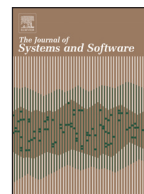




Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Data stream classification using random feature functions and novel method combinations

Diego Marrón^{a,*}, Jesse Read^{b,c}, Albert Bifet^c, Nacho Navarro^a^a Department of Computer Architecture, Universitat Politècnica de Catalunya and with the Department of Computer Science, Barcelona Supercomputing Center, Spain^b Aalto University and HIIT, Finland^c LTCI, CNRS, Télécom Paris Tech

ARTICLE INFO

Article history:

Received 30 September 2015

Revised 29 March 2016

Accepted 4 June 2016

Available online xxx

Keywords:

Data stream mining

Big data

Classification

GPUs

ABSTRACT

Big Data streams are being generated in a faster, bigger, and more commonplace. In this scenario, Hoeffding Trees are an established method for classification. Several extensions exist, including high-performing ensemble setups such as online and leveraging bagging. Also, k -nearest neighbors is a popular choice, with most extensions dealing with the inherent performance limitations over a potentially-infinite stream.

At the same time, gradient descent methods are becoming increasingly popular, owing in part to the successes of deep learning. Although deep neural networks can learn incrementally, they have so far proved too sensitive to hyper-parameter options and initial conditions to be considered an effective 'off-the-shelf' data-streams solution.

In this work, we look at combinations of Hoeffding-trees, nearest neighbor, and gradient descent methods with a streaming preprocessing approach in the form of a random feature functions filter for additional predictive power.

We further extend the investigation to implementing methods on GPUs, which we test on some large real-world datasets, and show the benefits of using GPUs for data-stream learning due to their high scalability.

Our empirical evaluation yields positive results for the novel approaches that we experiment with, highlighting important issues, and shed light on promising future directions in approaches to data-stream classification.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

There is a trend towards working with big and dynamic data sources. This tendency is clear both in real world applications and the academic literature. Many modern data sources are not only dynamic but often generated at high speed and must be classified in real time. Such contexts can be found in sensor applications (e.g., tracking and activity monitoring), demand prediction (e.g., of electricity), manufacturing processes, robotics, email, news feeds, and social networks. Real-time analysis of data streams is becoming a key area of data mining research as the number of applications in this area grows.

The requirements for a classifier in a data stream are to

- Be able to make a classification at any time
- Deal with a potentially infinite number of examples
- Access each example in the stream just once

These requirements can in fact be met by variety of learning schemes, including even batch learners (e.g., Qu et al., 2009), where batches are constantly gathered over time, and newer models replace older ones as memory fills up. Nevertheless, incremental methods remain strongly preferred in the data streams literature, and particularly the Hoeffding tree (HT) and its variations (Domingos and Hulten, 2000; Bifet et al., 2010b), k -nearest neighbors (k NN) (Shaker and Hüllermeier, 2012). Support for these options is given by large-scale empirical comparisons (Read et al., 2012), where it is also found that methods such as naive Bayes and stochastic gradient descent-based (SGD) are relatively poor performers.

* Corresponding author.

E-mail addresses: dmarron@ac.upc.edu (D. Marrón), jesse.read@telecom-paristech.fr (J. Read), albert.bifet@telecom-paristech.fr (A. Bifet), nacho@ac.upc.edu (N. Navarro).

<http://dx.doi.org/10.1016/j.jss.2016.06.009>

0164-1212/© 2016 Elsevier Inc. All rights reserved.

Classification in data streams is a major area of research, in which Hoeffding trees have long been a favored method. The main contribution of this paper is to show that random feature function can be leveraged by other algorithms to obtain similar or even improved performance over tree-based methods.

With the recent popularity of Deep Learning (DL) methods we also want to test how a random feature in the form of random projection layer performs on Deep Neural Networks (DNNs).

DL aims for a better data representation at multiple layers of abstraction, and for each layer the network needs to be fine-tuned. In classification, a common algorithm to fine-tune the network is the SGD which tries to minimize the error at the output layer using an objective function, such as Mean Squared Error (MSE). A Gradient vector is used to back-propagate the error to previous layers. This gradient nature of the algorithm makes it suitable to be trained incrementally in batches of size one, similar to how incremental training is done. Unfortunately, DNN are very sensitive to hyper-parameters such as learning rate (η), momentum (μ), number of number neurons per level, or the number of levels. It is then not straight forward to provide an off-the-shelf method for data streams.

Propagation between layers is usually done in the form of matrix-vector or matrix-matrix multiplications, which are computational intensive operation. Often hardware accelerators such as FPGAs or GPUs are used to accelerate the calculations. Despite some efforts, acceleration of HT and kNN algorithms for data streams on the GPUs are has some limitations. We talk briefly about this in [Section 2](#).

In recent years, Extreme Learning Machines ([Huang, 2015](#)) (ELMs) have emerged as a popular framework in Machine Learning. ELMs are a type of feed-forward neural networks characterized by a random initialization of their hidden layer, combined with a fast training algorithm. Our random feature method is based on this approach.

We made use of the MOA (Massive Online Analysis) framework ([Bifet et al., 2010a](#)), a software environment for implementing algorithms and running experiments for online learning from data streams in Java. It implements a large number of modern methods for classification in streams, including HT, kNN, and SGD-based methods. We make use of MOA's extensive library of methods to form novel combinations with these methods and further employ an extremely rapid preprocessing technique of projecting the input into a new space via random feature functions (similar to ELMs). We then took the methods purely related to Neural Networks (those which proved most promising under random projections) and implemented them using NVIDIA GPUs and CUDA 7.0; comparing performance to the methods in MOA.

This paper is organized as follows: [Section 2](#) introduces related work on tree based approaches, neural networks, and data streams on GPU. We discuss the use of random features in [Sections 3](#) and [4](#) for HT/kNN methods and neural networks respectively. We first present the evaluation of tree-based methods in [Section 5](#) and later in [Section 6](#) we extend the SGD method in the form of DNNs, using different activation functions. We finally conclude the paper in [Section 7](#).

2. Related work

Hoeffding trees ([Domingos and Hulten, 2000](#)) are state-of-the-art in classification for data streams and they predict by choosing the majority class at each leaf. However, these trees may be conservative at first and in many situations naive Bayes method outperforms the standard Hoeffding tree initially, although it is eventually overtaken ([Holmes et al., 2005](#)). A proposed hybrid adaptive method by [Holmes et al. \(2005\)](#) is a Hoeffding tree with naive Bayes at the leaves, i.e., returning a naive Bayes prediction at the

leaves, if it has been so far more accurate overall than the majority class. Given it's widespread acceptance, this is the default in MOA, and we denote this method in the experimental Section simply as HT. In fact, the naive Bayes classification comes for free, since it can be made with the same statistics that are collected anyway by the tree.

Other established examples include using principal component analysis (reviewed also in [Hastie et al., 2001](#)) for this transformation, and also Restricted Boltzmann Machines (RBMs) ([Hinton and Salakhutdinov, 2006](#)). RBMs can be seen as a probabilistic binary version of PCA, for finding higher-level feature representations. They have received widespread popularity in recent years due to their use in successful deep learning approaches. In this case, $\mathbf{z} = \phi(\mathbf{x}) = f(\mathbf{W}^T \mathbf{x})$ for some non-linearity f : a sigmoid function is typical, but more recently rectified linear units (ReLU, [Nair and Hinton, 2010](#)) have fallen into favor. The weight matrix \mathbf{W} is learned with gradient-based methods ([Hinton, 2000](#)), and the projected output should provide a better feature representation for a neural network or any off-the-shelf method. This approach was applied to data streams already in [Read et al. \(2015\)](#), but concluded that the sensitivity to hyper-parameters and initial conditions prevented good 'out-of-the-box' deployment in data streams.

Approaches such as the so-called extreme learning machines (ELMs) ([Huang et al., 2011](#)) avoid tricky parametrizations by simply using random functions (indeed, ELMs are basically linear learners on top of non-linear data transformations). Despite the hidden layer weights being random, it has been proven that ELMs is still capable of universal approximation of any non-constant piecewise continuous function ([Huang et al., 2006](#)).

Also an incremental version of ELMs is proposed in [bin Huang et al. \(2008\)](#). It starts with a small network, and new neurons are added at each step until a stopping criterion of size or residual error is reached. The difference with our incremental build is that we use one instance at time simulating they arrive in time, and we incrementally train the network. Also our number of neurons is fixed during the training, in other words, we don't add/remove any neuron during the process.

Nowadays, in 2015, it is difficult when talking about DL and DNNs not to mention GPUs. They are a massive parallel architectures providing an outstanding performance for High Performance Computing and a very good performance/watt ratio, as their architecture suits very fine to their needs of DNNs computations. Many tools include a back-end to offload the computation to the GPU. NVIDIA has its own portal for deep learning on GPUs at <https://developer.nvidia.com/deep-learning>.

GPUs has not only used to accelerate DL/DNN computations due to its performance, it has been also been used to successfully accelerate HT and ensembles. However, few works are provided in the context of data streams and GPUs.

The only work we are aware of regarding to HT in the context of online real-time data streams mining is [Marron et al. \(2014\)](#), where the authors present a parallel implementation of HT and Random Forests for binary trees and data streams achieving goods speedups, but with limitations on the size and with high memory consumption. More generic HT implementation of Random Forests is presented in [Grahn et al. \(2011\)](#). In [Schulz et al. \(2015\)](#) the authors introduced an open source library, available at github, to predict images labeling using random forests. The library is also tested their on a cell phone with VGA resolution in real-time with good results.

Also, kNN has already been successfully ported to GPUs ([Garcia et al., 2008](#)). That paper presented one of the first implementations of the "brute force" kNN on GPUs, and compared with several CPU-based implementations with speedups up to two orders of magnitude. kNN is also used in business intelligence ([Huang et al., 2012](#)) and has also its implementation on the GPU.

The same way as with HT, a tool for machine learning (including kNN) is described in Liu et al. (2015).

3. Tree based random feature functions

Transforming the feature space prior to learning and classification is an established idea in the statistical and machine learning literature (Hastie et al., 2001), for example with basis (or feature-) functions. Suppose the input instance is \mathbf{x} of length d . This vector is transformed to a new space $\mathbf{z} = \phi(\mathbf{x})$ via function ϕ , creating new vector \mathbf{z} of length h . Any off-the-shelf model now treats \mathbf{z} as if it were the input. The functions can be either chosen suitably by a domain expert, or simply chosen to achieve a more dimensioned representation of the input. Polynomials and splines are a typical choice.

Regarding HTs with additional algorithms in the leaves (as described in Section 2), this filter can either be placed before the HT, or before the method in the leaves, or both.

In this paper we adapt this methodology to deal with other classifiers in a similar way, namely kNN and an SGD-based method (rather than naive Bayes) at the leaves. We denote these cases HT-kNN and HT-SGD, respectively. For example, in HT-SGD, a gradient descent learner is employed in the leaves of each tree. Similarly to HT, predictions by the kNN and an SGD-based method are only used if they are more accurate on average than the majority class.

3.1. Ensembles in data streams

Bagging is an ensemble method used to improve the accuracy of classifier methods. Non-streaming bagging (Breiman, 1996) builds a set of M base models, training each model with a bootstrap sample of size N created by drawing random samples with replacement from the original training set. Each base model's training set contains each of the original training example K times where $P(K = k)$ follows a binomial distribution. This binomial distribution for large values of N tends to a Poisson($\lambda = 1$) distribution, where $\text{Poisson}(\lambda = 1) = \exp(-1)/k!$. Using this fact, Oza and Russell (2001b); 2001a) proposed Online Bagging, an online method that instead of sampling with replacement, gives each example a weight according to Poisson(1). ADWINBagging (Bifet et al., 2009) is an adaptive version of Online Bagging that uses a change detector to decide when to discard under-performing ensemble models.

Leveraging Bagging (LB, Bifet et al., 2010b) improves ADWINBagging, increasing the weights of this resampling using a larger value λ to compute the value of the Poisson distribution. The Poisson distribution is used to model the number of events occurring within a given time interval. It proved very competitive.

Again, we can run a filter on the input instances before entering the ensemble of trees, or at the leaves. It is even possible to run the filter again on the output of an ensemble (i.e., the votes), before running an additional stacking procedure. This kind of methodology can give way to rather 'deep' classifiers. Fig. 1 illustrates a possible setup. In this sense of multiple levels we could also call our approach *deep learning*. It is debatable whether decision trees can be called a deep method (their levels involve partitioning an existing feature set rather than because they simply partition a space rather than create higher-level feature space). However, several of the methods we investigate have at least multiple levels of feature transformation, which is behind the power of most deep methods. In the following Section we investigate the empirical performance of several novel combinations based on the methodology described so far.

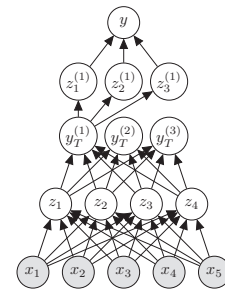


Fig. 1. An example setup: Input \mathbf{x} is filtered (i.e., projected) to random layer \mathbf{z} (first layer of connections), which goes to an ensemble of, for example, HTs (second layer), wherein instances are partitioned to the leaves and are again filtered (third layer) and used as training for, say, SGD, producing (in the fifth layer of connections) final vote y . Note, however, that we only draw the final two layers wrt to the first of the HT models.

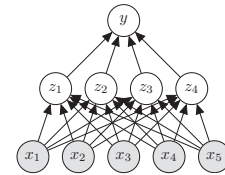


Fig. 2. Random Projection Layer: Input \mathbf{x} is projected to a random layer \mathbf{z} (first layer of connections), which is trained to produce the final vote y .

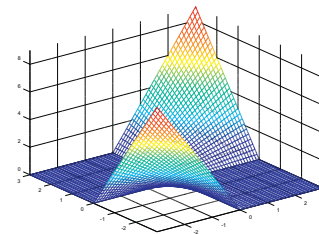


Fig. 3. Terrain of ReLU basis function on two input attributes x_1, x_2 the feature function z is given on the vertical axis.

4. Neural networks with random projections for data streams

Data streams are potentially infinite, and so, they can evolve with time. This means the statistical distribution of the data we are interested on can change. The idea behind the random layer is to improve data localization across the space the trained layer sees. Imagine the instance is a tiny luminous point on the space, with enough random neurons acting as a mirror we hope the trained layer can capture better the data movement. The strategy used by the random projection layer is shown in Fig. 2

Except for the fact it is never trained, the random layer is a normal layer and need its activation functions, in this work sigmoid, ReLU, ReLU incremental and a Radial Basis Function are used.

The sigmoid function used is the standard one, with $\sigma(x) \in [-1, 1]$:

$$\sigma(a_k) = \frac{1}{1 + e^{-a_k}}$$

where $a_k = \mathbf{W}_k^T \mathbf{x}$ is the k -th activation function and \mathbf{W} is the weight $d \times h$ matrix (d input attributes, h output features).

ReLU functions are defined as:

$$z_k = f(a_k) = \max(0, a_k)$$

As stated in Section 2, ReLUs activation are very efficient as they require only a comparison. In our random projection we expect near 50% of the neurons to be active for a single instance (the terrain of a ReLU is exemplified in Fig. 3).

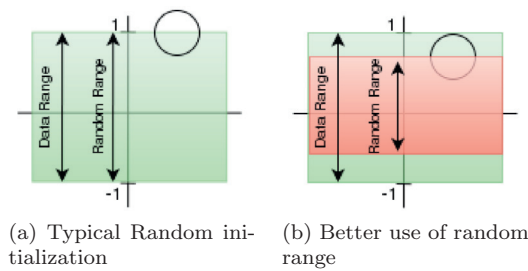


Fig. 4. Random number initialization strategies. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

One variation we can do to the standard ReLU is to use the mean value of the attribute as a threshold. The mean value is calculated incrementally as instances arrive. We call this variant ReLU incremental, and is defined as:

$$f(a_k) = \max(\bar{a}_k, a_k)$$

The last activation function we are using is the Radial Basis Function (RBF):

$$\phi(x) = e^{-\frac{(x-c_i)^2}{2\sigma^2}}$$

where x is an input instance attribute value and c_i is a random value set at initialization time. Each neuron in the random layer has its own set of c_i , the length of both vector x and c are the same. So, we can see the operation $(x - c_i)^2$ as the euclidean distance of the current instance to a randomly positioned center. The σ^2 is a free parameter. A simplification we can do to this notation is:

$$\gamma = \frac{1}{2\sigma^2}$$

In our experiments we try different γ values passed at command line. We use the following notation in our experiments:

$$\phi(x) = e^{-\gamma(x-c_i)^2}$$

All matrices and vectors in our model are initialized using random numbers. Matrices are used as normal weight matrix, but the function of the vectors are activation function dependent. Usually initialization is done using random numbers with $\mu = 0$ and $\sigma = 1$. Assuming our data range $\in [-1, 1]$ if we put a Gaussian centered at one of the endpoints, half of its area of influence area is wasted and will never see a point making it harder to fill the whole space and so the discovering of points.

If a smaller range is used, $\sigma \in (0, 1)$ (note the open interval), we can improve each neuron's area of influence, as shown in Fig. 4b. In red the random numbers range is smaller than data range so if we put a Gaussian at the random endpoint can improve its influence area. In this example we used a Gaussian function as an example, but we the idea extends the same for activation functions. In fact this is what we do in Section 6, specially when talking about the sigmoid neurons as they are always used at the trained layer.

5. Random feature function evaluation

Among the methods we investigate (e.g., HT, kNN, SGD¹), different levels of filters and ensembles and possibly additional classification in the leaves (in the case of HT), there are a multitude of possible combinations. We first investigate the viability of random

Table 1

Data sources used in the experimental evaluation. Synthetic datasets are listed first.

Dataset	#Attributes	#Instances
RBF1	10	100,000
HYP1	10	100,000
LED1	24	100,000
LOC1	25	100,000
LOC2	2500	100,000
Poker	10	829,201
Electricity	8	45,312
CoverType	54	581,012
SUSY	8	5,000,000

feature functions and their effect on the different classifiers (comparing these common methods with their 'filtered' versions that we denote HT-SGD, kNN-F, and SGD-F. This study led us to novel combinations, which we further compare to the benchmark methods and state-of-the-art Leveraging Bagging (LB-HT).

The random feature used in these evaluations are basically ELMs (Huang et al., 2011). In this Section, we use only ReLU (explained in Section 4) as the activation function. In Section 4 we extend the functions used within the random feature to define a random projection layer for DNNs.

Our random feature is based on ELMs, which are defined using Radial Basis Functions, but instead in this section we use ReLU as the activation functions. Both functions are defined in detail in Section 4

All experiments in this section were carried out using the MOA framework (Bifet et al., 2010a) with prequential evaluation: each individual example is used to test the model before it is used for training, and from this the accuracy can be incrementally updated. We used an 8-core (3.20GHz each) desktop machine allowing up to 1 gigabyte of RAM per run (all methods were able to finish).

Table 1 lists the data sources used. A thorough description of most of the datasets is given in Read et al. (2012). Of the others, LOC1 and LOC2 are datasets dealing with classifying the location of an object in a grid of 5×5 and 50×50 pixels respectively, described in Read and Hollmén (2014). Baldi et al. (2014) has features that are kinematic properties measured by particle detectors in an accelerator. The binary class distinguishes between a signal process which produces supersymmetric particles and a background process which does not. It is one of the largest datasets in the UCI repository that we could find.

For the feature filter we used parameters $h = 5d$ hidden units for kNN-F and $h = 10d$ for SGD-F and HT-F (a decision based on the relative computational sensitivity of kNN to a larger attribute space – for LOC2 this means 25,000 attributes in the projected space for SGD-F, and half of that for kNN-F) – except where this is varied in Fig. 5. For kNN we used a buffer size of 5000. For LB we specify 10 models. In other cases, the default parameters in MOA are used.

Fig. 5 displays the results of varying the relative size of the new feature space (wrt to the original feature space) on two real-world datasets. Note that the feature space is different, so even when this ratio is 1: 1, performance may differ.

With regard to kNN, performance improves with more feature functions. In one of the two cases, this is sufficient to overtake kNN on the original feature space. Unfortunately, kNN is particularly sensitive to the number of attributes, so complexity becomes an issue long before other methods. The new feature space does not help the performance of HT, and in neither case does it reach the performance of HT on the original feature space. In fact, it begins to decrease again. This is because too many features makes it difficult for HT to become confident enough to split on, and may

¹ We refer, in this case, to the instantiation with default parameters in MOA, i.e., minimizing hinge loss.

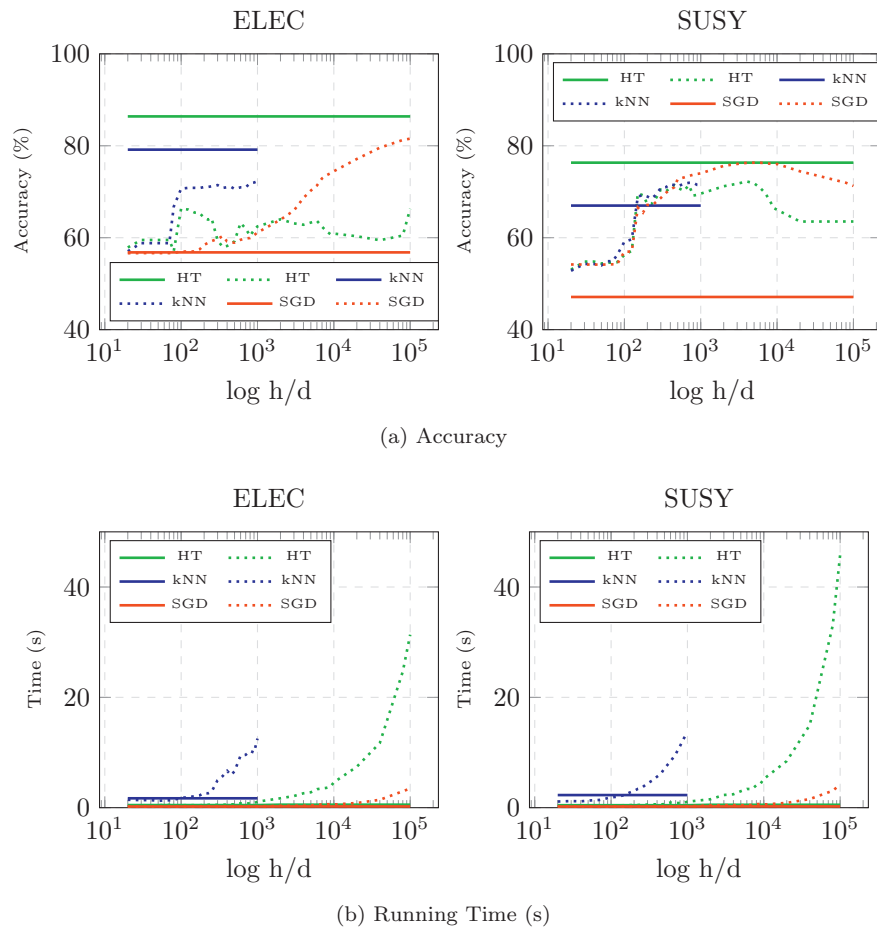


Fig. 5. Accuracy (Fig. 5a) and Running Time (Fig. 5b) on a 10,000-instance segment of two real-world datasets (ELEC and SUSY) for varying proportions of h (number of hidden units / basis functions) wrt d . kNN has been cut out after $h = 1000/d$ due to scalability reasons. Note the log scale on the horizontal axis.

split poorly. Also, by partitioning the feature space, interaction between the features is lost. SGD reacts best to a new feature space. As noticed earlier (Read et al., 2012), SGD is a poor performer compared to HTs, however, working in a feature space of random RE-LUs, SGD-F actually reaches HT performance (on SUSY, and looks promising under ELEC) with similar time complexity. Even at 1000 times the original feature space, running time is acceptable (only a several seconds per 10,000 instances). On the other hand, the increased memory use is significant across all methods. SGD requires 1000 times more memory in this setting.

From this initial investigation we formulate several method combinations for a more extensive evaluation. Table 2 displays the final accuracy over the data stream. The first four columns represent the baselines and state-of-the-art (LB-HT), and remaining columns are a selection of new method combinations. Fig. 6 gives a more detailed over-time view of the largest dataset (SUSY), with the average performance plotted over the entire stream over 100 intervals, and also the first 1/10th of the data (again, over 100 intervals). The second plot gives more of an idea about how models respond to fresh concepts. Learning new concepts is a fundamental part in data streams of adapting to concept drift.

Regarding this experiment some of the most important observations and conclusions are as follows:

- SGD-F (i.e., SGD with random feature functions), even in this first analysis, out-competes established methods like kNN on several datasets.
- kNN benefits relatively less (than SGD) from the feature functions filter. This is expected, since kNN is already a non-linear learner.

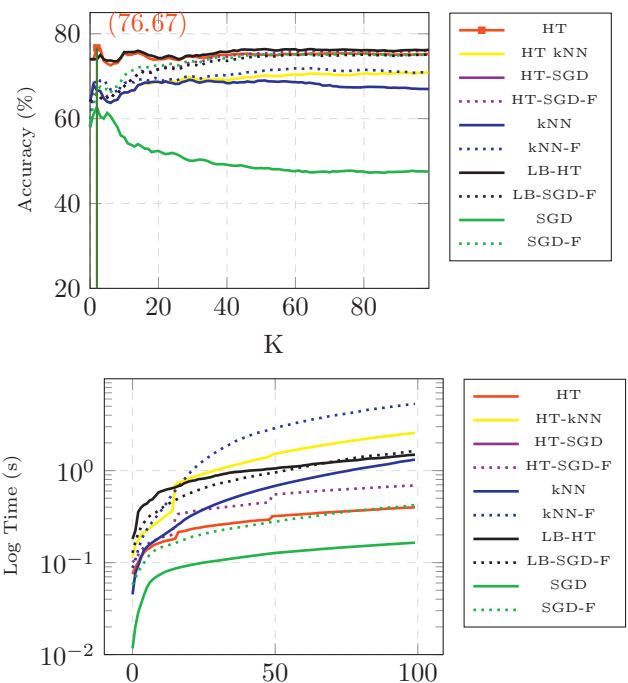


Fig. 6. Performance over first 50,000 examples (right) of the SUSY data, in each divided into 100 windows.

Table 2

Final accuracy and running times. The dataset-wise ranking is given in (parentheses) and the average of these ranks is given in the final row.

(a) Accuracy									
Dataset	HT	SGD	kNN	LB-HT	LB-SGD-F	kNN-F	SGD-F	HT-kNN	HT-SGD-F
RBF1	75.0 (5)	54.5 (9)	92.0 (2)	88.7 (4)	72.0 (8)	90.4 (3)	72.0 (7)	92.6 (1)	73.7 (6)
Rbfd	65.7 (5)	51.3 (9)	88.6 (1)	79.5 (4)	59.8 (8)	86.3 (2)	59.9 (7)	84.9 (3)	59.9 (6)
HYP1	87.7 (1)	50.3 (9)	82.9 (4)	85.7 (2)	67.2 (7)	77.0 (5)	67.2 (7)	83.3 (3)	67.9 (6)
LED1	73.1 (1)	10.3 (9)	62.8 (3)	72.0 (2)	15.6 (6)	49.0 (5)	15.5 (7)	62.8 (3)	15.5 (7)
POKR	76.1 (6)	68.9 (9)	69.3 (8)	87.6 (1)	82.3 (2)	81.5 (4)	81.9 (3)	74.8 (7)	80.1 (5)
LOC1	85.5 (8)	80.4 (9)	91.0 (2)	90.5 (6)	90.7 (4)	88.8 (7)	90.7 (3)	91.3 (1)	90.7 (5)
LOC2	56.3 (5)	51.5 (9)	75.7 (2)	52.6 (8)	56.8 (4)	74.5 (3)	55.9 (7)	75.9 (1)	56.1 (6)
ELEC	79.2 (4)	57.6 (9)	78.4 (5)	89.8 (1)	74.8 (7)	74.2 (8)	74.8 (6)	82.5 (2)	81.8 (3)
COVT	80.3 (5)	60.7 (9)	92.2 (1)	91.7 (2)	78.7 (6)	91.6 (3)	78.7 (7)	91.2 (4)	78.3 (8)
SUSY	78.2 (3)	76.5 (7)	67.5 (9)	78.7 (1)	77.7 (4)	71.2 (8)	77.7 (5)	77.2 (6)	78.4 (2)
avg rank	4.30	8.80	3.70	3.10	5.60	4.80	5.90	3.10	5.40

(b) Running time (s)									
Dataset	HT	SGD	kNN	LB-HT	LB-SGD-F	kNN-F	SGD-F	HT-kNN	HT-SGD-F
RBF1	0 (3)	0 (1)	3 (6)	3 (5)	4 (8)	14 (9)	0 (2)	4 (7)	1 (4)
Rbfd	1 (3)	0 (1)	3 (6)	2 (5)	4 (8)	15 (9)	0 (2)	4 (7)	1 (4)
HYP1	0 (2)	0 (1)	3 (6)	2 (5)	4 (7)	13 (9)	0 (3)	4 (8)	1 (4)
LED1	0 (2)	0 (1)	7 (6)	2 (5)	17 (8)	40 (9)	1 (3)	8 (7)	1 (4)
POKR	9 (2)	3 (1)	455 (8)	91 (5)	279 (6)	1539 (9)	21 (3)	422 (7)	26 (4)
LOC1	1 (2)	0 (1)	8 (7)	2 (5)	21 (8)	48 (9)	1 (3)	8 (6)	2 (4)
LOC2	9 (2)	4 (1)	1276 (7)	93 (3)	1917 (8)	2270 (9)	367 (5)	1230 (6)	350 (4)
ELEC	1 (3)	0 (1)	14 (7)	10 (6)	9 (5)	49 (9)	1 (2)	19 (8)	2 (4)
COVT	19 (2)	11 (1)	605 (6)	220 (3)	4119 (9)	3998 (8)	233 (4)	727 (7)	250 (5)
SUSY	45 (2)	25 (1)	1464 (8)	530 (5)	1040 (6)	4714 (9)	118 (3)	1428 (7)	159 (4)
avg rank	2.30	1.00	6.70	4.70	7.30	8.90	3.00	7.00	4.10

However, on a few datasets accuracy is 5-10 percentage points higher with the filter.

- kNN can be used effectively in the leaves of HT instead of the default of naive Bayes. There is an additional computational cost involved, but results showed this to be highly competitive method – best equal overall in predictive performance tied with state-of-the-art LB-HT
- HT is difficult to improve on using feature functions (at least with the ReLUs that we experimented with). Again, this can be attributed to HT being a non-linear learner. Peak accuracy is reached in relatively short space of time.
- SGD takes longer than HT or LB-HT to reach competitive accuracy, but the gap narrows significantly with more examples (for example, under SUSY). On the largest datasets, the final average accuracy is within a percentage point – and this average includes initial poorer performance. Therefore, on particularly big data streams (which are increasingly common), HTs could find themselves increasingly challenged to stay ahead of these methods.
- HT-SGD-F is comparable to the state of the art LB-HT on several datasets, but demonstrates more favorable running times.
- Unlike many deep learning techniques, these random functions do not require sensitive calibration.
- Unsurprisingly, kNN-based methods perform best on the dataset Rbfd which has a drifting concept, since they automatically phase out older concepts. We did not look into detail about dealing with concept drift in this paper, but this can be dealt with by ‘meta methods’, e.g., (Bifet and Gavalda, 2007).
- Employing random feature functions as a ‘filter’ in the MOA framework is a convenient and flexible way to apply it in a range of different data-stream methods.

6. GPU extended evaluation

In the previous Section evaluations, we noticed SGD methods have the strongest advantage from random feature functions. This added to the increasing popularity of DL methods, we elected this

Table 3

Random numbers initialization strategy for the different activation functions.

Activation	Weight matrix	Bias vector
RBF	mean = 0.0 and std = 1.0	gamma
Sigmoid	mean = 0.0 and std = 0.9	mean = 0.0 and std = 0.2
ReLU	mean = 0.0 and std = 1.0	mean = 0.0 and std = 0.1
ReLU Inc	mean = 0.0 and std = 1.0	0.0

strategy for further investigation and experimentation in this Section. A natural choice for implementing DNNs is to use GPUs to accelerate the calculations. Our experiments were evaluated on an NVIDIA Tesla K40c with 12GB of RAM each, 15 SMX and up to 2880 simultaneous threads and CUDA 7.0.

Another motivation to use to GPUs the selection of the network hyper parameters by using cross-validation: for each dataset and activation functions different configurations are tested and best performing one is chosen. In turned out this was a high number of combinations and a way to accelerate the process is using GPUs.

The random projection layer is implemented using an standard two layers feed-forward fully connected network. The input is fed to the random layer, which is never trained, and the output from this layer is forwarded to the trained layer. In this work we use the SGD and MSE as the training algorithm and objective function respectively for the last layer.

We use three of the data sources from Table 1 Coverttype (COVT), Electricity (ELEC), and SUSY. This way we can compare the accuracy obtained in this Section against well-known state-of-the-art algorithms.

The initialization of each layer depends on the activation function used, we tried different random number initialization strategies and those for which we achieved the best results are summarized in Table 3. Most of the weight matrices are initialized using random numbers with mean = 0 and $\sigma = 1.0$, except for the sigmoid activation function. The bias vector purpose and usage is activation function dependent.

Table 4
GPU tests; best results.

Dataset	Best algorithms			Random projection				
	Alg	Acc (%)	Time (s)	Activ	Size	Acc (%)	Time (s)	speedup
ELEC	LB-HT	89.8	10	Sigmoid	100	85.33	1.2	8x
COVT	kNN	92.2	605	ReLU	2000	94.59	32	17x
SUSY	LB-HT	78.7	530	RBF	600	77.63	172	3x

Different activation functions have been tested at the random layer: RBF-gamma, Sigmoid, ReLU, incremental ReLU. Sigmoid and ReLU are used in the standard way. As we can see in Table 3 bias vector for RBF stores the gammas, in our evaluations we use $\gamma = \{0.001, 0.01, 0.1, 1.0, 10.0\}$. ReLU incremental used the bias vector to store the incremental mean for each output attribute. At the trained layer always, we always use the standard sigmoid as the activation function.

The same way as in Section 5, the network is built incrementally using prequential learning; we visit each instance only one time. This is in contrast to typical DNNs training, where instances are loaded in batches and the algorithm iterates over them given number of times and, every time the error is reduced the model is checkpointed and spread to be used.

Table 4 summarizes the best results we obtained, and it compares them with the best results obtained in Section 5 evaluations. We choose the algorithms by accuracy, and compared the time to run against them. Configuration were chosen by cross-validation using the following parameters: $\mu \in [0.1, 1.0]$ with an increment of 0.1 and a similar range for learning rate. Sizes tested: [10, 100] increment of 10, [100, 1000] increment of 100, and two more sizes: 1500, 2000.

For the electricity dataset random projection layer (RPL) obtained an accuracy of 85.33% using a random layer of 100 neurons and a sigmoid activation function. As we can see in Table 4 the best performing algorithm is the LB-HT which achieved a 89.8%. If compared with results at Table 2, we can see our method is the second best result, 4.47 percentage points less.

In the covertype dataset evaluation RPL obtained the best result for this dataset with an accuracy of 94.59%, improving 2.39 percentage points the kNN algorithm using a ReLU activation functions.

Finally, our RPL performed relatively poorly in the SUSY dataset using 600 random neurons. We obtained a 77.63%, 1.07 percentage points less than the LB-HT. This distance is less than the distance obtained with the electricity algorithm, but if we rank out results with those in Table 2 we are at the sixth position.

With regard the time to complete, we can see the GPU is faster in all of the three datasets. For the electricity dataset RPL is 8 times faster, for the CoverType dataset 17 times faster and 3 times faster for the SUSY dataset.

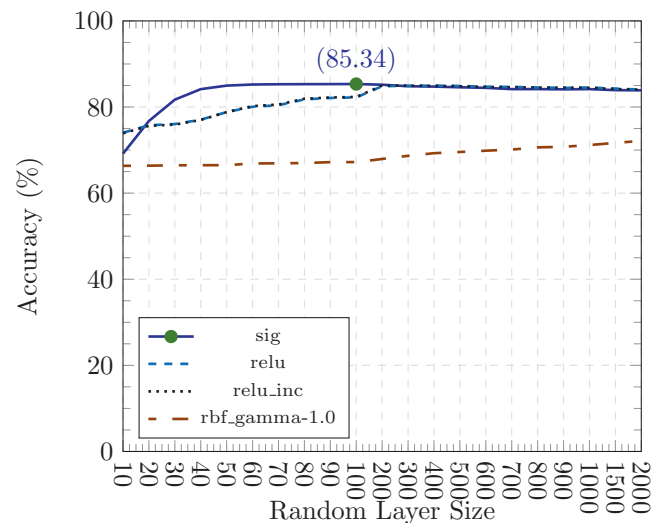
Now we detail for each dataset the activation curves, the momentum and learning rate for this figure are the same across all sizes and we used the ones for the best results to see how size affects the accuracy.

6.1. Electricity dataset

Table 5 summarizes the best results for each activation functions, and its configurations. As saw previously sigmoid activation function performed better than the others for this dataset. In the second position we find ReLU and ReLU inc activation functions which gave similar results, and slightly worse than the sigmoid. Regarding the RBF, all configurations we tried performed worse if compared to sigmoid, ReLU and ReLU inc, but very similar for the

Table 5
ELEC evaluation.

Activation	Random neurons	μ	η	Accuracy(%)
SIG	100	0.3	0.11	85.33
ReLU	400	0.3	0.01	84.95
ReLU inc	200	0.3	0.01	84.97
RBF $\gamma=0.001$	2000	0.7	1.01	72.13
RBF $\gamma=0.01$	2000	0.7	1.01	72.13
RBF $\gamma=0.1$	2000	0.7	1.01	72.13
RBF $\gamma=1.0$	2000	0.7	1.01	72.13
RBF $\gamma=10.0$	2000	0.7	1.01	72.13

**Fig. 7.** Elec dataset.**Table 6**
COV evaluation.

Activation	Random neurons	μ	η	Accuracy(%)
SIG	1000	0.4	0.11	94.45
ReLU	2000	0.4	0.01	94.59
ReLU inc	2000	0.4	0.01	94.58
RBF $\gamma=0.001$	90	0.9	1.01	73.18
RBF $\gamma=0.01$	90	0.9	1.01	73.18
RBF $\gamma=0.1$	90	0.5	1.01	73.18
RBF $\gamma=1.0$	90	0.8	1.01	73.18
RBF $\gamma=10.0$	90	1.0	1.01	73.18

different gammas. In Fig. 7 we can see how accuracy changes with different sizes we tested.

6.2. CoverType dataset

Table 6 give us the best results for the COVT dataset each activation function. We can see a similar pattern as with the ELEC evaluation, SIG, ReLU and ReLU inc performed much better than the RBFs, and all three can beat results shown in Table 2. This time the best result is obtained with the ReLU activation function at the random layer.

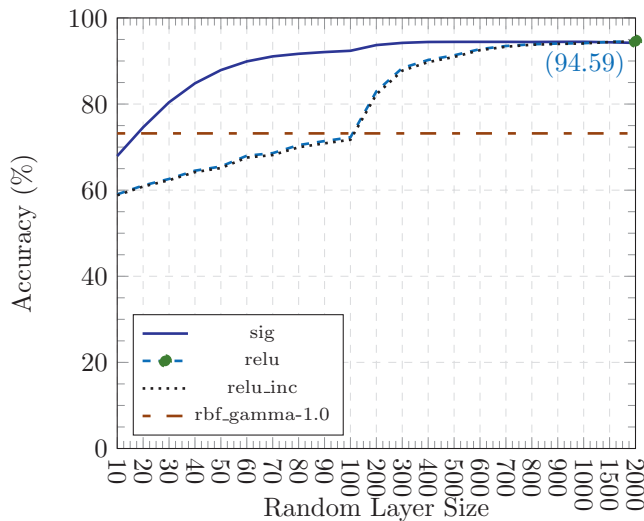


Fig. 8. COV normalized dataset.

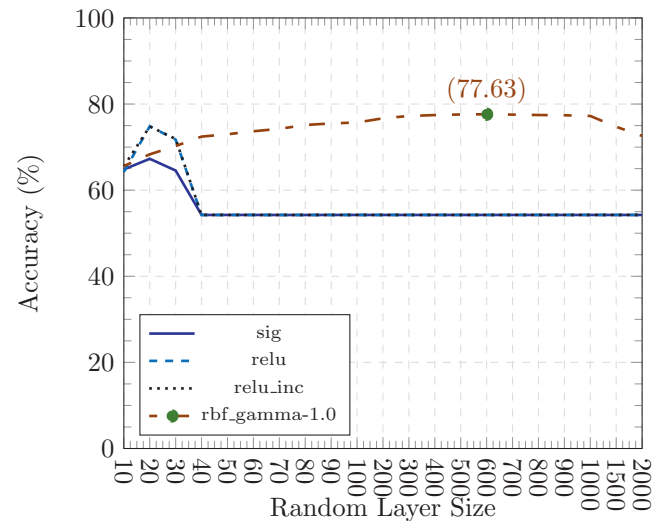


Fig. 9. SUSY dataset.

Table 7
SUSY evaluation.

Activation	Random neurons	μ	η	Accuracy(%)
SIG	20	1	0.61	67.28
ReLU	20	1	0.61	74.84
ReLU inc	20	1	0.91	74.80
RBF $\gamma=0.001$	600	1	0.71	77.63
RBF $\gamma=0.01$	600	1	0.71	77.63
RBF $\gamma=0.1$	600	1	0.71	77.63
RBF $\gamma=1.0$	600	1	0.71	77.63
RBF $\gamma=10.0$	600	1	0.71	77.63

In Fig. 8 we can see the activation curves. Although we got the best result with ReLU, the sigmoid has a better learning curve and it is very close to the ReLU accuracy. ReLU inc has a very similar learning curve as the standard ReLU. The different RBFs for the same momentum and learning with different sizes gives very similar (if not equal) results, so we chose the lower sizes.

6.3. SUSY dataset

Table 7 shows the best results for the SUSY dataset and activation function, and Fig. 9 the learning curves. The most noticeable effect is sigmoid, ReLU and ReLU inc the stop learning very soon, with only 20 random neurons ReLU reached its maximum peak with 74.85%. The RBFs which performed poorly in previous evaluations, here are those with the best results.

One curious result we can see is that the RBFs are performing around 7x% in all 3 evaluations. Even if 2 of the 3 results are not very good, it seems they are not very sensitive to the different datasets, and somehow the results are stable across different data distributions.

7. Conclusions

In this paper, we studied combinations of Hoeffding trees, nearest neighbour, and gradient descent methods adding a layer based on a random feature function filter. We found that this random layer can turn a simple gradient descent learner into a competitive method for real-time data analysis. With this first attempt we could even improve on current state-of-the-art algorithms, scoring the best and the second best results for two out of three datasets tested. Like Hoeffding Trees and nearest neighbour methods, but

unlike many other gradient descent-based methods, the random layer works well without intensive parameter tuning.

We successfully extended and implemented on GPUs, obtaining powerful predictive performance. This suggests that using GPUs for data stream mining is a promising research topic for obtaining new fast and adaptive machine learning methodologies.

In the future we intend to look for adding and pruning units incrementally in the stream over time to respond to make more efficient use of memory and adapt to drifting concepts. Also we would like to continue studying how to obtain new high scalable methods using GPUs.

Acknowledgment

This work was supported in part by the Aalto University AEF research programme <http://energyefficiency.aalto.fi/en/>, by NVIDIA through the UPC/BSC GPU Center of Excellence, and the Spanish Ministry of Science and Technology through the TIN2012-34557.

References

- Baldi, P., Sadowski, P., Whiteson, D., 2014. Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun.* 5 (4308).
- Bifet, A., Gavalda, R., 2007. Learning from time-changing data with adaptive windowing. In: *SIAM International Conference on Data Mining*.
- Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B., 2010. MOA: Massive Online Analysis. *J. Mach. Learn. Res. (JMLR)*. <http://moa.cs.waikato.ac.nz/>.
- Bifet, A., Holmes, G., Pfahringer, B., 2010. Leveraging bagging for evolving data streams. In: *ECML PKDD'10*. Springer-Verlag, Berlin, Heidelberg, pp. 135–150.
- Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., Gavalda, R., 2009. New ensemble methods for evolving data streams. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*, pp. 139–148.
- Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24 (2), 123–140. <http://dx.doi.org/10.1023/A:1018054314350>.
- Domingos, P., Hulten, G., 2000. Mining high-speed data streams. In: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 71–80.
- Garcia, V., Debreuve, E., Barlaud, M., 2008. Fast k nearest neighbor search using gpu. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6.
- Grahn, H., Lavesson, N., Lapajne, M.H., Slat, D., 2011. Cudarf: a cuda-based implementation of random forests. In: Siegel, H.J., El-Kadi, A. (Eds.), *AICCSA. IEEE Computer Society*, pp. 95–101.
- Hastie, T., Tibshirani, R., Friedman, J., 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.
- Hinton, G., 2000. Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14 (8), 1711–1800.
- Hinton, G., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. *Science* 313 (5786), 504–507.
- Holmes, G., Kirkby, R., Pfahringer, B., 2005. Stress-testing Hoeffding trees. In: *9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD '05)*, pp. 495–502.

- Huang, G., 2015. What are extreme learning machines? filling the gap between frank rosenblatt's dream and john von neumann's puzzle. *Cogn. Comput.* 7 (3), 263–278. doi:[10.1007/s12559-015-9333-0](https://doi.org/10.1007/s12559-015-9333-0).
- bin Huang, G., bin Li, M., Chen, L., kheong Siew, C., 2008. Incremental extreme learning machine with fully complex hidden nodes. *Neurocomputing* 71 (4-6), 576–583.
- Huang, G.-B., Chen, L., Siew, C.-K., 2006. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Netw.* 17 (4), 879–892. doi:[10.1109/TNN.2006.875977](https://doi.org/10.1109/TNN.2006.875977).
- Huang, G.-B., Wang, D., Lan, Y., 2011. Extreme learning machines: a survey. *Int. J. Mach. Learn. Cybern.* 2 (2), 107–122. doi:[10.1007/s13042-011-0019-y](https://doi.org/10.1007/s13042-011-0019-y).
- Huang, L., Liu, Z., Yan, Z., Liu, P., Cai, Q., 2012. An implementation of high performance parallel knn algorithm based on gpu. In: 2012 Third International Conference on Networking and Distributed Computing (ICNDC), pp 30–30. doi:[10.1109/ICNDC.2012.15](https://doi.org/10.1109/ICNDC.2012.15).
- Liu, D., Chen, T., Liu, S., Zhou, J., Zhou, S., Teman, O., Feng, X., Zhou, X., Chen, Y., 2015. Pudiannao: a polyvalent machine learning accelerator. In: Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems, pp. 369–381.
- Marron, D., Bifet, A., Morales, G.D.F., 2014. Random forests of very fast decision trees on gpu for mining evolving big data streams. In: 21st European Conference on Artificial Intelligence 2014.
- Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21–24, 2010, Haifa, Israel, pp. 807–814.
- Oza, N., Russell, S., 2001. Online bagging and boosting. In: *Artificial Intelligence and Statistics 2001*. Morgan Kaufmann, pp. 105–112.
- Oza, N.C., Russell, S.J., 2001. Experimental comparisons of online and batch versions of bagging and boosting. In: *KDD*, pp. 359–364.
- Qu, W., Zhang, Y., Zhu, J., Qiu, Q., 2009. Mining multi-label concept-drifting data streams using dynamic classifier ensemble. In: *Asian Conference on Machine Learning*. Springer, pp. 308–321.
- Read, J., Bifet, A., Pfahringer, B., Holmes, G., 2012. Batch-incremental versus instance-incremental learning in dynamic and evolving data. *11th Int. Symposium on Intelligent Data Analysis*.
- Read, J., Hollmén, J., 2014. A deep interpretation of classifier chains. In: *Advances in Intelligent Data Analysis XIII - 13th International Symposium, IDA 2014*, pp. 251–262.
- Read, J., Perez-Cruz, F., Bifet, A., 2015. Deep learning in multi-label data-streams. *Symposium on Applied Computing*. ACM.
- Schulz, H., Waldvogel, B., Sheikh, R., Behnke, S., 2015. CURFIL: random forests for image labeling on GPU. In: *VISAPP 2015 - Proceedings of the 10th International Conference on Computer Vision Theory and Applications*, Volume 2, Berlin, Germany, 11–14 March, 2015., pp. 156–164.
- Shaker, A., Hüllermeier, E., 2012. Instance-based classification and regression on data streams. In: *Learning in Non-Stationary Environments*. Springer New York, pp. 185–201.



Diego Marrón is a Ph.D. student at the Computer Architecture department at Universitat Politècnica de Catalunya (UPC) and works as researcher at the Barcelona Supercomputing Center NVIDIA GPU Center of Excellence. His research interests include GPU computing, scalable programming, data streams and neural networks.



Jesse Read obtained his Ph.D. in Computer Science from the University of Waikato, New Zealand, in 2010. He later carried out postdoc research in the Department of Signal Theory and Communications at the Carlos III University of Madrid and the Department of Information and Computer Science at Aalto University in Finland. He is currently working at Tèlècom ParisTech in France. His main research interests include multi-label classification and data streams.



Albert Bifet is Associate Professor at Telecom ParisTech. He is the author of a book on Adaptive Stream Mining and Pattern Learning and Mining from Evolving Data Streams. He is one of the leaders of MOA and Apache SAMOA software environments for implementing algorithms and running experiments for online learning from evolving data streams. He is serving as Co-Chair of the Industrial track of IEEE MDM 2016, ECML PKDD 2015, and as Co-Chair of BigMine (2016, 2015, 2014, 2013, 2012), and ACM SAC Data Streams Track (2016, 2015, 2014, 2013, 2012).