Department of Computer Science

REAL-TIME FAST PREDICTIVE RULE INDUCTION DIRECTLY FROM CONTINUOUS STREAMING DATA

Dr Frederic Stahl

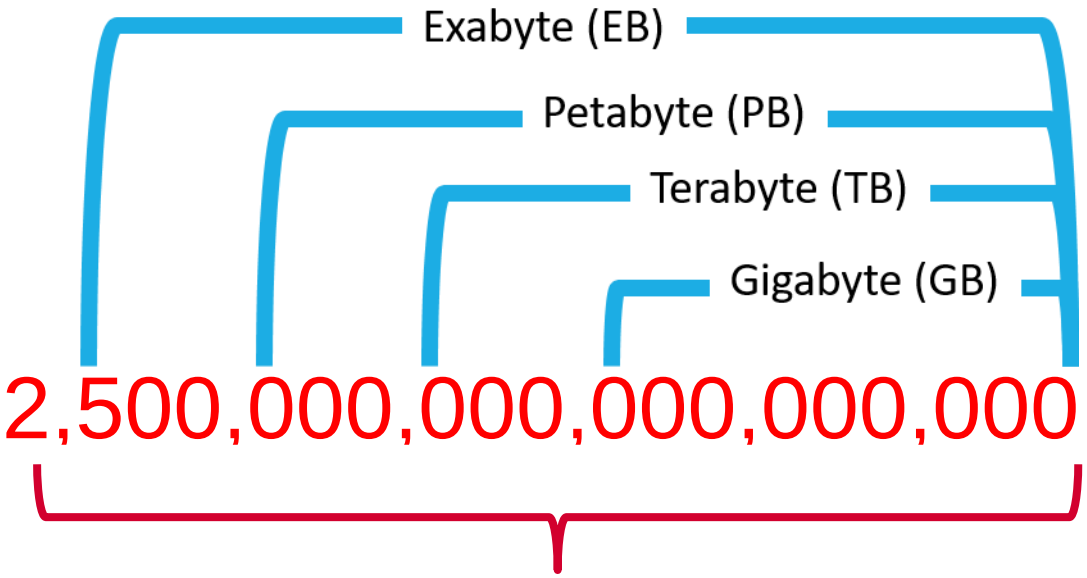
Lecturer in Data Science

OUTLINE

- Data Stream Mining
  - Data Stream Sources
  - Concept Drift
  - Data Stream Classification
- eRules Family of Algorithms: Expressive Rule Induction from Data Streams
  - Modular Rule Induction
  - Real-time Adaptive Rule Induction from Streaming Data
  - Expressive Rule Terms
- Evaluation
- Conclusions
- Resources

DATA STREAM MINING

BIG DATA



2.5 Quintillion bytes of data is generated every day, most of which is never captured, never collected, with no corresponding action taken. [IBM/Cisco]

SOURCES OF DATA STREAMS

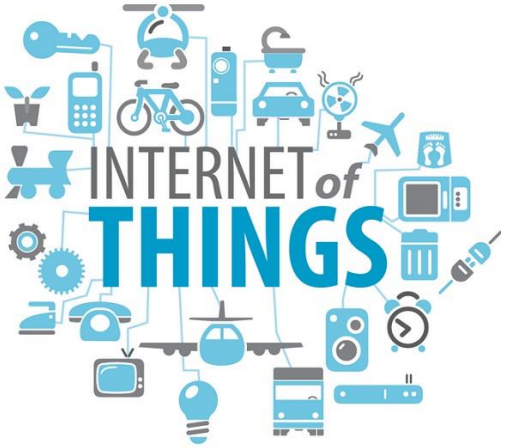
The advances in data acquisition hardware and the emergence of applications that process continuous flow of data records have led to the data stream phenomenon.

Spam Detection

- Emails:
  - 1.9 billion email users
  - 90 trillions emails sent a year
  - 90% are either spams or viruses [7]

Personalization

- Facebook:
  - 1.35 billion active users
  - 4.75 billion pieces of content shared
  - 300 millions photos uploaded per day [8]



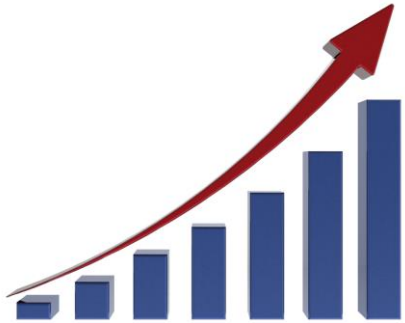
Internet of Things

- By year-end 2018, 25% of durable good manufacturers will utilize data generated by smart machines. And by 2018, 6 billion “Things” will request support. [Gartner]

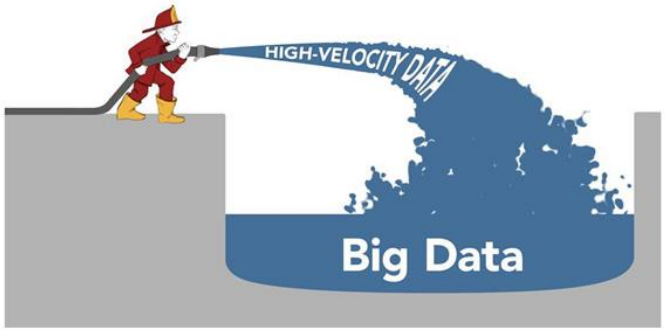
STATIC VERSUS STREAMING DATA

A data stream is *a continuous, rapid flow of data that challenge our state-of-the-art processing and communication infrastructure.*

Static Data	Streaming Data
• Random Access • Secondary Storage • Little or No Time Requirements • Assumed Complete Data	• Sequential Access • Limited Memory • Real-Time Requirements • Assumed Outdated/ Inaccurate Data

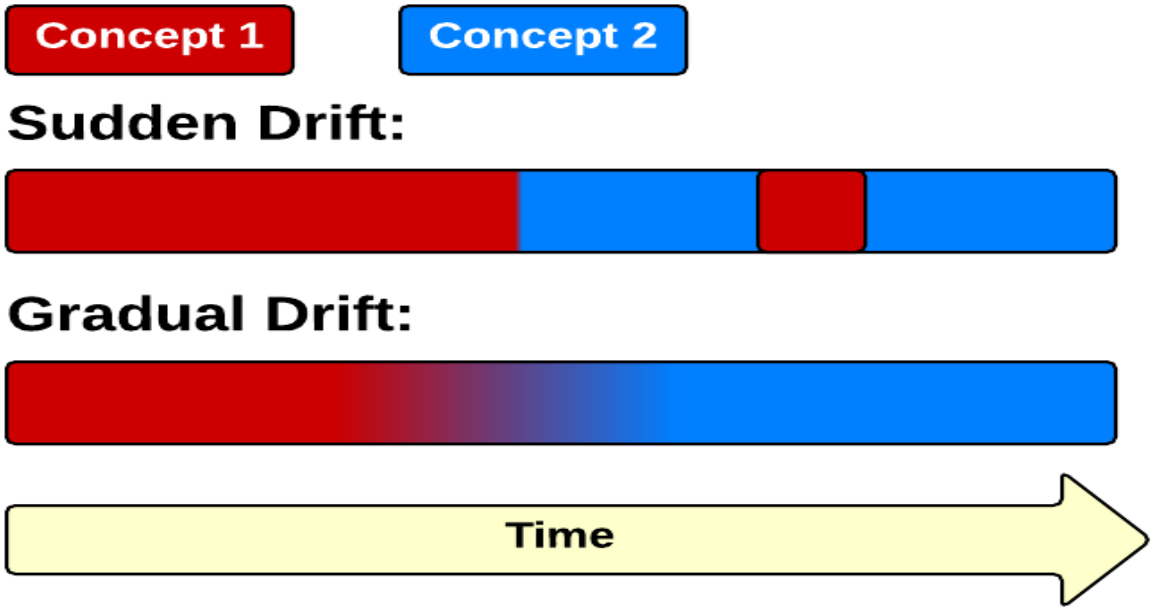


Volume and **Velocity**



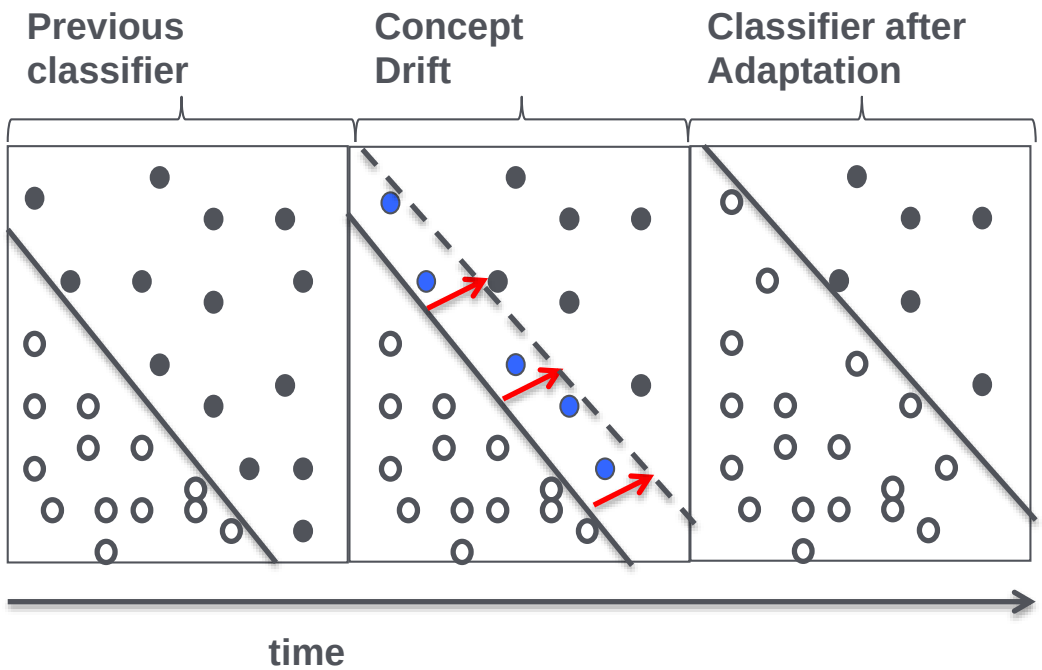
CONCEPT DRIFT

- Underlying concept defining the knowledge being learned begins to shift over time
- The concept change is unforeseen and unpredictable
- Concepts in the past may re-occur in the future
- Concept drift exists in real-life problems:
  - Seasonal weather
  - Stock market rallies because of breaking news
  - etc.



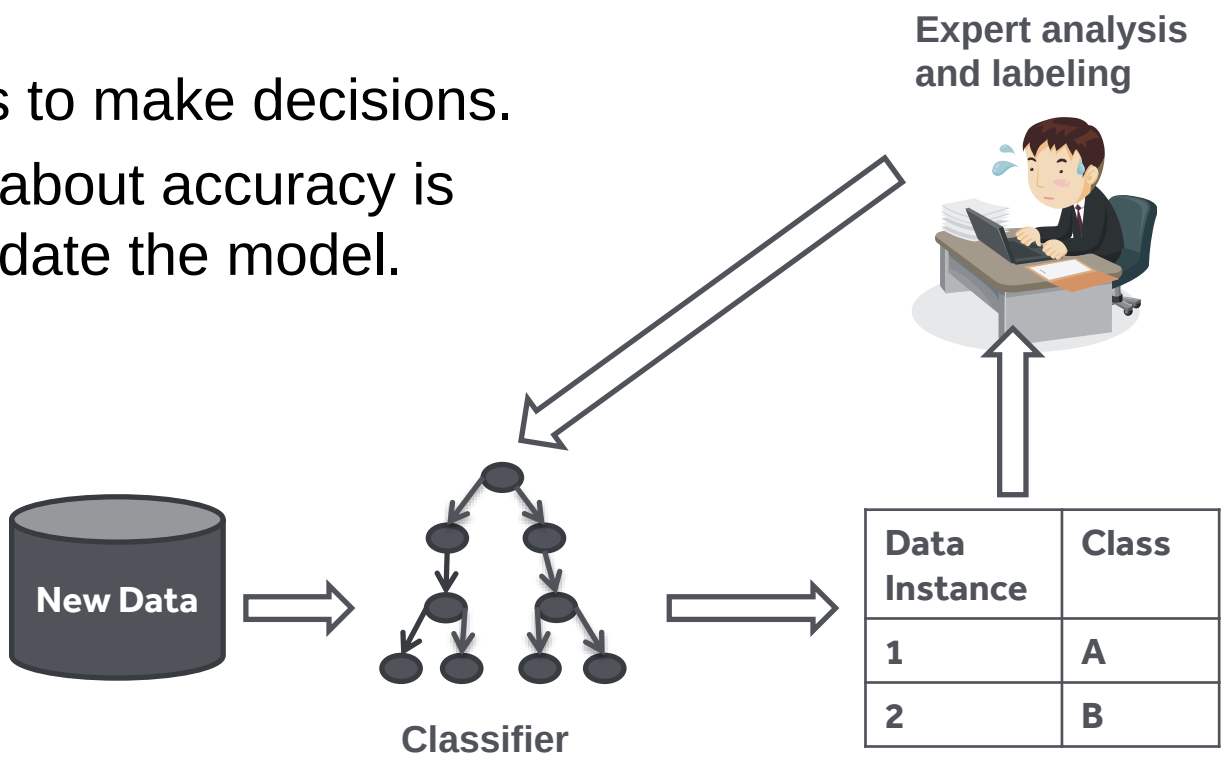
CONCEPT DRIFT

- A model should always reflect the time-changing concept.
- A model should be rendered invalid if it becomes inaccurate
- Robustness and adaptability: a model should quickly recover/adjust after or better during concept drift.



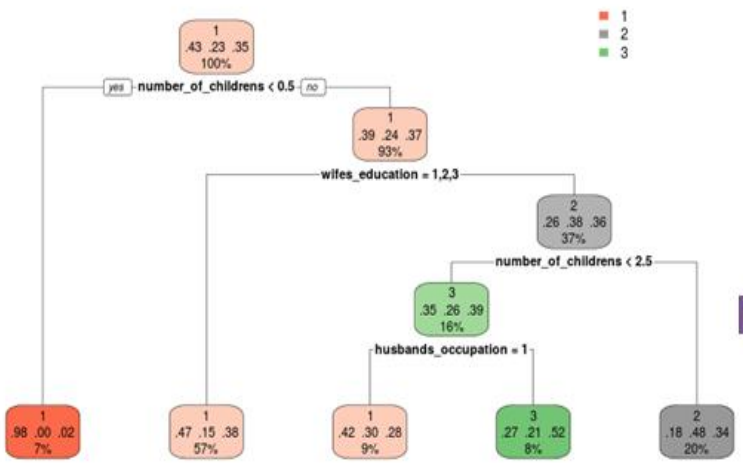
PREDICTIVE DATA STREAM MINING

- A classifier is built with past labeled data.
- It predicts the label of future instances.
- Thus helps to make decisions.
- Feedback about accuracy is used to update the model.



BLACK BOX VS. HUMAN UNINTERPRETABLE MODELS

- **Black box models provide no clear understandings** of internal working.
- The user has **no** or **little control over** the process that produces output (i.e. classifications) for a given input.
- They are difficult to **interpret** and **understand** by the user or even domain expert experts

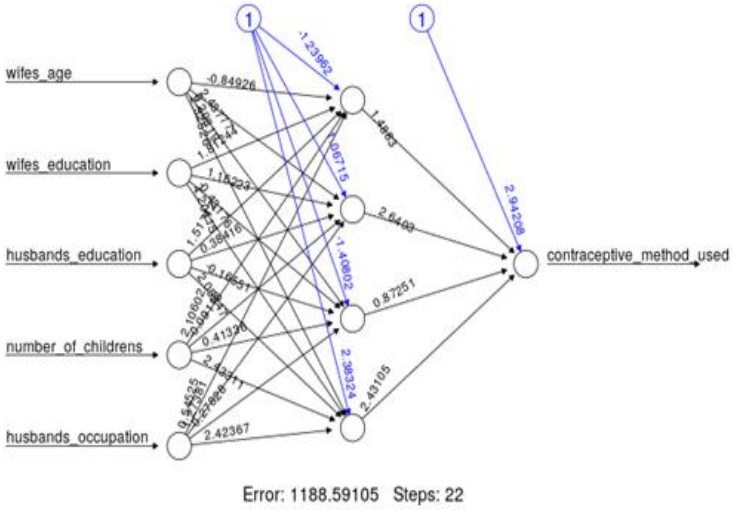


Decision Tree

<===

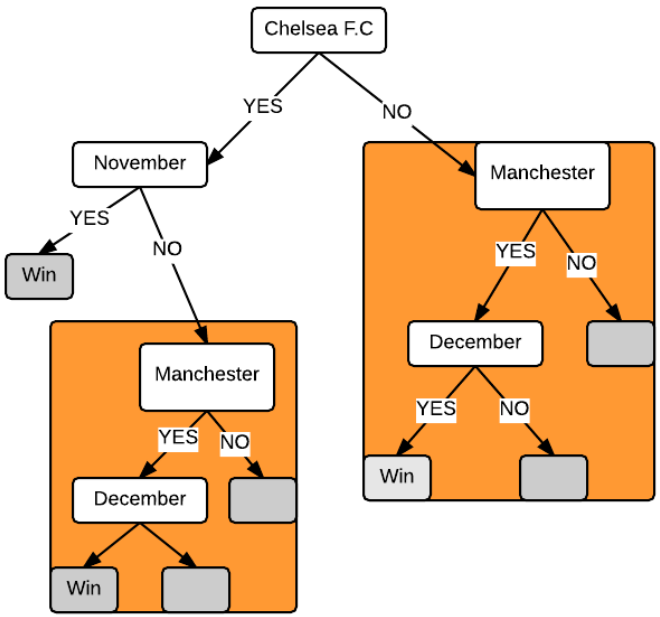
Neural Network

===>



EXPRESSIVE CLASSIFICATION RULES

- Each rule represents an independent piece of information
- A rule can easily be read and interpreted by a human
- Rules are much more compact than decision trees
- New rules can be added without disturbing the rule library



Rules:

IF **Chelsea=YES** AND **November=YES** THEN **Win**

Feature-Value **Target Class**

IF **Manchester=YES** and **December=YES** THEN **Win**

OBJECTIVES

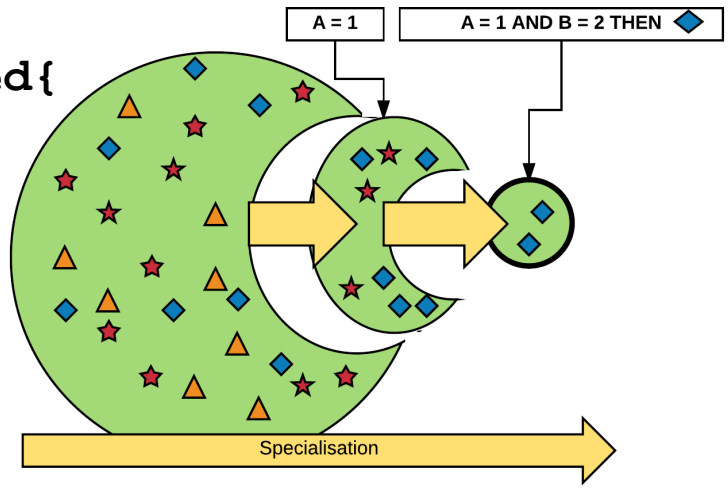
- Development of a predictive data stream mining techniques that are:
  - Expressive (human readable)
  - Adaptive to concept drift
  - Scalable with respect to existing predictive data stream mining techniques.

TOWARDS EXPRESSIVE RULE INDUCTION FROM DATA STREAMS WITH THE *eRules* FAMILY OF ALGORITHMS

RULE INDUCTION: SEPARATE AND CONQUER

- Each rule represents an independent piece of information
- A rule can easily be read and interpreted by a human
- Decision rules are much more compact than decision trees
- New rules can be added without disturbing the rule library

```
Rule_Set rules = new Ruleset();
While Stopping Criterion not satisfied{
    • Rule = learnRule();
    • Remover all instances
      covered form rule
    • rules.add(rule);
}
```



TRAINING DATA

University of Reading

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
young	myope	no	0.5	none
young	myope	no	0.9	soft
young	myope	yes	0.5	none
young	myope	yes	0.9	hard
young	hypermetrope	no	0.3	none
young	hypermetrope	no	0.9	soft
young	hypermetrope	yes	0.3	none
young	hypermetrope	yes	1	hard
pre-presbyopic	myope	no	0.3	none
pre-presbyopic	myope	no	0.8	soft
pre-presbyopic	myope	yes	0.5	none
pre-presbyopic	myope	yes	1	hard
pre-presbyopic	hypermetrope	no	0.4	none
pre-presbyopic	hypermetrope	no	0.8	soft
pre-presbyopic	hypermetrope	yes	0.4	none
pre-presbyopic	hypermetrope	yes	0.9	none
presbyopic	myope	no	0.4	none
presbyopic	myope	no	0.8	none
presbyopic	myope	yes	0.3	none
presbyopic	myope	yes	1	hard
presbyopic	hypermetrope	no	0.5	none
presbyopic	hypermetrope	no	1	soft
presbyopic	hypermetrope	yes	0.4	none
presbyopic	hypermetrope	yes	0.8	none

PRISM [5]

- Example contact lenses

-Classes: hard, soft, none

ALGORITHM

- Rule: IF ? THEN recommendation = hard
- Conditions:

Age = young	**2/8**
Age = Pre-presbyopic	**1/8**
Age = Presbyopic	**1/8**
Spectacle prescription = Myope	**3/12**
Spectacle prescription = Hypermetrope	**1/12**
Astigmatism = no	**0/12**
Astigmatism = yes	**4/12**
Tear production rate <= 0.5	**0/12**
Tear production rate > 0.5	**4/12**

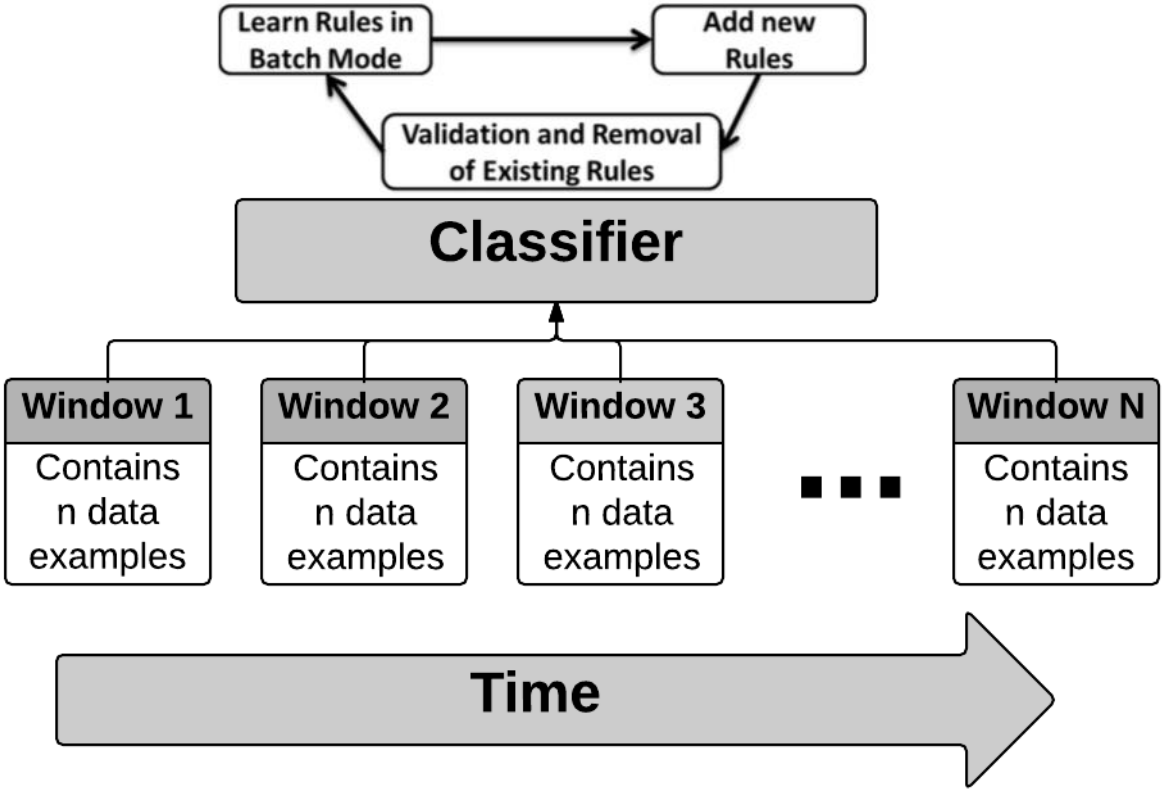
Rule: IF (astigmatism = yes) THEN recommendation = hard

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
young	myope	yes	0.5	none
young	myope	yes	0.9	hard
young	hypermetrope	yes	0.3	none
young	hypermetrope	yes	1	hard
Pre-presbyopic	myope	yes	0.4	none
Pre-presbyopic	myope	yes	1	hard
Pre-presbyopic	hypermetrope	yes	0.5	none
Pre-presbyopic	hypermetrope	yes	0.8	none
Presbyopic	myope	yes	0.4	none
Presbyopic	myope	yes	1	hard
Presbyopic	hypermetrope	yes	0.3	none
Presbyopic	hypermetrope	yes	0.8	none

Take covered instances and use these to generate the next term.

=> IF (astigmatism = yes) AND (age = young) THEN recommendation = hard

CONCEPTUAL eRules ALGORITHM



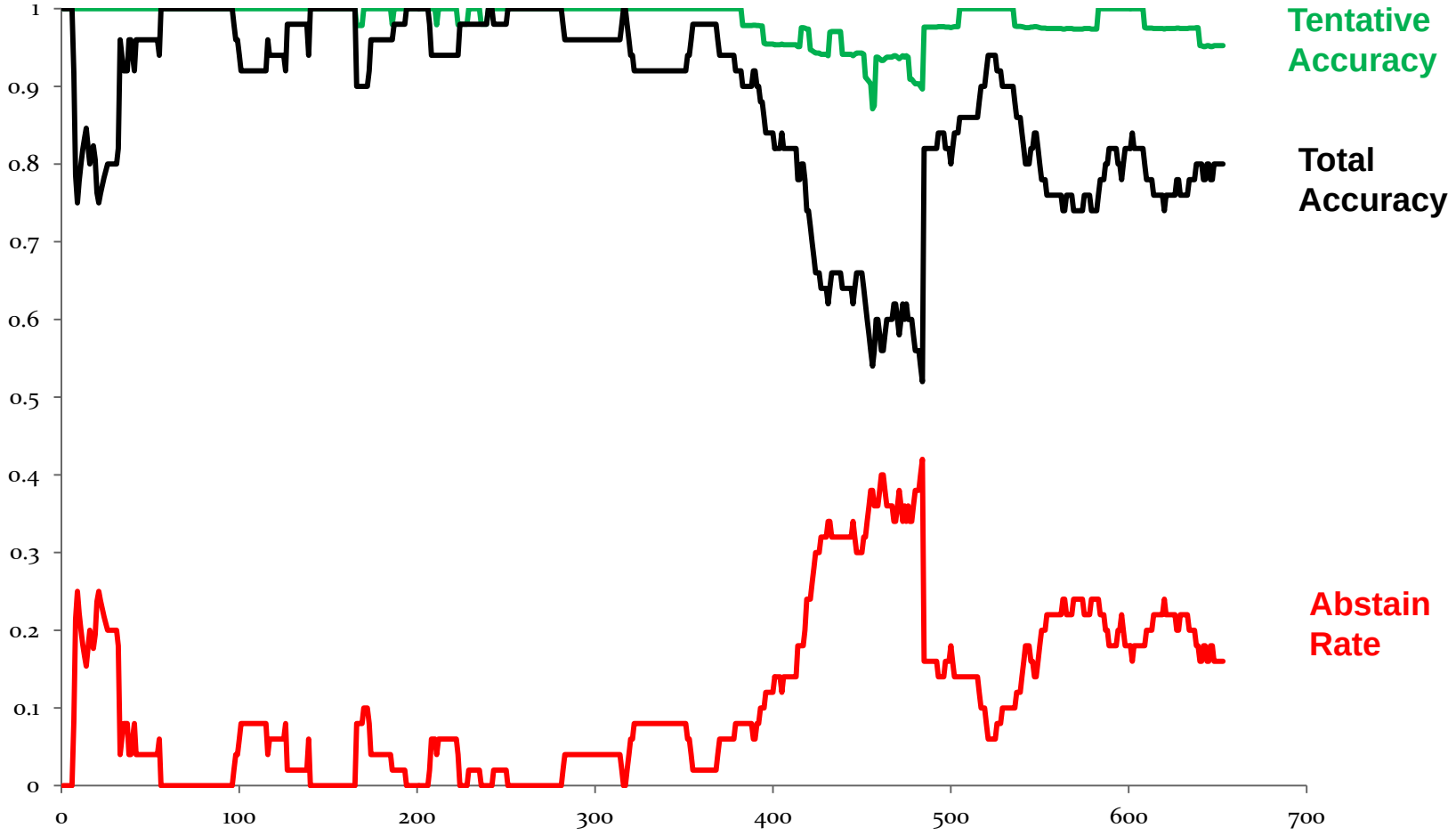
- 1st Process: buffer instances and generate initial ruleset
- 2nd Process: buffer recently abstained instances and induce new rules
- 3rd Process: remove rules with low accuracy

eRule's Abstaining Feature

- This is a notable feature of the eRules classifier, which is not available in most predictive data stream mining algorithms.
- eRules refuses to classify an unseen data example if it is uncertain about its class label
- Popular decision tree based approaches cannot abstain from classification.
- This abstaining feature maybe highly desirable in applications where miss-classification is costly and irreversible, such as in medical and financial applications.

[1]

A Typical Performance Plot of eRules



CONTINUOUS DATA

- Can be any value within a finite and an infinite range
- Many real-worlds problems are best expressed by continuous numeric values
  - E.g.: person's height, number of children, time in a race, a dog's weight, temperature and many more
- Most classifiers in the literature are able to deal with continuous feature by discretizing its continuous values into categorical values and treat them as a categorical feature
- Continuous attributes often require more computational efforts compared to categorical attributes

CUT-POINT CALCULATIONS TO INDUCE RULE TERMS

Attribute, x (Continuous)	**Attribute, y** (Categorical)	**Class Label, w** (A or B)
1	YES	B
2	NO	A
3	NO	A
4	YES	B
5	NO	A
6	NO	B

Categorical Attribute:

$P(w = A|y = YES) = 0.0$
$P(w = A|y = NO) = 0.75$

2 Calculations

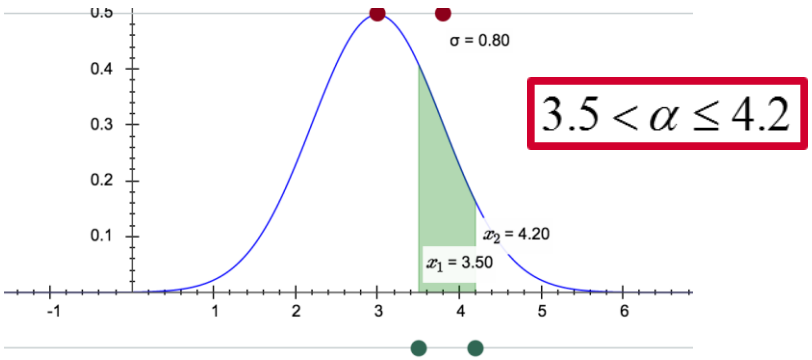
Continuous Attribute:

$P(w = A|x \leq 2) = 0.50$	**AND**	$P(w = A|x > 2) = 0.50$
$P(w = A|x \leq 3) = 0.66$	**AND**	$P(w = A|x > 3) = 0.33$
$P(w = A|x \leq 4) = 0.50$	**AND**	$P(w = A|x > 4) = 0.50$
$P(w = A|x \leq 5) = 0.60$	**AND**	$P(w = A|x > 5) = 0.00$
$P(w = A|x \leq 6) = 0.50$	**AND**	$P(w = A|x > 6) = N/A$

10 Calculations

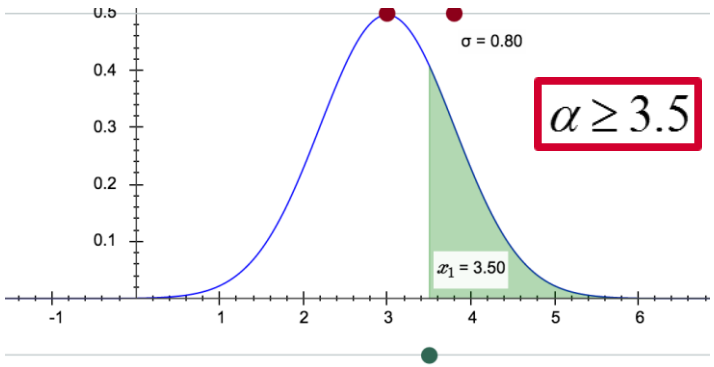
INDUCING RULE TERMS FOR CONTINUOUS ATTRIBUTE USING GAUSSIAN DISTRIBUTION

Rule term structure of **G-eRules**



- Rule term based on density estimation from Gaussian distribution per classification
- Maximize the coverage of the rule for a given target class
- This form of rule term indicates only highly relevant value ranges

Original rule term structure of **eRules**:

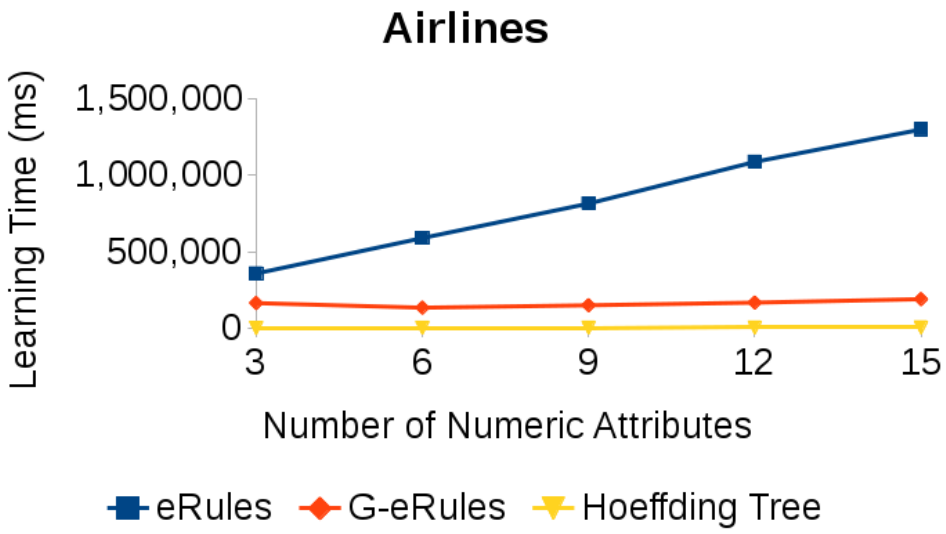


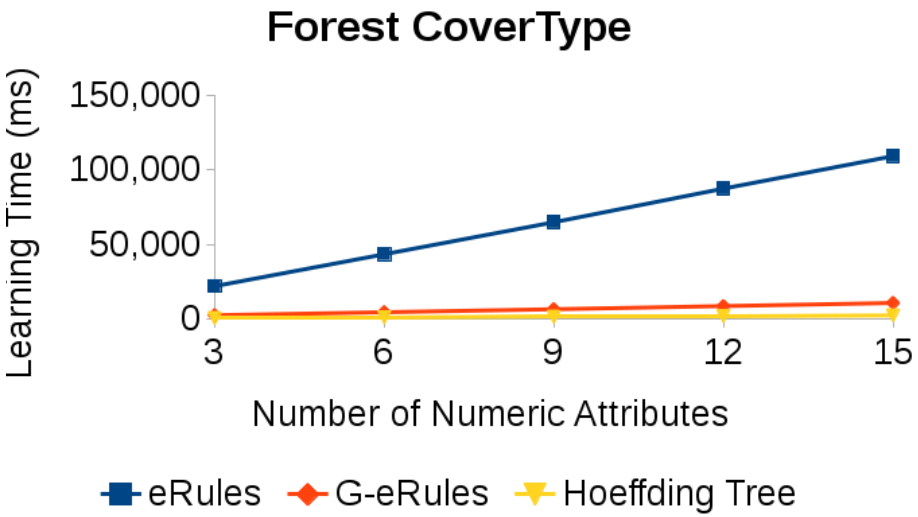
- Repeated calculations required to find best rule term for each iteration
- Rule term is more likely to cover several classifications
- Many more rule terms may be needed to introduce a complete rule

EVALUATION SETUP

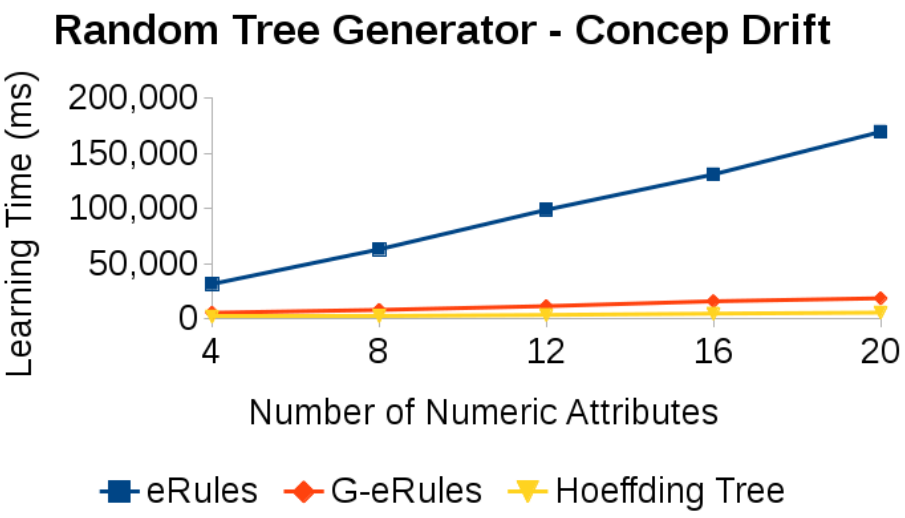
- Massive Online Analysis (MOA) [6] has been used for generating datasets with concept drift: SEA Concepts Generator; RandomTree Generator;
  - 50,000 instances
  - Drift at position 15,000 lasting 1000 instances.
- Real Datasets: Covertype (581,012 instances) and Airlines (500,000 instances)
- Setting of eRules & G-eRules:
  - Window Size 500
  - Minimum rule accuracy 0.8 and minimum classification attempts 5.
- Competitor Algorithms:
  - VFDR: a rule based stream classifier available in MOA [3]
  - Hoeffding Tree: state of the art decision tree classifier available in MOA [4]

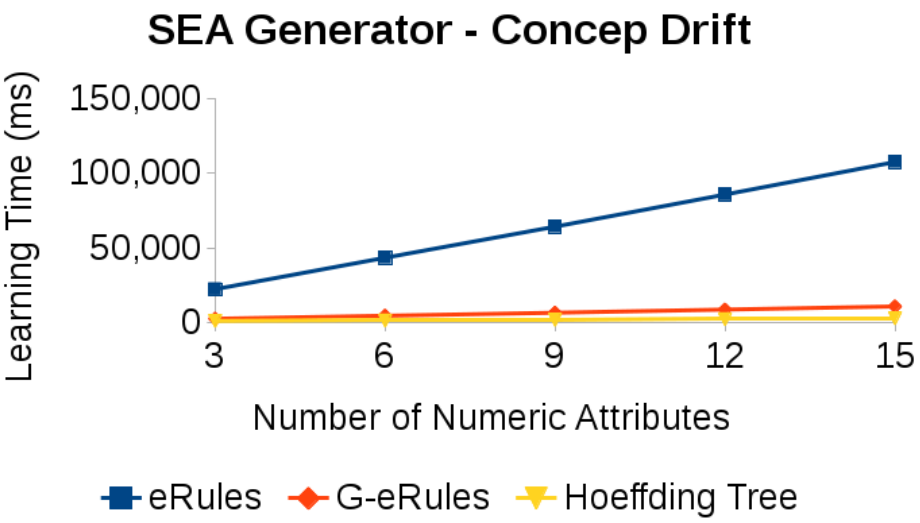
LEARNING TIME



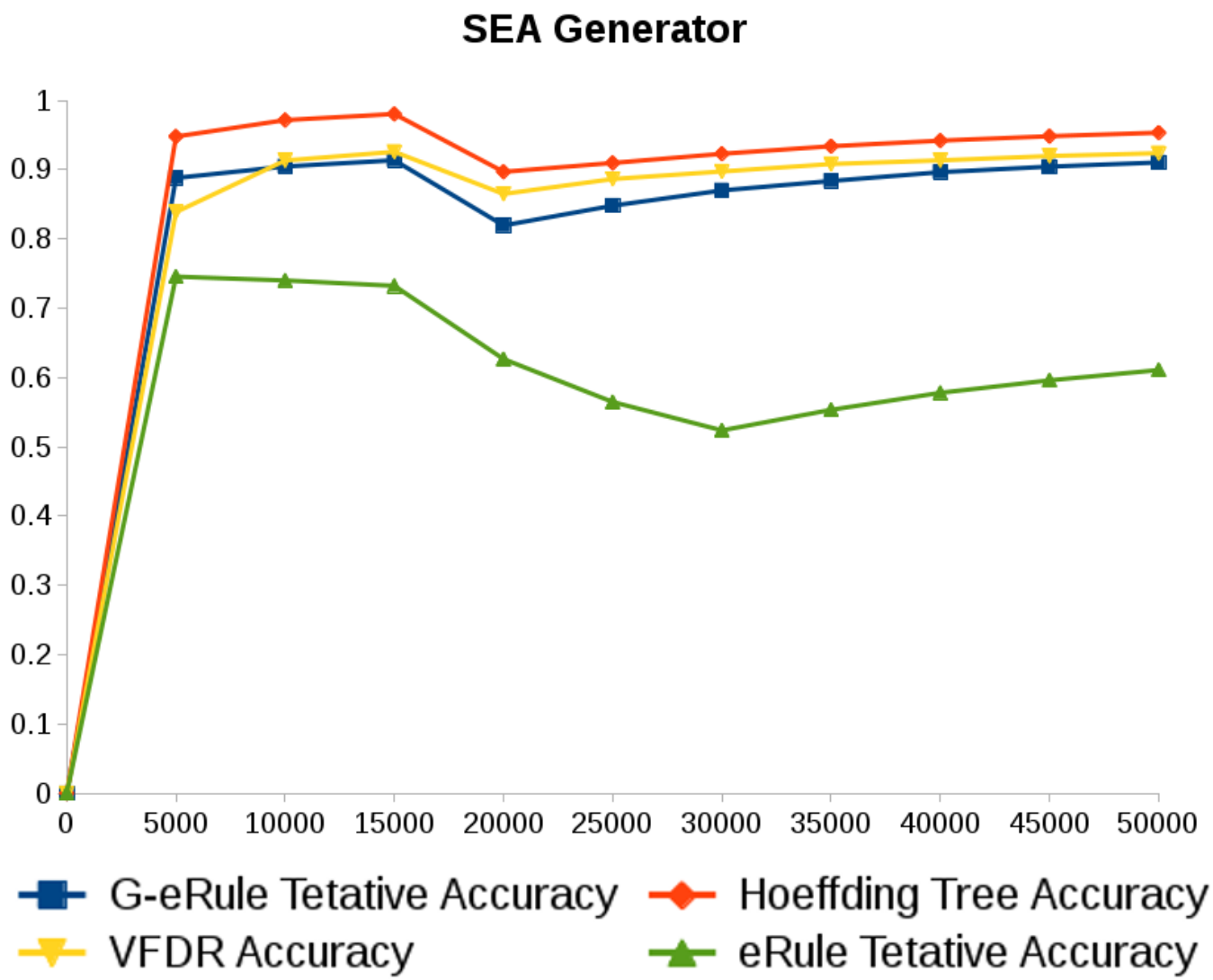


LEARNING TIME

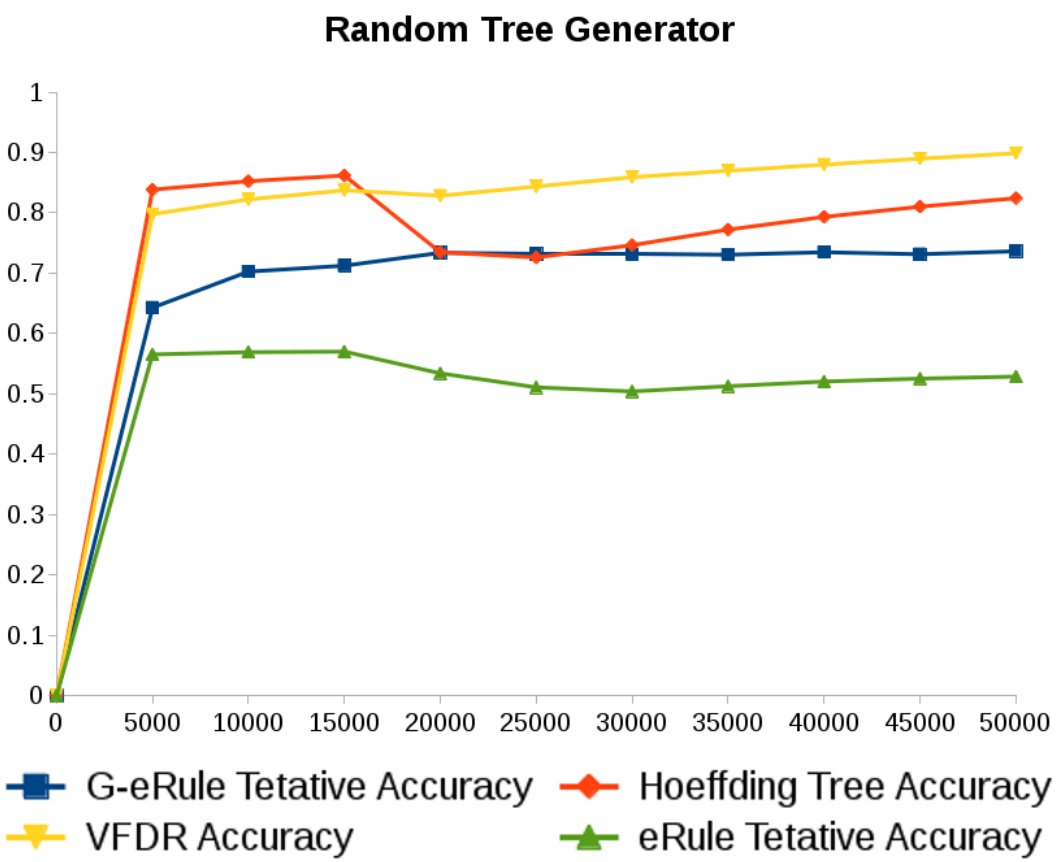




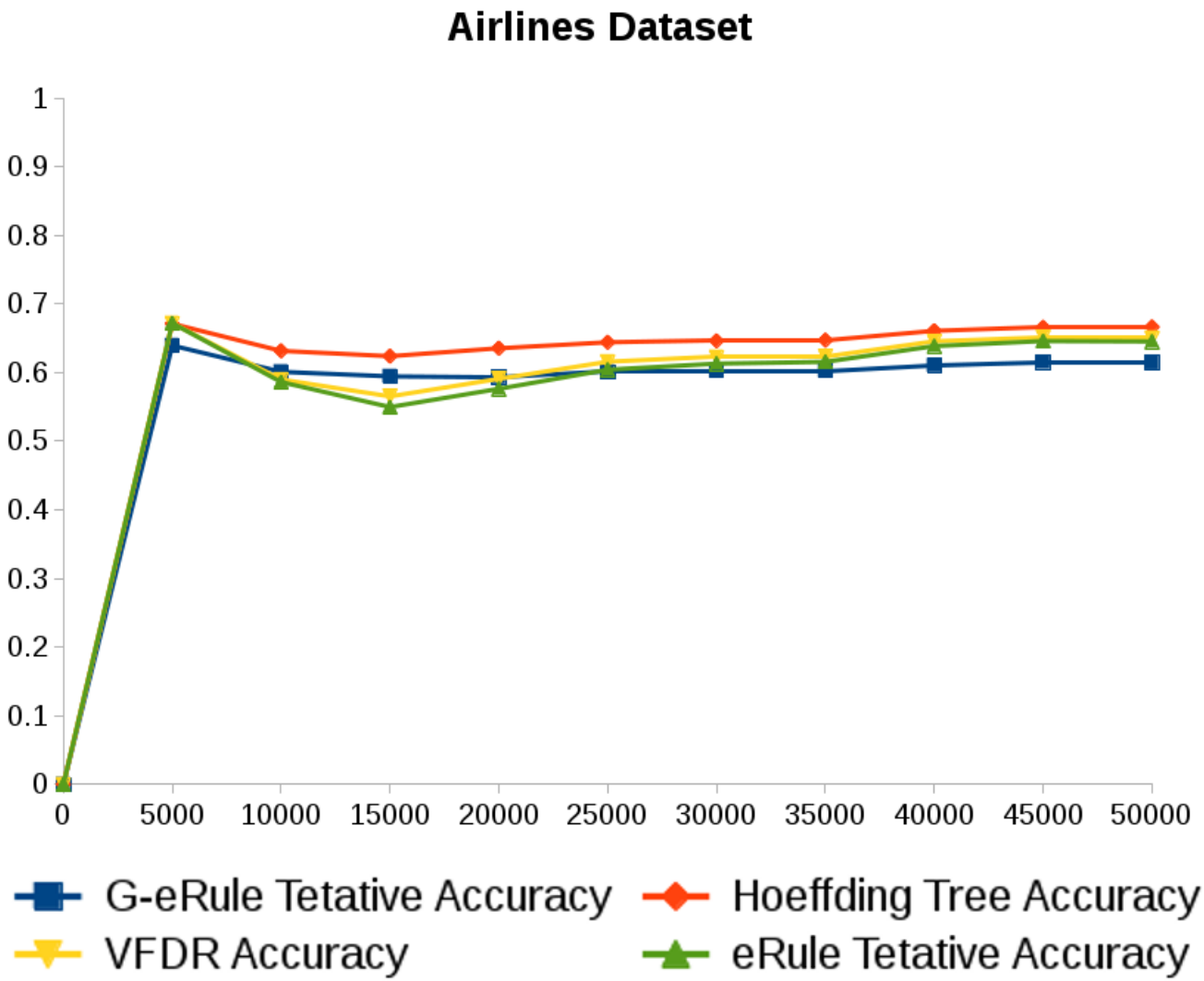
ACCURACY AND ADAPTATION



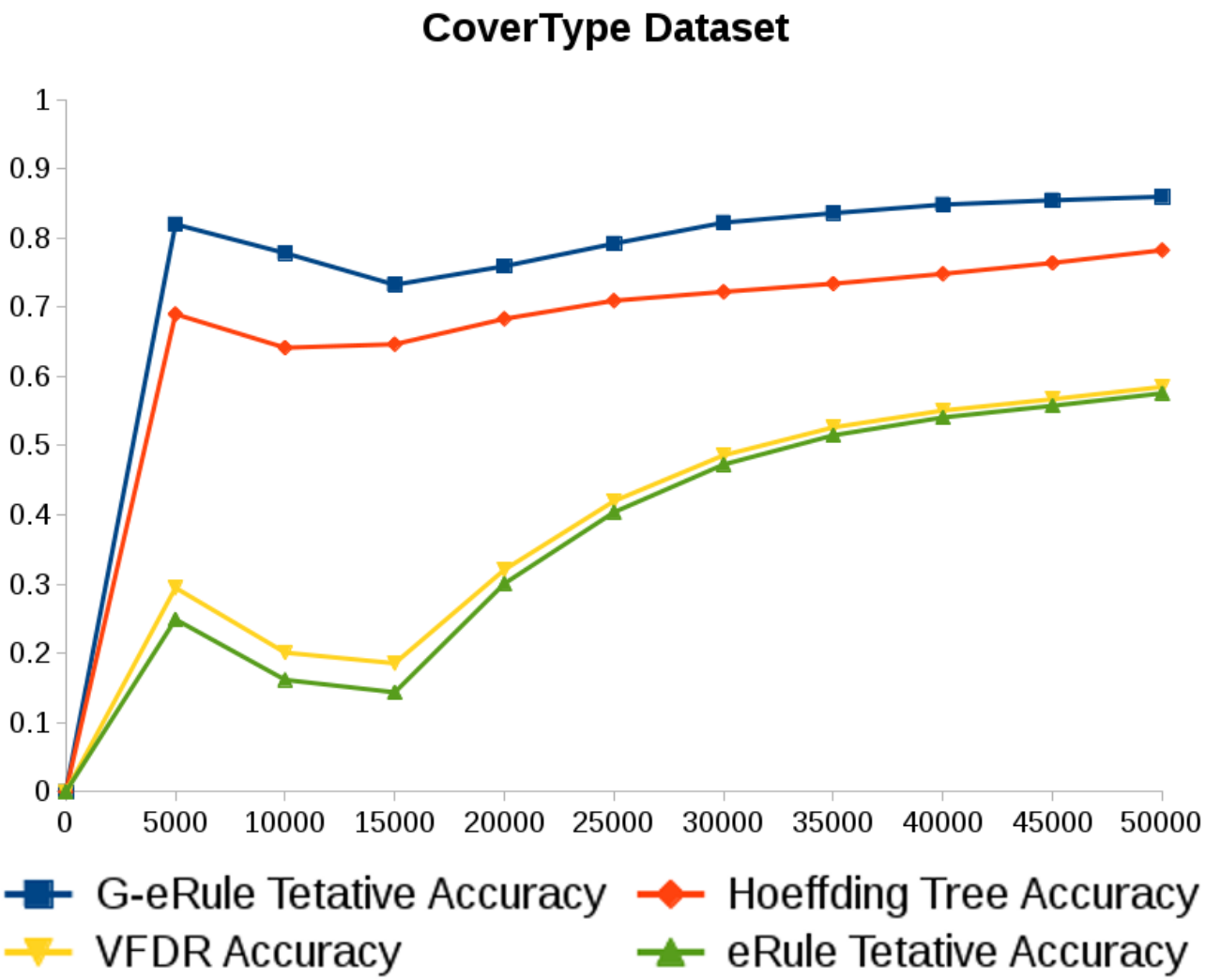
ACCURACY AND ADAPTATION



ACCURACY AND ADAPTATION



ACCURACY AND ADAPTATION



CONCLUSIONS

- Data from data streams is challenging to analyse due to concept drift, real-time requirements and infinite data.
- The work presented develops an expressive approach for predictive analytics on streaming data by inducing adaptive rulesets
- The developed approach is robust to concept drift, produces a competitive classification accuracy and is computationally efficient.

ACKNOWLEDGEMENTS

EPSRC
Engineering and Physical Sciences Research Council

This research has been supported by the UK Engineering and Physical Research Council EPSRC grant EP/M016870/1

RESOURCES

University of Reading

Publications:

- Stahl, F., Gaber, M. M. and Salvador, M. M. (2012) *eRules: a modular adaptive classification rule learning algorithm for data streams.* In Proceedings of The Thirty-second SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge. Springer, pp. 65-78. ISBN 9781447147381 doi: 10.1007/978-1-4471-4739-8-5
- Le, T., Stahl, F., Gomes, J.B., Gaber, M.M., and Di Fatta, G. (2014) *Computationally Efficient Rule-Based Classification for Continuous Streaming Data*. In Proceedings of the Thirty-Fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge. Springer, pp 21-34. ISBN 978-3-319-12068-3 doi: 10.1007/978-3-319-12069-0-2.
- Le, T., Stahl, F., Gomes, J.B., Gaber, M.M., and Di Fatta, G. (in Press) *An Adaptive and Fast Rule-based Classification Algorithm for Streaming Data*. Expert Systems, Wiley.

Source Code:

RESOURCES (CONT.)

Source Code:

G-eRules on Github:

https://github.com/thienle2401/G-eRules

Generalised Rule Induction on Github:

https://github.com/thienle2401/GeneralisedRulesAlgorithm

QUESTIONS?

University of Reading | 1926 2016
NINETY YEARS OF EXCELLENCE

Links & References

[1] http://www.extremetech.com/wp-content/uploads/2015/07/NeuralNetwork.png

[2] http://cdn.edureka.co/blog/wp-content/uploads/2015/01/tree2.png

[3] Gama, J., Kosina, P. (2011). Learning Rules From Data Streams, In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence , pp. 1255-1260, ACM.

[4] Domingos, P., & Hulten, G. (2000). Mining high-speed data streams. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining , pp. 71-80, ACM.

[5] J. Cendrowska. PRISM: an algorithm for inducing modular rules. International Journal of Man-Machine Studies, 27(4):349–370, 1987.

[6] http://moa.cms.waikato.ac.nz/

[7] Tschabitscher, H. (2014). Ever Wondered How Many Emails Get Sent Worldwide Every Day? Here's the Answer. [online] About. Available at: http://email.about.com/od/emailtrivia/f/emails_per_day.htm [Accessed 1 Dec. 2014].

[8] Noyes, A. and Noyes, D. (2014). The Top 20 Valuable Facebook Statistics - Updated October 2014 - Zephoria Inc.. [online] Zephoria Inc. Available at: https://zephoria.com/social-media/top-15-valuable-facebook-statistics/ [Accessed 1 Dec. 2014].

Contact Details

https://fredericstahl.wordpress.com/

@fred_stahl

F.T.Stahl@reading.ac.uk