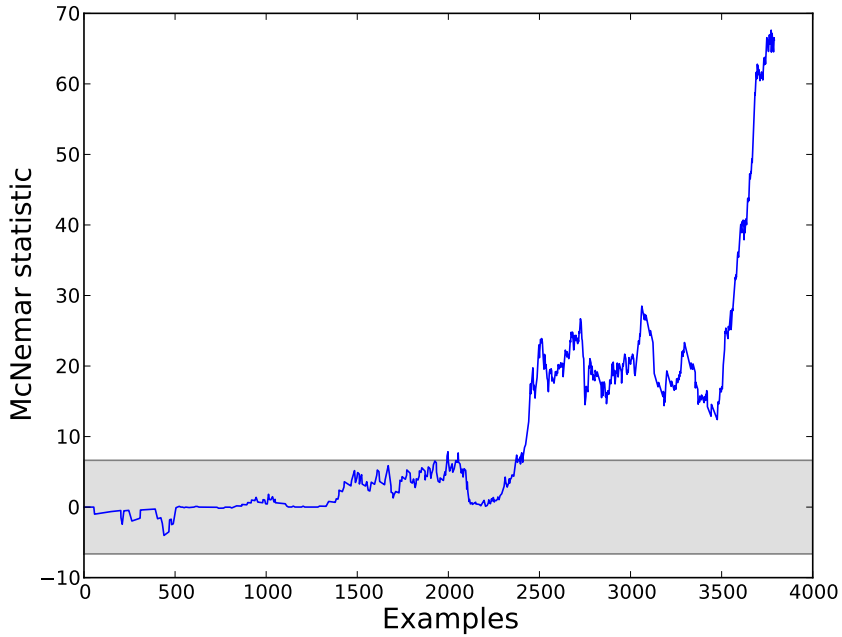
(b) Fading factors preq. (α = 0.995)

Figure: Prequential based results for kitchen-l

Adapted McNemar test



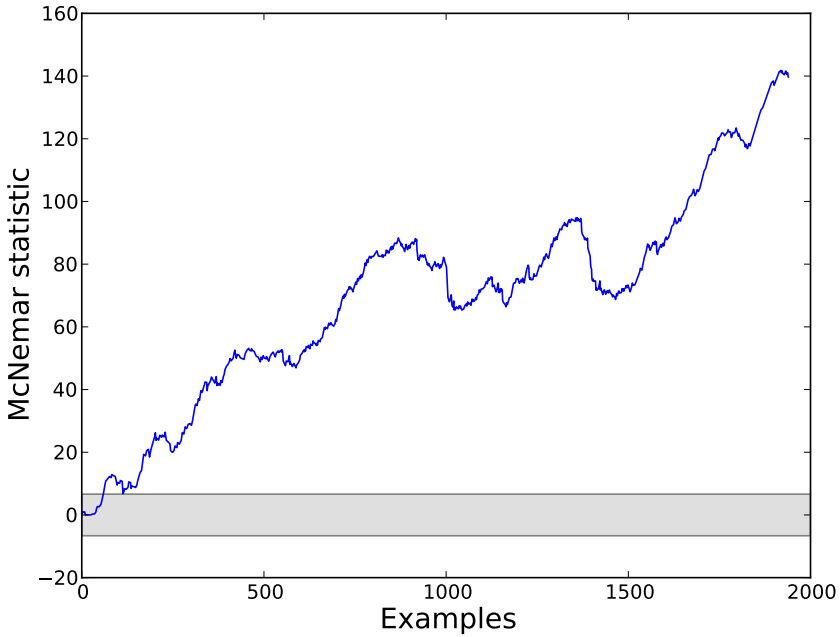
(a) beck-s



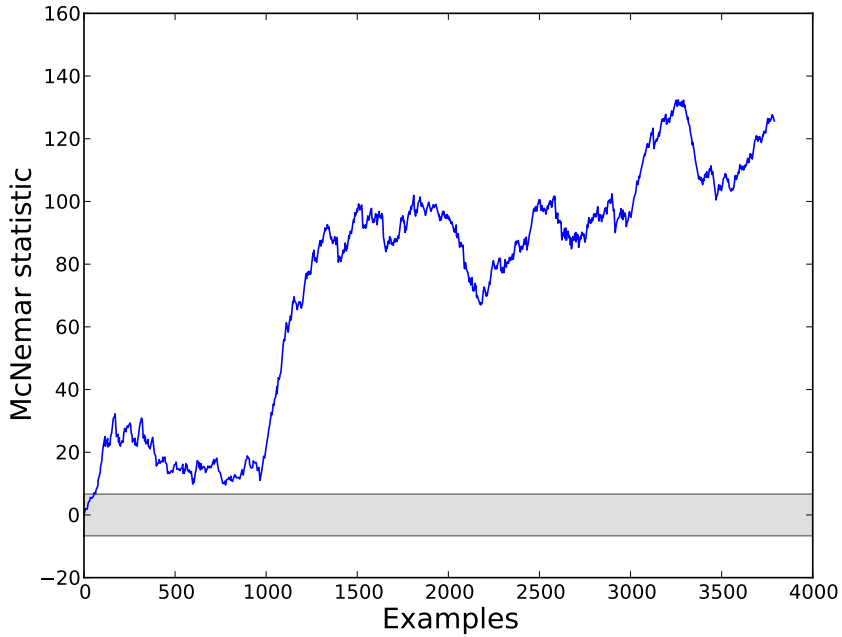
(b) kitchen-l

Figure: OzaBag vs. NN-ge using fading factors with $\alpha = 0.995$

Adapted McNemar test



(a) beck-s



(b) kitchen-l

Figure: OzaBag vs. Hoeffding tree using fading factors with $\alpha = 0.995$

Content

Conclusion and Future Work

- We have presented different methods to evaluate data stream algorithms.
- We have incorporated to GNUsmail recently proposed evaluation methods.
- Such evaluation methods improve prequential error measures.
- McNemar test is adequate as a tool to compare the online performance in the domain of email classification.
- Current online learning algorithm implementations needs to known all the attributes before the learning itself.
- Future methods should support online addition of new features.

UNIVERSIDAD
DE MÁLAGA

Using GNUsmail to Compare Data Stream Mining Methods for On-line Email Classification

José M. Carmona-Cejudo[1], Manuel Baena-García[1], José del Campo-Ávila[1], João Gama[2], Albert Bifet[3] and Rafael Morales-Bueno[1]

[1]Universidad de Málaga, Spain
[2]University of Porto, Portugal
[3]University of Waikato, New Zealand

Manuel Baena-García
Castro Urdiales, October 2011