CDI: A Complex Event Processing Distributed System for Time Series Forecasting Applied to Air Quality

FACULDADE DE TECNOLOGIA

Luís Fernando L. Grim & André Leon S. Gradvohl

lgrim@ifsp.edu.br, gradvohl@ft.unicamp.br

School of Technology – UNICAMP

UNICAMP

Introduction

The Online Sequential Extreme Learning Machine algorithm is suitable for forecasting Data Streams with Concept Drifts. Nevertheless, data streams forecasting require high-performance implementations due to the high incoming samples rate.

In this preliminary work, we analyzed the scalability of parallel implementations for the Online Sequential Extreme Learning Machine in C programming language, with OpenBLAS, Intel MKL, and MAGMA libraries, that we proposed.

Goals

- Compare the performance – prediction error and execution time – of the proposed implementations when forecasting concentrations of breathable Particulate Matter (PM_{10}) in the air, on the resources that HPI Future Soc Lab made available to us;
- Based on the results obtained in this work, the future intention is to construct an ensemble that distributes several models with one of the implementations developed as kernels, functioning as operators of a Complex Event Processing System.

Proposed Implementations

We have developed three distinct C implementations: OS-ELM with OpenBLAS (OS-ELMob), OS-ELM with MKL (OS-ELMmkl) and OS-ELM whit MAGMA (OS-ELMmgm).

OpenBLAS and Intel MKL provide functions that explore the multithread features in multicore processors, which also enables the parallelism in multiprocessors architectures. MAGMA offers functions that run parallel in hybrid architectures for Multicore + GPU.

The dataset and pre-processing steps

- Hourly samples of PM_{10} concentration, provided by QUALAR System (CETESB), from 01 January 1998 to 23 November 2017 from Cubatão – Vila Parisi station;
- 174,397 samples, 0 as the minimum value, 1,470 μm/m^3 as the maximum value, and 8,011 missing samples filled with the most probable value and linear interpolation;
- Outlier replacement considering the maximum value of PM_{10} reported for the monitoring station and Hampel filter; Data normalization between $[0, 1]$;
- Data organized in hourly samples: the current sample and the last five instants of time, $x_n = [s_{n-5} \ldots, s_{n-1}, s_n]$, are used to predict the next instant $y = [s_{n+1}]$.

Configurations of the OS-ELM implementations

- Cases with 100 hidden neurons and with Block Sizes (BS) of 1, 10, 100, 1,000, 2,500 and 5,000 instances. For the initial training dataset, we used the 5,000 older instances, and simulated a data stream with the remaining 169,397 instances.
- We evaluated the performance measuring the Root Mean Square Error (RMSE) and the Total Real Times (TRT) in seconds. We repeated the experiments for 50 trials.

Future SOC Lab resources

GPU Cluster: 32 cores Intel (R) Xeon (R) CPU E5-2620 v4 @ 2.10 GHz; 128 GB of RAM; Ubuntu 16.04.2 O.S.; OpenBLAS 0.2.20 and Intel MKL 2018.1.163 with default configuration (using 16 threads):

- 6 Tesla K80 @ 824 MHz and 12.4 GB;
- MAGMA 2.2.0 compiled for CUDA 8.0 and OpenBLAS 0.2.20;
- OS-ELMmgm uses use only 1 of the 6 available GPUs.

Virtual Machines: vCPUs Intel Core i7 9xx (Nehalem Class Core i7) @ 2.39 GHz:

- Small VM: 1 vCPU, 1 GB RAM;
- Medium VM: 2 vCPUs, 2 GB RAM;
- Large VM: 4 vCPUs, 4 GB RAM;
- XL VM: 8 vCPUs, 8 GB RAM;

Ubuntu 16.04.2 O.S.; OpenBLAS 0.2.20 and Intel MKL 2018.1.163 with default configuration (the number of threads follows the number of cores of the VM).

Experimental Results

- The RMSEs are practically the same in all implementations, regardless of the execution environment, as well as the BS and also related to the VM size, with the means ranging from 0.1080 to 0.1081 and standard deviations varying from 0.00003 to 0.00005.
- The following tables show the TRTs obtained. The figures indicate the speedups.

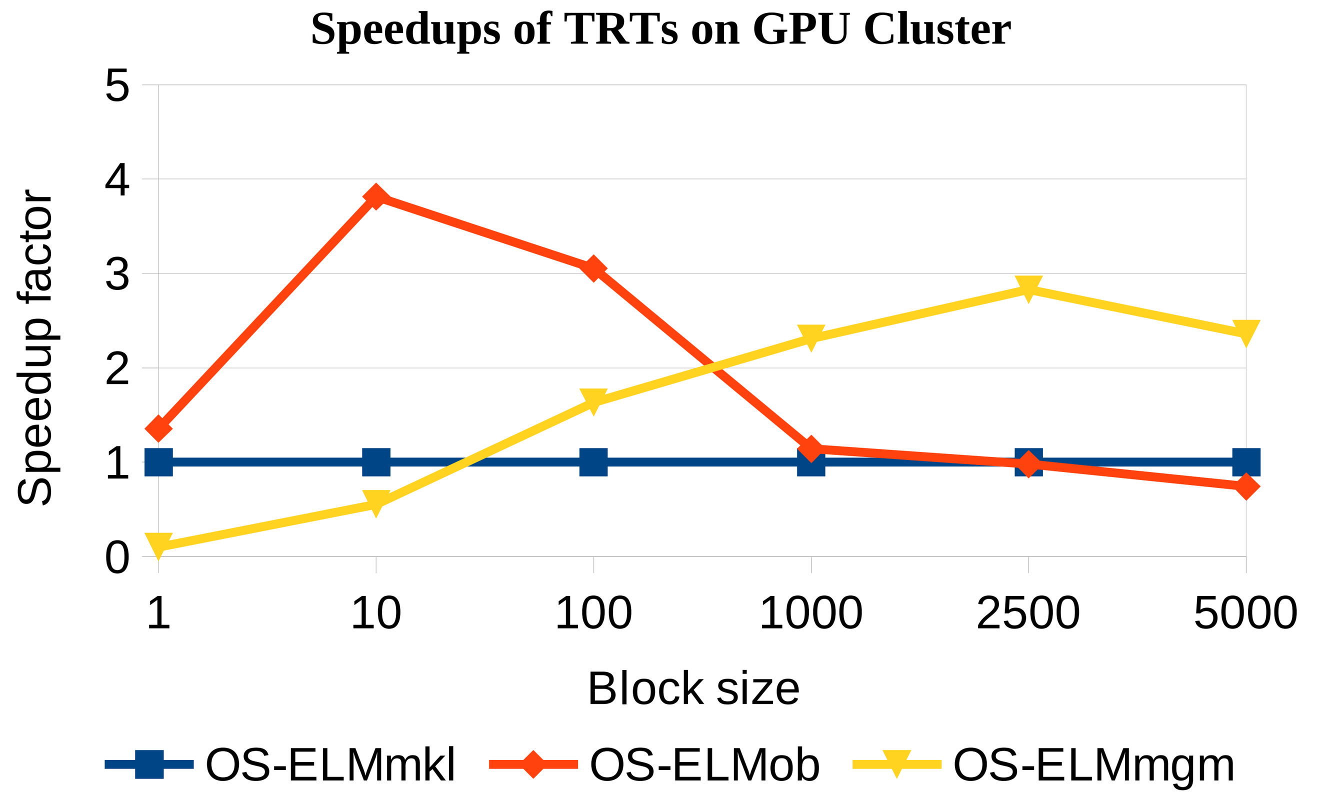
TRTs of Implementations executed on GPU Cluster

BS	OS-ELMob	OS-ELMmkl	OS-ELMmgm
1	9.90 ± 0.47	13.42 ± 0.88	134.95 ± 0.70
10	2.48 ± 0.02	9.47 ± 1.20	17.13 ± 0.07
100	2.29 ± 0.02	6.98 ± 0.90	4.27 ± 0.02
1000	10.31 ± 0.07	11.77 ± 0.39	5.09 ± 0.03
2500	34.92 ± 0.15	34.17 ± 0.62	12.08 ± 0.08
5000	105.33 ± 1.43	78.04 ± 1.69	33.08 ± 0.17

TRTs of OS-ELMmkl executed on Virtual Machines

BS	Small	Medium	Large	XL
1	5.07 ± 0.07	5.32 ± 0.59	5.20 ± 0.35	6.81 ± 0.24
10	3.54 ± 0.09	2.88 ± 0.15	2.49 ± 0.23	139.03 ± 7.09
100	4.57 ± 0.17	3.61 ± 0.72	12.45 ± 2.01	68.83 ± 3.16
1000	97.93 ± 0.70	54.15 ± 1.16	33.44 ± 0.44	22.56 ± 1.56
2500	542.61 ± 12.26	285.28 ± 2.75	161.27 ± 0.99	94.03 ± 2.54
5000	3697.97 ± 203.33	1082.12 ± 8.28	574.06 ± 8.13	330.40 ± 7.01

Speedups of TRTs on GPU Cluster



Speedups of TRTs on Virtual Machines