Adaptive Model Rules from Data Streams [Regresión]

Decision rules are one of the most expressive languages for machine learning. In this paper we **present Adaptive Model Rules (AMRules), the first streaming rule learning algorithm for regression problems**. **In AMRules the antecedent of a rule is a conjunction of conditions on the attribute values, and the consequent is a linear combination of attribute values. Each rule uses a Page-Hinkley test to detect changes in the process generating data and react to changes by pruning the rule set.** In the experimental section we report the results of AMRules on benchmark regression problems, and compare the performance of our system with other streaming regression algorithms.

Apartado 4.1 Experimentación

All our algorithms were implemented in java using the **Massive Online Analysis (MOA)** data stream software suite [2]. The performance of the algorithms is measured using the standard metrics for regression problems: **Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)** [24]. We used two evaluation methods. **When no concept drift is assumed, the evaluation method we employ uses the traditional train and test scenario.**

The experimental datasets include both artificial and real data, as well sets with continuous attributes. We use ten regression datasets from the **UCI Machine Learning** Repository [1] and other sources:

Table 1. Summary of datasets

Datasets	# Instances	# Attributes
2dplanes	40768	11
Ailerons	13750	41
Puma8NH	8192	9
Puma32H	8192	32
Pol	15000	49
Elevators	8752	19
Fried	40769	11
bank8FM	8192	9
kin8nm	8192	9
Airline	115Million	11

Comparison with Other State-of-the-Art Regression Algorithms. We compared AMRules with other non-incremental regression algorithms from **WEKA** [9]. We use the **standard method of ten-fold cross-validation**, using the same folds for all the algorithms included.

6 citas (2013) Capítulo de libro. Joao Gama
https://link.springer.com/chapter/10.1007/978-3-642-40988-2_31

***Evaluation methods and decision theory for classification of streaming data with temporal dependence

Predictive modeling on data streams plays an important role in modern data analysis, where data arrives continuously and needs to be mined in real time. In the stream setting the data distribution is often evolving over time, and models that update themselves during operation are becoming the state-of-the-art. **This paper formalizes a learning and evaluation scheme of such predictive models.** We theoretically **analyze evaluation of classifiers on streaming data with temporal dependence.** Our findings suggest that the commonly accepted data stream classification measures, such as classification accuracy and Kappa statistic, fail to diagnose cases of poor performance when temporal dependence is present, therefore they should not be used as sole performance indicators. Moreover, **classification accuracy can be misleading if used as a proxy for evaluating change detectors with datasets that have temporal dependence.** We formulate the decision theory for streaming data classification with temporal dependence and develop a new evaluation methodology for data stream classification that takes temporal dependence into account. We propose a combined measure for classification performance, that takes into account temporal dependence, and we recommend using it as the main performance measure in classification of streaming data.

...

We experiment with four real datasets often used in evaluating data stream classification.

1. The Electricity dataset (Elec2) (Harries 1999)
2. The Forest Covertype (Cover) (Bache and Lichman 2013)
3. The KDD cup intrusion detection dataset (KDD99) (Bache and Lichman 2013)
4. The Ozone dataset (Ozone) (Bache and Lichman 2013)

Table 4
Classifiers used in the experiments

	Adaptation	Base classifier	Number of models
Naive Bayes (NB)	Non-adaptive	Naive Bayes	One
Hoeffding tree (HT) Domingos and Hulten (2000)	Non-adaptive	Hoeffding tree	One
Drift detection (DDM) Gama et al. (2004)	Adaptive	Naive Bayes	One
Hoeffding adaptive tree (HAT) Bifet and Gavalda (2009)	Adaptive	Hoeffding tree	One
Leveraged bagging (LBAG) Bifet et al. (2010)	Adaptive	Hoeffding tree	Ensemble

Table 5
Accuracies of adaptive classifiers on the Electricity dataset reported in the literature

Algorithm name	Accuracy (%)	Reference
DDM	89.6[a]	Gama et al. (2004)
Learn++.CDS	88.5	Ditzler and Polikar (2013)
KNN-SPRT	88.0	Ross et al. (2012)
GRI	88.0	Tomczak and Gonczarek (2013)
FISH3	86.2	Zliobaite (2011)
EDDM-IB1	85.7	Baena-Garcia et al. (2006)

We compare the **accuracies of five intelligent classifiers (NB, DDM, HT, HAT, LBAG) with two established baselines Majority Class and Persistent classifiers**, which give important indications about the performance of intelligent classifiers with respect to class imbalance and temporal dependence in the data

19 citas (2014)
https://link.springer.com/article/10.1007/s10994-014-5441-4

Online tree-based ensembles and option trees for regression on evolving data streams [Regresión]

The emergence of ubiquitous sources of streaming data has given rise to the popularity of algorithms for online machine learning. In that context, **Hoeffding trees represent the state-of-the-art algorithms for online classification.** Their popularity stems in large part from **their ability to process large quantities of data with a speed that goes beyond the processing power of any other streaming or batch learning algorithm**. As a consequence, Hoeffding trees have often been used as base models of many ensemble learning algorithms for online classification. However, despite the existence of many algorithms for online classification, **ensemble learning algorithms for online regression do not exist.** In particular, the field of **online any-time regression analysis seems to have experienced a serious lack of attention**. In this paper, we address this issue through a study and an **empirical evaluation of a set of online algorithms for regression, which includes the baseline Hoeffding-based regression trees, online option trees, and an online least mean squares filter.** We also design, implement and evaluate two novel ensemble learning methods for online regression: online bagging with Hoeffding-based model trees, and an online RandomForest method in which we have used a randomized version of the online model tree learning algorithm as a basic building block. Within the study presented in this paper, we **evaluate** the proposed algorithms along several dimensions: **predictive accuracy and quality of models, time and memory requirements, bias–variance and bias–variance–covariance decomposition of the error, and responsiveness to concept drift.**

Table 1. The tree-based algorithms for online regression used in the empirical evaluation.

Algorithm	Description
LMS	Least mean squares adaptive filter
FIMT-DD, [6]	Fast and incremental model tree with drift detection
ORTO-A, [8]	Online option tree with averaging
ORTO-BT, [8]	Online option tree with best model's prediction
OBag	Online Bagging of FIMT-DD trees
ORF	Online Random Forest of FIMT-DD trees

Apartado 5.4. Datasets

Use all the available regression datasets of sufficient size, as we are interested in studying how the learning methods behave in different scenarios. From the **UCI Machine Learning Repository** [32], the **Delve Repository** and the **StatLib System′s site**, **we chose the nine largest available "benchmark" datasets on which no concept drift has been reported**. We **also** evaluate the studied approaches on **three real-world datasets**, described below:

- Infobiotics PSP data sets: The PSP repository contains real-world tunable scalability benchmark datasets for protein structure prediction.
- City traffic data set: This dataset was generated for the data mining competition that was part of IEEE ICDM 2010.
- Airline data set: The last dataset represents a non-stationary real-world problem and was created using the data from the Data Expo competition (2009)

Table 2. Characteristics of the datasets used for the experimental evaluation.

Dataset name	Size	Number of attributes		Values of the target
		Nominal	Numerical	Mean ±Std.
Abalone	4177	1	8	9.934 ±3.224
Cal housing	20,640	0	9	206855.817 ±115395.616
Elevators	16,599	0	19	0.022 ±0.007
House 8L	22,784	0	9	50074.44 ±52843.476
House 16H	22,784	0	17	50074.44 ±52843.476
Mv Delve	40969	3	8	−8.856 ±10.420
Pol	15,000	0	49	28.945 ±41.726
Wind	6574	2	13	15.599 ±6.698
Winequality	4898	0	12	5.878 ±0.886

Muchas tablas comparativas de resultados
0 citas (2014) Joao Gama
https://www.sciencedirect.com/science/article/pii/S0925231214012338

***A survey on data stream clustering and classification

Nowadays, with the advance of technology, many applications generate huge amounts of data streams at very high speed. Examples include network traffic, web click streams, video surveillance, and sensor networks. Data stream mining has become a hot research topic. Its goal is to extract hidden knowledge/patterns from continuous data streams. Unlike traditional data mining where the dataset is static and can be repeatedly read many times, **data stream mining algorithms face many challenges and have to satisfy constraints such as bounded memory, single-pass, real-time response, and concept-drift detection.** This paper presents a comprehensive **survey of the state-of-the-art data stream mining algorithms with a focus on clustering and classification because of their ubiquitous usage**. It identifies mining constraints, proposes a general model for data stream mining, and depicts the relationship between traditional data mining and data stream mining. Furthermore, it analyzes the advantages as well as limitations of data stream algorithms and suggests potential areas for future research.
...
There is some useful, **open-source software for data stream mining research**:

- WEKA
- MOA

● RapidMiner

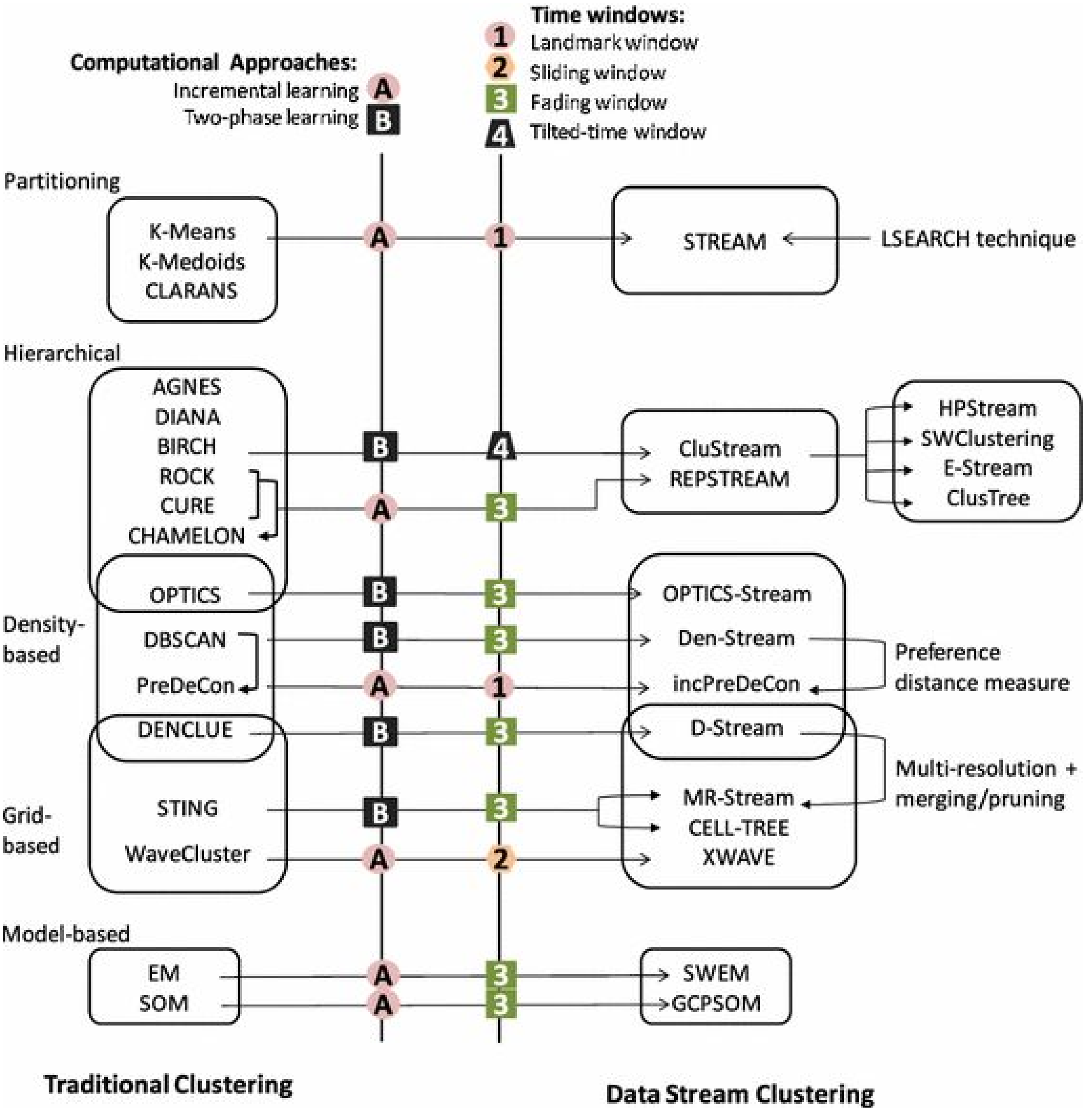
...

Some **synthetic datasets** with unlimited number of examples are created, for example, Random Tree Generator [35], SEA Concepts Generator [90], and Rotating Hyperplane [53]. These data generators are implemented in the MOA software package [17].

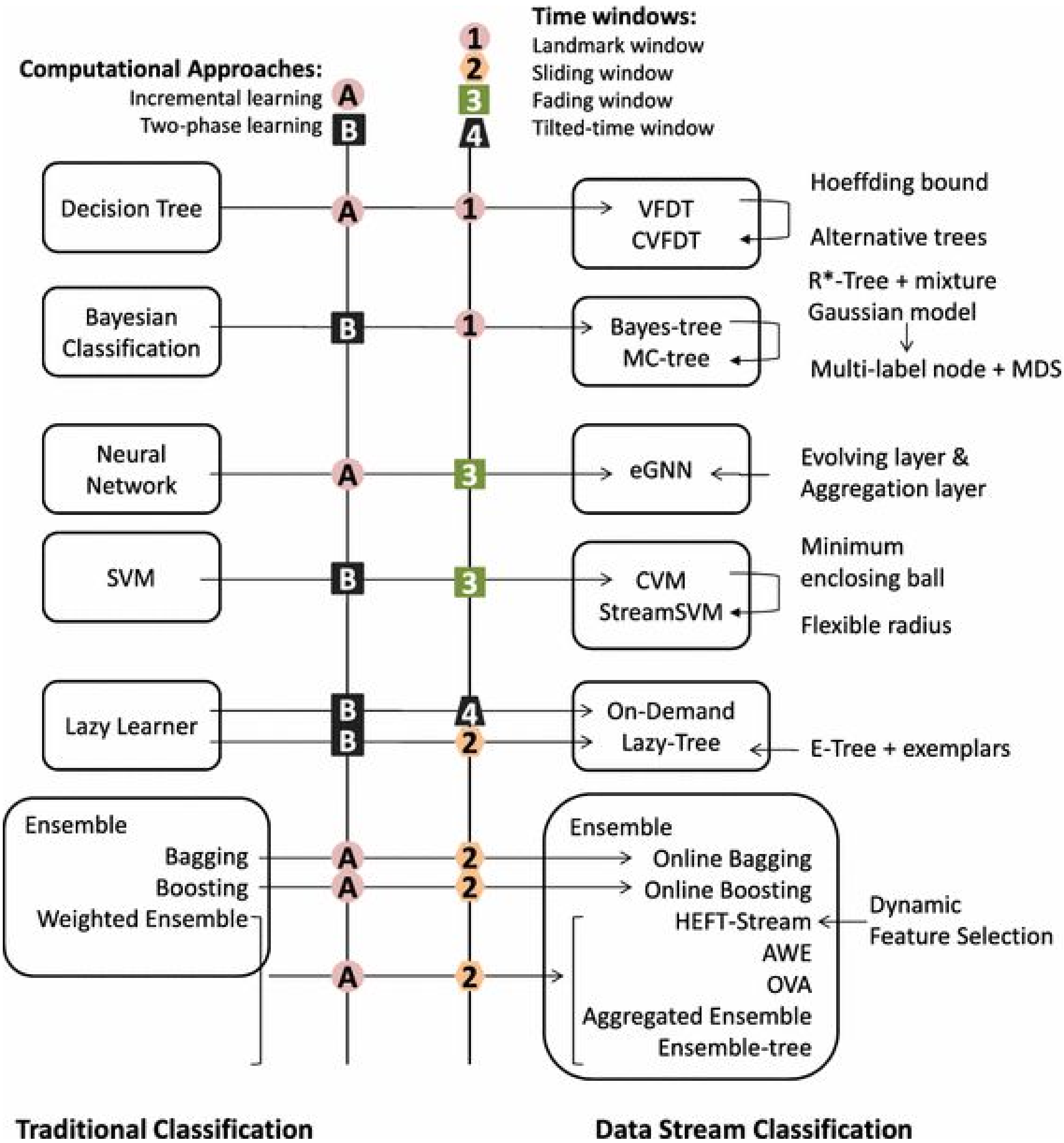
We also recommend some (**real datasets**) data repositories with large datasets for stream evaluation:

● UCI Machine Learning Repository: Forest Covertype, Poker-Hand, and Electricity
● KDD Cup Center: The computer network intrusion dataset (KDD'99)

Clustering



Classification



28 citas (2014)
https://link.springer.com/article/10.1007/s10115-014-0808-1

A similarity-based approach for data stream classification

Solo pongo esta tabla comparativa de los datasets que utilizan:

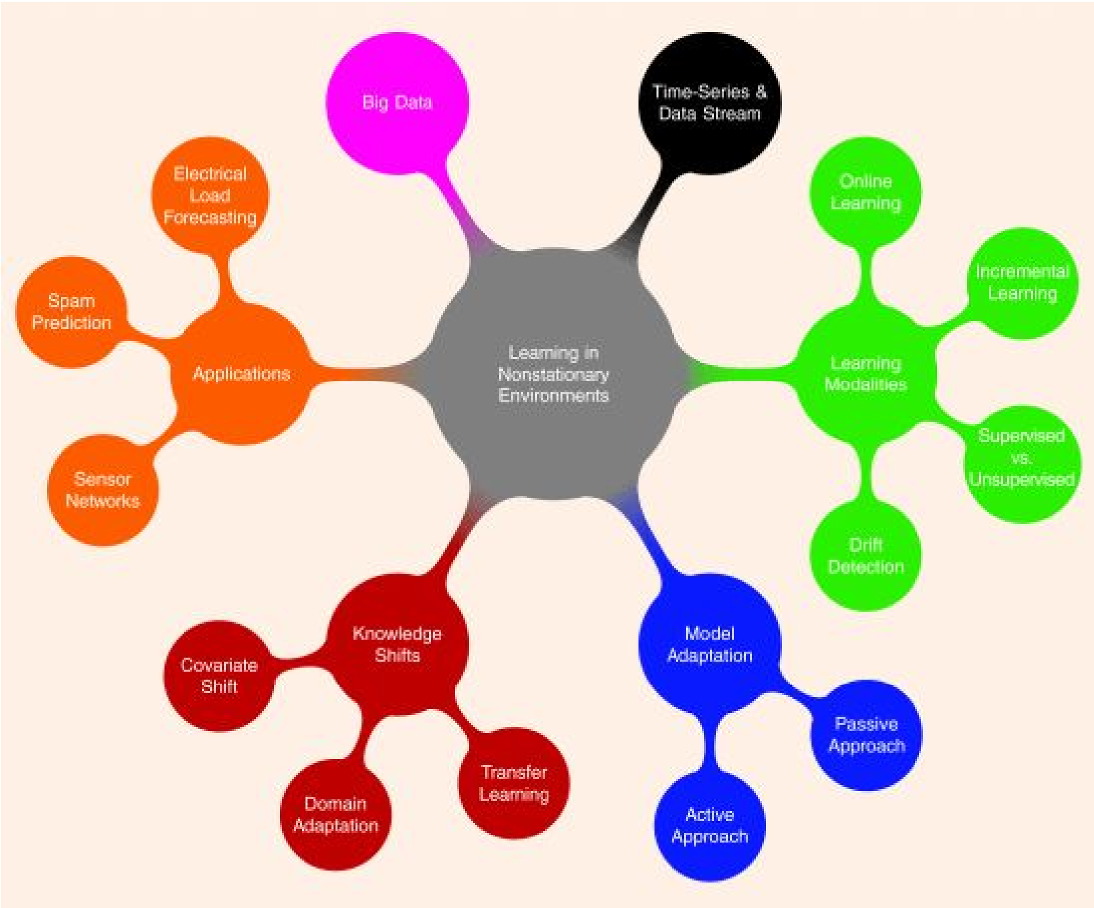
Table 1. Data streams.

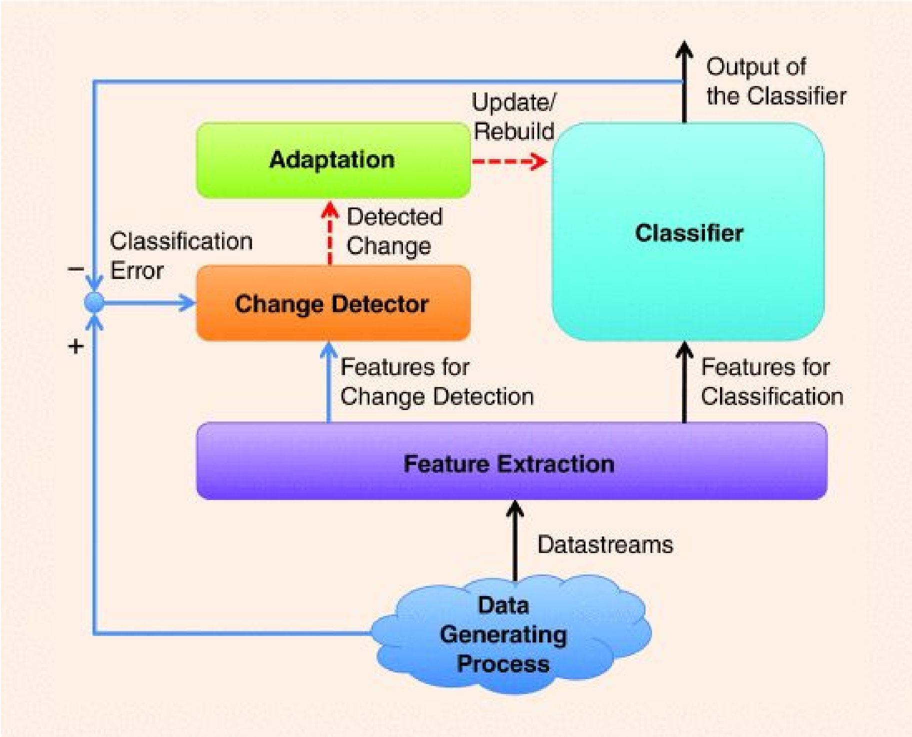
Data streams	No. attribute	Type	Class	No. instances
MOA				
Agrawal	9	Mix	2	20,000
LED24	24	Nom	10	20,000
RDG1	10	Nom	2	20,000
RandomRBF	10	Num	2	20,000
RandomTree	10	Mix	2	100,000
Stagger	3	Nom	2	100,000
Real				
Airlines	7	Mix	2	539,383
Bank	16	Mix	2	45,210
Poker Hand	10	Mix	10	25,010
UCI				
Annealing	39	Mix	5	20,000
Balance	5	Num	3	20,000
Breast	10	Num	6	20,000
Car	7	Nom	4	20,000
Pages-bloks	11	Num	5	20,000
Pen Digits	17	Num	10	20,000
Solar-Flare	13	Nom	6	20,000
Vehicle	19	Num	4	20,000
Vowel	14	Mix	11	20,000

0 citas (2014) (Expert Systems with Applications)
https://www.sciencedirect.com/science/article/pii/S0957417413010300

***Learning in Nonstationary Environments: A Survey

The prevalence of mobile phones, the internet-of-things technology, and networks of sensors has led to an enormous and ever increasing amount of data that are now more commonly available in a streaming fashion [1]-[5]. Often, it is assumed - either implicitly or explicitly - that the process generating such a stream of data is stationary, that is, the data are drawn from a fixed, albeit unknown probability distribution. **In many real-world scenarios,** however, such an assumption is simply not true, and **the underlying process generating the data stream is characterized by an intrinsic nonstationary (or evolving or drifting) phenomenon.** The nonstationarity can be due, for example, to seasonality or periodicity effects, changes in the users' habits or preferences, hardware or software faults affecting a cyber-physical system, thermal drifts or aging effects in sensors. **In such nonstationary environments, where the probabilistic properties of the data change over time, a non-adaptive model trained under the false stationarity assumption is bound to become obsolete in time, and perform sub-optimally at best, or fail catastrophically at worst.**





Active approach for learning a classifier in nonstationary environments.

Open Source Software and Available Benchmarks

- Hierarchical ICI-based Change-Detection Tests (Matlab)
  http://home.deib.polimi.it/boracchi/Projects/HierarchicalICI-basedCDT.html
- Learn++.NSE (Matlab) https://github.com/gditzler/IncrementalLearning
- Massive Online Analysis (Java) http://moa.cms.waikato.ac.nz/
- Scalable Advanced Massive Online Analysis (Java)
  http://jmlr.org/papers/v16/morales15a.html
- Online Nonstationary Boosting (Java)
  http://www.cs.man.ac.uk/+pococka4/ONSBoost.html

The following **datasets and code for generating datasets are commonly used for assessing the performances of proposed concept drift algorithms.**

- Minku & Yao's Concept Drift Generator (Matlab)
  http://www.cs.bham.ac.uk/+minkull/opensource.html
- Kuncheva's Concept Drift Generator (Matlab)
  http://pages.bangor.ac.uk/+mas00a/EPSRC_simulation_framework/changing_environments_stage1a.htm
- Airlines Flight Delay Prediction
  http://sourceforge.net/projects/moa-datastream/files/Datasets/Classification/airlines.arff.zip
- Spam Classification http://www.comp.dit.ie/sjdelany/Dataset.htm
- Chess.com https://sites.google.com/site/zliobaite/resources-1
- KDD Cup 1999 http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

- POLIMI Rock Collapse and Landslide Forecasting
  http://roveri.faculty.polimi.it/software-and-datasets

More software and data
http://github.com/gditzler/ConceptDriftResources and
http://roveri.faculty.polimi.it/software-and-datasets.

64 citas (2015)
http://0-ieeexplore.ieee.org.fama.us.es/document/7296710/

Multi-target regression from high-speed data streams with adaptive model rules [Regresión]

Many real life prediction problems involve predicting a structured output. Multi-target regression is an instance of structured output prediction whose task is to predict for multiple target variables. Structured output algorithms are usually computationally and memory demanding, hence are not suited for dealing with massive amounts of data. Most of these algorithms can be categorized as local or global methods. Local methods produce individual models for each output component and combine them to produce the structured prediction. Global methods adapt traditional learning algorithms to predict the output structure as a whole. **We propose the first rule-based algorithm for solving multi-target regression problems from data streams. The algorithm builds on the adaptive model rules framework.** In contrast to the majority of the structured output predictors, **this particular algorithm does not fall into the local and global categories. Instead, each rule specializes on related subsets of the output attributes. To evaluate the performance of the proposed algorithm, two other rule-based algorithms were developed, one using the local strategy and the other using the global strategy. These methods were compared considering their prediction error, memory usage, computational time, and model complexity.** Experimental results on synthetic and real data show that the local-strategy algorithm usually obtains the lowest error. However, the proposed and the global-strategy algorithms use much less memory and run significantly much faster at the cost of a slightly increase in the error, which make them very attractive when computation resources are an important factor. Also, the models produced by the latter approaches are much easier to understand since considerably less rules are produced.

Datasets

Datasets	#Examples	#Numeric Inputs	#Nominal Inputs	#Outputs
Bicycle	17,377	12	0	3
Eunite	8,064	29	0	5
2Dplanes	256,000	20	0	8
FriedD	256,000	10	0	4
FriedAsyncD	256,000	10	0	4
MV	256,000	15	5	9
Airline	15,235,031	6	3	6
RF1	9,005	64	0	8
RF2	7,679	576	0	8
SCM1d	9,803	280	0	16
SCM20d	8,966	61	0	16

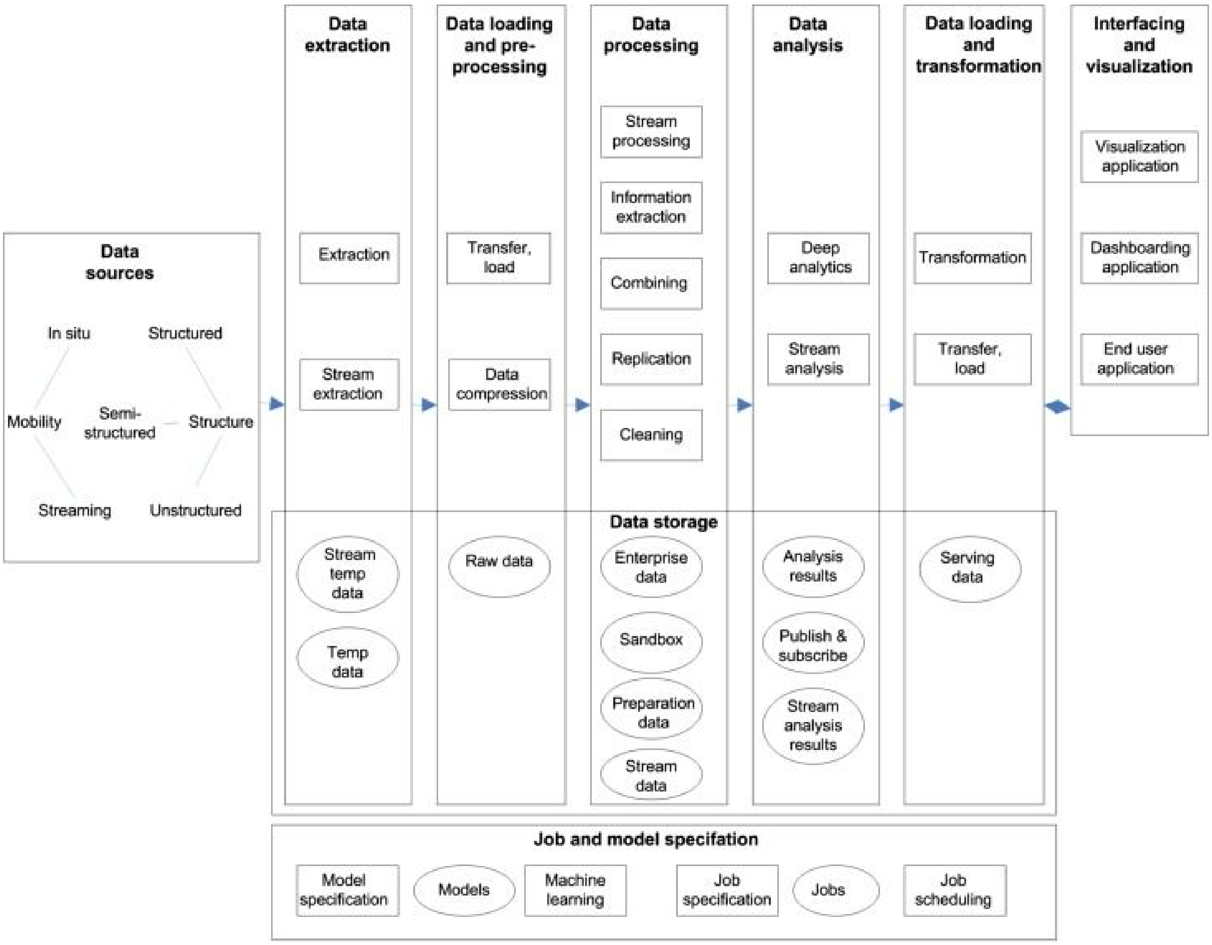
Algunos resultados

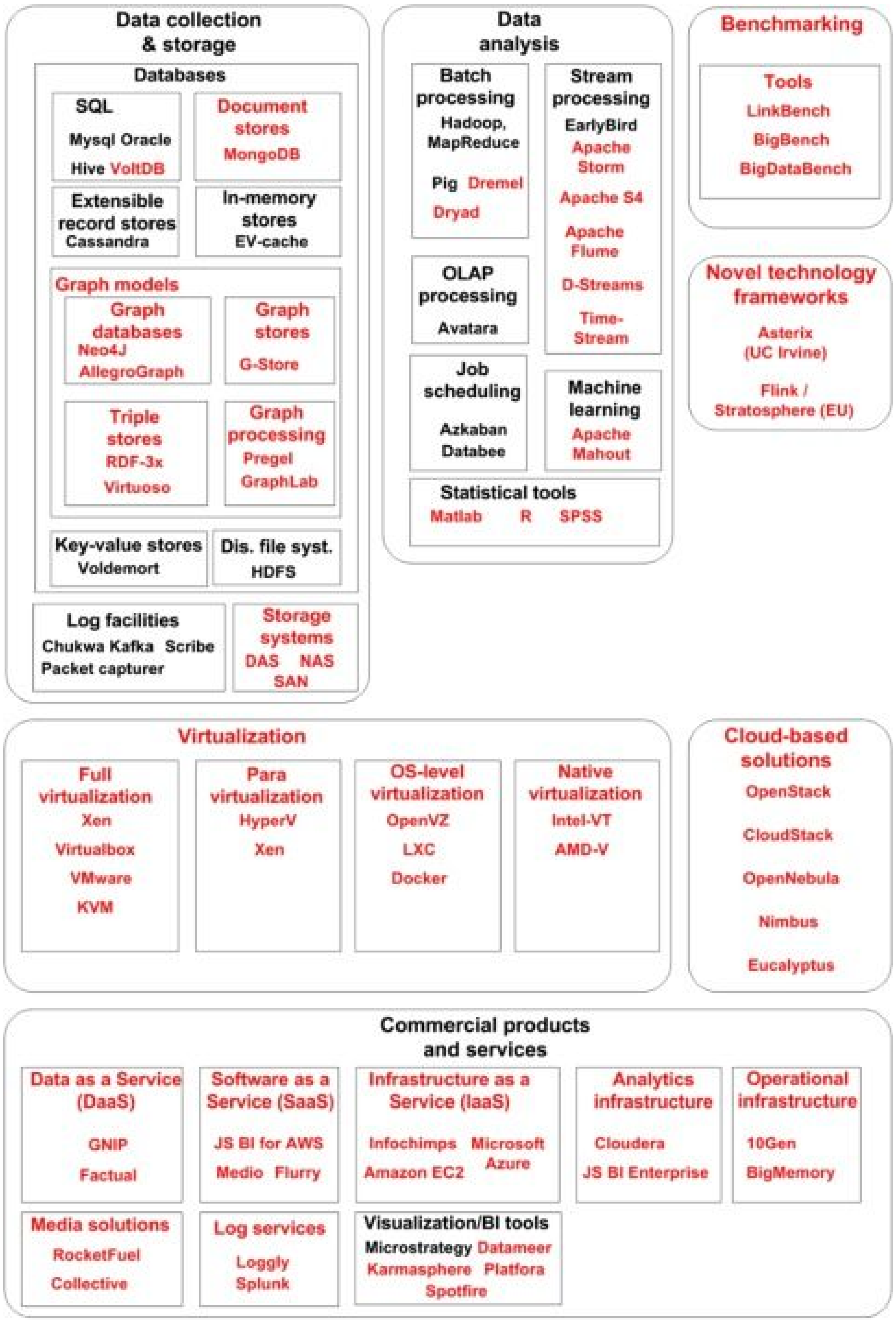
Fig. 4. Average relative mean absolute errors.

0 citas (Diciembre 2015). Joao Gama
http://0-ieeexplore.ieee.org.fama.us.es/xpls/icp.jsp?arnumber=7344900

Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems

Many business cases exploiting big data have been realised in recent years; Twitter, LinkedIn, and Facebook are examples of companies in the social networking domain. Other big data use cases have focused on capturing of value from streaming of movies (Netflix), monitoring of network traffic, or improvement of processes in the manufacturing industry. Also, implementation architectures of the use cases have been published. However, conceptual work integrating the approaches into one coherent reference architecture has been limited. **The contribution of this paper is technology independent reference architecture for big data systems, which is based on analysis of published implementation architectures of big data use cases. An additional contribution is classification of related implementation technologies and products/services, which is based on analysis of the published use cases and survey of related work.** The reference architecture and associated classification are aimed for facilitating architecture design and selection of technologies or commercial solutions, when constructing big data systems.





Incluye esquemas de las infraestructuras BigData de Linkedin, Facebook, Twitter, etc.

0 citas (2015)
https://www.sciencedirect.com/science/article/pii/S2214579615000027

The Great Time Series Classification Bake Off: An Experimental Evaluation of Recently Proposed Algorithms. Extended Version

In the last five years there have been a large number of new time series classification algorithms proposed in the literature. These algorithms have been evaluated on subsets of the 47 data sets in the University of California, Riverside time series classification archive. The archive has recently been expanded to 85 data sets, over half of which have been donated by researchers at the University of East Anglia. Aspects of previous evaluations have made comparisons between algorithms difficult. For example, several different programming languages have been used, experiments involved a single train/test split and some used normalised data whilst others did not. The relaunch of the archive provides a timely opportunity to thoroughly evaluate algorithms on a larger number of datasets. **We have implemented 18 recently proposed algorithms in a common Java framework and compared them against two standard benchmark classifiers (and each other) by performing 100 resampling experiments on each of the 85 datasets. We use these results to test several hypotheses relating to whether the algorithms are significantly more accurate than the benchmarks and each other. Our results indicate that only 9 of these algorithms are significantly more accurate than both benchmarks and that one classifier, the Collective of Transformation Ensembles, is significantly more accurate than all of the others.** All of our experiments and results are reproducible: we release all of our code, results and experimental details and we hope these experiments form the basis for more rigorous testing of new algorithms in the future.

Código fuente disponible en: http://timeseriesclassification.com/ [Muy completa]
2016
https://arxiv.org/abs/1602.01711

Active Learning Classifier for Streaming Data

This work reports the research on active learning approach applied to the data stream classification. The chosen characteristics of the proposed frameworks were evaluated on the basis of the wide range of computer experiments carried out on the three benchmark data streams. Obtained results confirmed the usability of proposed method to the data stream classification with the presence of incremental concept drift.

Experiments: The experiments were carried out for the 3 artificially generated data streams available in MOA (WaveformGeneratorDrift, RandomTreeGenerator, LEDGeneratorDrift) and 3 online learners as minimal distance classifier (k-NN), Naïve Bayes and Perceptron.

0 citas (2016)
https://link.springer.com/chapter/10.1007%2F978-3-319-32034-2_16

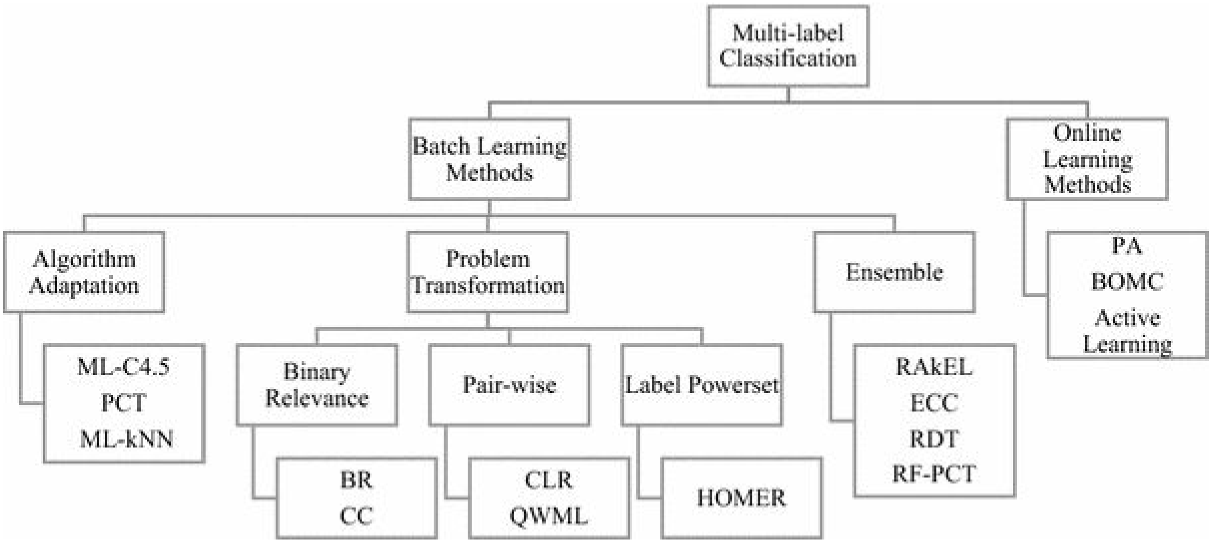
Ensemble learning for data stream analysis: A survey

In many applications of information systems learning algorithms have to act in **dynamic environments where data are collected in the form of transient data streams**. Compared to static data mining, processing streams imposes new computational requirements for algorithms to **incrementally process incoming examples while using limited memory and time**. Furthermore, due to the non-stationary characteristics of streaming data, **prediction models are often also required to adapt to concept drifts**. Out of several new proposed stream algorithms, ensembles play an important role, in particular for non-stationary environments. **This paper surveys research on ensembles for data stream classification as well as regression tasks**. Besides presenting a comprehensive spectrum of ensemble approaches for data streams, we also discuss advanced learning concepts such as imbalanced data streams, novelty detection, active and semi-supervised learning, complex data representations and structured outputs. The paper concludes with a discussion of open research problems and lines of future research.

Tablas comparativas
0 citas (2017) (Information Fusion)
https://www.sciencedirect.com/science/article/pii/S1566253516302329

***A novel online multi-label classifier for high-speed streaming data application

In this paper, a **high-speed online neural network classifier based on extreme learning machines for multi-label classification is proposed**. In multi-label classification, each of the input data sample belongs to one or more than one of the target labels. The traditional binary and multi-class classification where each sample belongs to only one target class forms the subset of multi-label classification. Multi-label classification problems are far more complex than binary and multi-class classification problems, as both the number of target labels and each of the target labels corresponding to each of the input samples are to be identified. **The proposed work exploits the high-speed nature of the extreme learning machines to achieve real-time multi-label classification of streaming data**. A new threshold-based online sequential learning algorithm is proposed for high speed and streaming data classification of multi-label problems. The proposed method is experimented with **six different datasets from different application domains such as multimedia, text, and biology. The hamming loss, accuracy, training time and testing time of the proposed technique is compared with nine different state-of-the-art methods.** Experimental studies shows that the proposed technique outperforms the existing multi-label classifiers in terms of performance and speed.



Multi-label methods

Table 1
Dataset specifications

Dataset	Domain	No. of features	No. of samples	#Train	#Test	No. of labels	LC	LD
Yeast	Biology	103	2417	1600	817	14	4.24	0.3
Scene	Multimedia	294	2407	2000	407	6	1.07	0.17
Corel5k	Multimedia	499	5000	4500	500	374	3.52	0.0
Enron	Text	1001	1702	1200	502	53	3.38	0.0
Medical	Text	1449	978	700	278	45	1.25	0.0

Table 2
Methods used for comparison

Method name	Method category	Machine learning category
Classifier chain (CC)	PT	SVM
QWeighted approach for multi-label learning (QWML)	PT	SVM
Hierarchy of multi-label ClassifiERs (HOMER)	PT	SVM
Multi-label C4.5 (ML-C4.5)	AA	Decision trees
Predictive clustering trees (PCT)	AA	Decision trees
Multi-label k-nearest neighbors (ML-kNN)	AA	Nearest neighbors
Ensemble of classifier chains (ECC)	EN	SVM
Random forest predictive clustering trees (RF-PCT)	EN	Decision trees
Random forest of ML-C4.5 (RFML-C4.5)	EN	Decision trees

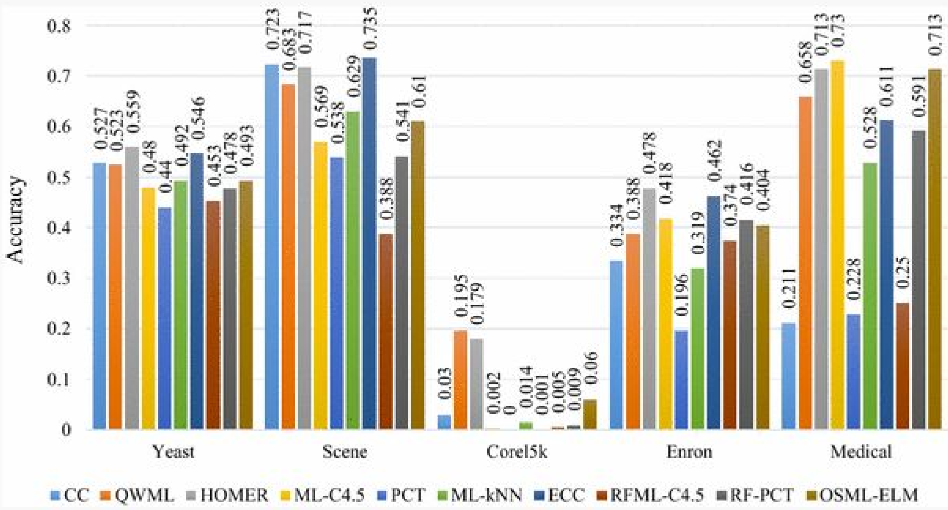


Fig. 3
Comparison of accuracy metric with state-of-the-art techniques

The proposed OSML-ELM classifier outperforms the existing state-of-the-arts multi-label classification techniques in terms of speed and performance.

Tablas comparativas de resultados muy completas
1 cita (Diciembre 2017)
https://link.springer.com/article/10.1007%2Fs12530-016-9162-8

Scalable real-time classification of data streams with concept drift

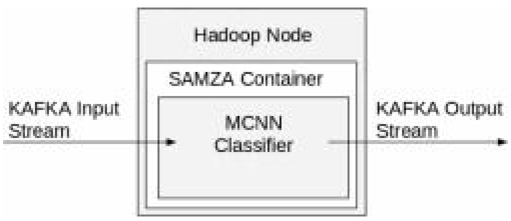
Inducing adaptive predictive models in real-time from high throughput data streams is one of the most challenging areas of Big Data Analytics. The fact that data streams may contain concept drifts (changes of the pattern encoded in the stream over time) and are unbounded, imposes unique challenges in comparison with predictive data mining from batch data. **Several real-time predictive data stream algorithms exist, however, most approaches are not naturally parallel and thus limited in their scalability. This paper highlights the Micro-Cluster Nearest Neighbour (MC-NN) data stream classifier. MC-NN is based on statistical summaries of the data stream and a nearest neighbour approach, which makes MC-NN naturally parallel. In its serial version MC-NN is able to handle data streams, the data does not need to reside in memory and is processed incrementally. MC-NN is also able to adapt to concept drifts.** This paper provides an empirical study on the serial algorithm's speed, adaptivity and accuracy. Furthermore, this paper discusses the new parallel implementation of MC-NN, its parallel properties and provides an empirical scalability study.

All synthetic data generators and algorithms evaluated in Section 3.2 were taken from the **Massive Online Analysis (MOA) framework** [38]. Also the here presented real-time KNN and MC-NN algorithms were implemented within the MOA framework.

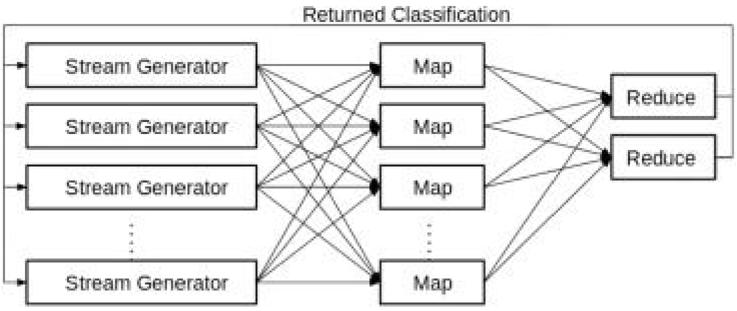
Four data streams have been utilised:

- SEA data stream
- Random Tree Generator
- Hyperplane generator
- Human Activity Recognition (HAR)

The architecture is based on integrating various distributed Open Source technologies such as **Hadoop, Samza and Kafka**. The paper describes the use of these technologies for parallel data stream processing and highlights issues and experiences.



Architecture of a MC-NN computer cluster node.



Cluster setup

0 citas (2018)
https://www.sciencedirect.com/science/article/pii/S0167739X17304685

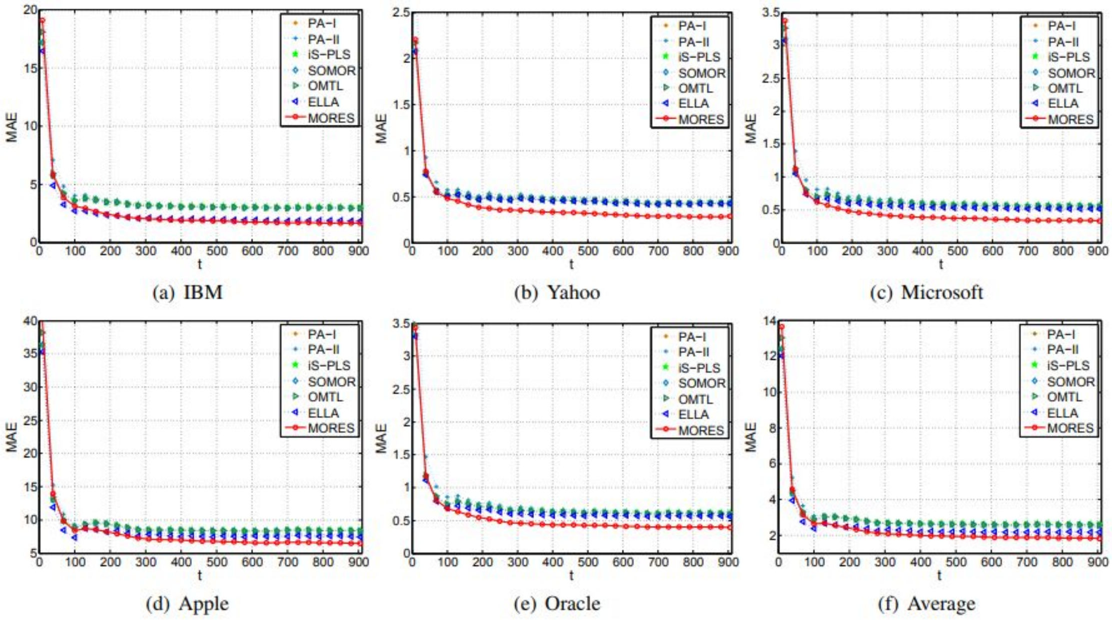
***Dynamic Structure Embedded Online Multiple-Output Regression for Streaming Data [Regresión]

Online multiple-output regression is an important machine learning technique for modeling, predicting, and compressing multi-dimensional correlated data streams. In this paper, we propose a **novel online multiple-output regression method, called MORES, for streaming data**. **MORES can dynamically learn the structure of the regression coefficients to facilitate the model's continuous refinement**. Considering that limited expressive ability of regression models often leading to residual errors being dependent, MORES intends to dynamically learn and leverage the structure of the residual errors to improve the prediction accuracy. Moreover, we introduce three modified covariance matrices to extract necessary information from all the seen data for training, and set different weights on samples so as to track the data streams' evolving characteristics. Furthermore, an efficient algorithm is designed to optimize the proposed objective function, and an efficient online eigenvalue decomposition algorithm is developed for the modified covariance matrix. Finally, we analyze the convergence of MORES in certain ideal condition. **Experiments on two synthetic datasets and three real-world datasets validate the effectiveness and efficiency of MORES. In addition, MORES can process at least 2,000 instances per second (including training and testing) on the three real-world datasets, more than 12 times faster than the state-of-the-art online learning algorithm.**

To evaluate the performance of MORES, we perform the experiments on both **one synthetic dataset and three real-world datasets: the stock price prediction dataset, the Barrett WAM dataset [40], and the weather dataset [41]**. We compare MORES with an online multiple-output regression method, iS-PLS [19], two online multitask learning methods, ELLA2 [30] and OMTL3 [13]. We also compare with two variants of PA algorithm in [26] called PA-I and PA-II, which are two classical online learning approaches for single regression tasks.

Apartado 3.2

Following previous studies in [44] and [20], **we apply our algorithms to stock data of companies for price prediction.** We choose the **daily stock price** data of five companies including **IBM, Yahoo, Microsoft, Apple, and Oracle** in the period from 2010 to 2013. The learned model can predict the stock prices in the future by using the stock prices in the past as inputs.



. MAEs of different approaches as a function of the number of samples t on different companies.

Mucha (muchísima) matemática

0 citas (2018)

http://0-ieeexplore.ieee.org.fama.us.es/document/8260965/

Otros datasets usados en algunos artículos

Comparison of Tree-Based Methods for Multi-target Regression on Data Streams. 2016

https://link.springer.com/chapter/10.1007/978-3-319-39315-5_2

Table 1.
Datasets used in the experiments and their properties. N – number of instances, T – number of targets.

Dataset	Domain	N	Input attr.	T
Bicycles [6]	service prediction	17379	12 numeric	3
EUNITE03	quality prediction	8064	29 numeric	5
Forestry Kras [18]	vegetation prediction	60607	160 numeric	11
Forestry Slivnica [19]	vegetation prediction	6218	149 numeric	2
RF1 [21]	environmental prediction	9005	64 numeric	8
RF2 [21]	environmental prediction	7679	575 numeric	8
SCM1d [21]	price prediction	9803	280 numeric	16
SCM20d [21]	price prediction	8966	61 numeric	16