Research and development of methods for Data Stream Mining based on Meta Heuristic, Statistic and Nonparametric Learning

Thuong Pham Thi[*], Xuan Hoai Nguyen[†], Tri Thanh Nguyen[††]
[*]University of Information and Communication Technology - Thai Nguyen University,
[†]Hanu IT Research and Development Center - Hanoi University,
[††]VNU University of Engineering and Technology - Vietnam National University, Hanoi.

Introduction

Develop adaptive learning algorithms for Evolving Data Streams is one of the major challenges that we face today. In this research, we propose improved methods aim at answering three main research questions includes:

(1) What to remember or forget?;

(2) When to do the model upgrade?;

(3) How to do the model upgrade?

These proposed methods are based on Meta Heuristic, Statistic and Nonparametric Learning.

Objectives

1. Propose a new sampling method for the first research question.

2. Propose a new change detection method for the second research question and a new adaptive learning algorithm for the third research question. Experimental evaluation the proposed methods with existing methods.

3. Built a framework for Evolving Data Streams Learning.

Methods

This research is based on Backgrounds:

1. Meta heuristic methods[5][3]:

- Genetic Programming – GP,
- Multi-Object Optimal.

2. Statistic Learning [1][2][4]:

- Bootstrap sampling,
- Online Random Forests.

2. Non-parametric Bayesian Methods [6]:

For handing the big &complex data streams

Results

1. A new method to quantify the over-fitting in Genetic Programming

```
Algorithm 1: Quantify the over-fitting
    $overfit(0) = 0$
    $btp = test\_fit(0)$
    $tbtp = training\_fit(0)$
    for $generationi = 1$ to $n - 1$
        if $(test\_fit(i) < tbp)$
            $overfit(i) = 0$
            $btp = test\_fit(i)$
            $tbtp = training\_fit(i)$
        else $overfit(i) = test\_fit(i) - btp$
    $OV = \sqrt{overfit(n-1) * (n - btp)/2}$
    return $OV$
```

⇒ Provide a suite of 140 instances of symbolic regression benchmarks with various types of noise, levels of noise grouped into clusters by increasing difficult levels (OV).

Table 1: Name of data set

	Name of data set (Index of data set)			
Cluster 0 (C0)	Kei2.My	Kei12.Ly	Kei12.Hy	Kei12.Hxy
	Kei12.Hx	Kei12.Mx	Nguyen_4.Ly	Kei11.My
	Kei11.Ly	Nguyen_4.Hy	Kei11.Lxy	Vla1.Hx
	Nguyen_4.Hx			
Cluster 1 (C1)	Vla1.Lxy	Nguyen_4.My	Kei12.Lx	Vla8.My
	Vla1.Hy	Nguyen_4.Lx	Kei10.Hy	Nguyen_3.Mxy
	Vla5.Lxy	Kei13.Hy	Kei14.Mxy	Vla8.Hxy
	Vla8.Hy	Vla6.Mx	Kei14.Lx	Vla8.Mx
	Kei10.Mxy	Kei13.My	Vla1.Mx	Kei13.My
	Vla6.My	Nguyen_4.Mx	Vla6.Hy	Kei13.Mxy
	Nguyen_2.Mx	Vla5.My	Nguyen_2.Ly	Kei13.Ly
	Vla1.Lx	Vla1.Ly	Vla6.Ly	Kei15.Mx
	Vla8.Lx	Kei10.Mx	Kei10.Ly	Kei10.Ly
	Kei14.Lxy	Kei10.Lxy	Kei10.F	Kei11.F
	Kei12.F	Kei13.F	Kei14.F	Kei15.F
	Vla1.F	Vla5.F	Vla6.F	Vla8.F
	Nguyen_1.tr10	Nguyen_2.F	Nguyen_3.F	Nguyen_4. F
	Kei13.Lx	Kei15.Lx	Vla6.Lx	Nguyen_3.Ly
	Vla5.Ly	Vla8.Ly	Vla6.Lxy	Vla8.Lxy
	Nguyen_1.Lxy	Vla6.Mxy	Nguyen_1.Mxy	
Cluster 2 (C2)	Nguyen_1.My	Kei14.My	Vla5.Hy	Nguyen_4.Lxy
	Kei15.Hx	Kei11.Mxy	Kei14.Hy	Kei14.Ly
	Nguyen_2.Lx	Vla5.Hxy	Vla6.Hxy	Kei14.Hx
	Vla8.Hx	Kei14.Mx	Nguyen_1.Hx	Vla1.My
	Vla5.Mxy	Nguyen_1.Mx	Nguyen_2.Hxy	Nguyen_3.Mx
	Kei10.Hxy	Nguyen_1.Hxy	Kei14.Hxy	Vla6.Hx
	Kei15.Lxy	Kei13.Hx	Kei10.Hx	Vla5.Hx
	Vla1.Hxy	Kei15.Ly	Nguyen_2.Lxy	Vla5.Lx
	Kei10.My	Vla5.Mx	Vla1.Mxy	Kei13.Lxy
	Nguyen_3.Lxy	Kei12.Lxy		
Cluster 3 (C3)	Kei11.Hy	Nguyen_3.Hy	Nguyen_4.Hxy	Nguyen_4.Mxy
	Kei11.Mx	Nguyen_2.Hy	Kei12.Mxy	Kei15.Hy
	Kei11.Lx	Kei11.Hxy	Nguyen_2.Mxy	Kei15.Mxy
	Nguyen_2.My	Nguyen_3.Ly	Nguyen_3.Hx	Nguyen_1.Ly
	Kei15.My	Nguyen_3.My	Nguyen_1.Lx	Nguyen_1.Hy
	Vla8.Mxy	Kei13.Hxy	Nguyen_2.Hx	Nguyen_3.Hxy
	Kei15.Hxy	Kei11.Hx		

2. Propose a new fitness representation in GP (Stochastic Fitness):

$$Stochastic\ Fitness{:}Std(bias, variance) \sim Std(\mu, \sigma^2).$$

Table 2: P values, Fittest error on Benchmark problems in Cluster 0

	P value		Fittest		
Data set	GP	BVGP	GP	BVGP	SFGP-RS
C0.01	**0.0001(+)**	**0.0053(+)**	3.88E+01	3.77E+01	3.58E+01
C0.02	**0.0005(+)**	**0.0005(+)**	1.74E+01	1.73E+01	1.46E+01
C0.03	**0.0001(+)**	**0.0001(+)**	3.99E+01	3.98E+01	3.79E+01
C0.04	**0.0097(+)**	**0.0097(+)**	5.62E+01	5.61E+01	5.51E+01
C0.05	**0.0000(+)**	**0.0001(+)**	4.55E+01	3.99E+01	3.79E+01
C0.06	0.0824	**0.0000(+)**	31.35558	35.07208	30.329617
C0.07	**0.0000(+)**	**0.0000(+)**	0.811338	0.794747	0.446406
C0.08	**0.0000(+)**	**0.0000(+)**	1.15E+00	1.16E+00	1.08E+00
C0.09	**0.0000(+)**	**0.0000(+)**	9.00E-01	8.90E-01	7.82E-01
C0.10	**0.0000(+)**	**0.0000(+)**	1.080646	1.080721	0.9027656
C0.11	**0.0000(+)**	**0.0000(+)**	1.04E+00	1.01E+00	9.35E-01
C0.12	**0.0000(+)**	0.0578	1.87E+00	0.1173112	1.24E-01
C0.13	0.5010	**0.0000(+)**	0.917431	1.080646	0.9027656

Conclusion

In this research:

- The major research challenges and Objectives are listed
- Backgrounds for investigating the new methods are outlined.
- Some preliminary results are shown in Results section. However, these results focus on Meta Heuristics and Statistics Learning.

We will focus on three main research questions in the further works.

References

[1] Albert Bifed; *Adaptive Stream Mining: Pattern learning and Mining from Evolving Data Streams*; ISO press, 2010.

[2] Indrë Žliobaitë; *ADAPTIVE TRAINING SET FORMATION* (thesis); Vilnius, 2010.

[3] Lones, Michael. "Sean Luke: essentials of Metaheuristics." *Genetic Programming and Evolvable Machines* 12.3 (2011): 333-334.

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[5] John R Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.

[6] Zoubin Ghahramani, *Non-parametric Bayesian Methods Uncertainty in Artificial Intelligence Tutorial,* July 2005.