Learning Specifications for Labelled Patterns[1]

Nicolas Basset[1] Thao Dang[1] Akshay Mambakam[1] José Ignacio Requeno Jarabo[2]

[1]VERIMAG/CNRS, Université Grenoble Alpes, Grenoble, France

[2]Department of Computing, Mathematics, and Physics, Western Norway University of Applied Sciences (HVL), Bergen, Norway

28/07/2020

[1] This work has been partially supported by the ANR Project ANR-15-IDEX-02

Outline

1. Introduction
2. Parametric Pattern Predictors for Labelling Signals
3. Error Measures and ϵ-count
4. Specification Learning Framework
    - Upset, Downset and their intersection
    - Algorithm to compute Intersection of Upset and Downset
5. Results
    - Characterization of patterns
    - Classification of patterns
6. Future Work

Introduction

- Working with labelled patterns. Learning specifications.
- Parametric Pattern Predictors.
- Minimize false positives and negatives.

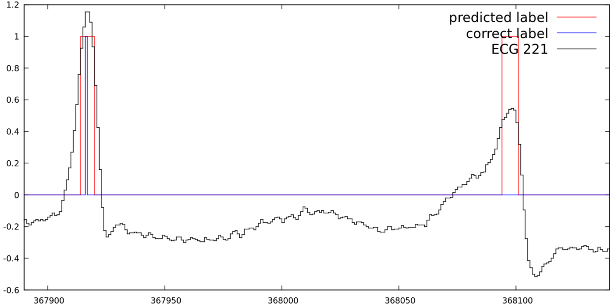


Figure 1: Labelled signal and prediction.

Parametric Predictors

- Labelling signal λ_s, a Boolean signal indicating the occurrence of a pattern.
- Parametric Pattern Predictor (Ψ_p): A parametric function that maps real-valued signals to Boolean-valued signals.
- Use extended STL (eSTL) having Min, Max operators.
- $\Psi^{ch}_{(p_1,p_2,p_3)} := ((\mathrm{Max}_{[-\mathrm{c},-\mathrm{p}_1]}\, s - \mathrm{Min}_{[-\mathrm{c},-\mathrm{p}_1]}\, s) \leq p_2) \ \wedge ((\mathrm{Max}_{[-\mathrm{p}_1,\mathrm{p}_1]}\, s) \geq -p_3) \wedge ((\mathrm{Max}_{[\mathrm{p}_1,\mathrm{c}]}\, s - \mathrm{Min}_{[\mathrm{p}_1,\mathrm{c}]}\, s) \leq p_2)$

False Positives, False Negatives and ϵ-count

- "How often" does a mismatch occur? How to quantify?
- The *false positive signal* indicates when the predictor predicts an occurrence when there is none.
- The *false negative signal* indicates when the predictor misses an actual occurrence.
- Lebesgue's measure or count edges?
- Lebesgue: Not convenient as a signal whose support is the disjoint union of many intervals of almost-null measure which are quite far apart gives a small measure.
- We want instead a big "count" because it can represent the number of mismatches.

Counting Edges is not Monotonic

s(t) < p. The Boolean signal $s(t) < 3$ is true on two intervals, while $s(t) < 2$ and $s(t) < 6$ are true on one interval.

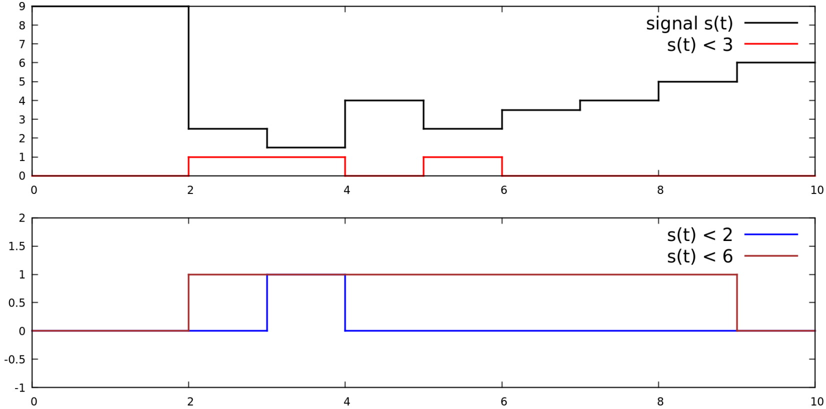


Figure 2: Non-monotonicity of interval count.

Frog Analogy for ϵ-count.

- For a Boolean signal w, the maximum number of ϵ-separated points that can be contained in $\mathtt{supp}(w)$ is ϵ-count.
- Close to the notion of ϵ-capacity by Kolmogorov et al [4].
- *Algorithm/Analogy:* A primordial one legged frog. `True` is land and `False` is ocean.
- It can jump by exactly ϵ when on land. It can swim any distance in water. The number of footsteps it leaves is ϵ-count.

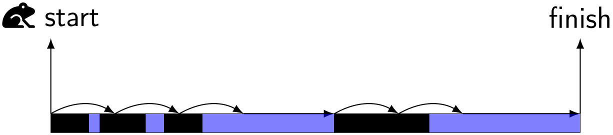


Figure 3

Specification Learning Framework

Parameter Identification: We formulate the problem as finding parameters for Parametric Pattern Predictor.

- $c_\epsilon(\neg\Psi_p(s) \land \lambda_s) \leq \mathtt{f}_-$
- $c_\epsilon(\Psi_p(s) \land \neg\lambda_s) \leq \mathtt{f}_+$
-

Specification Learning Framework

Parameter Identification: We formulate the problem as finding parameters for Parametric Pattern Predictor.

- $\texttt{Dom}-(\Psi, \mathcal{S}, \texttt{f}_-) = \{p \mid \forall (s, \lambda_s) \in \mathcal{S},\ c_\epsilon(\neg\Psi_p(s) \wedge \lambda_s) \leq \texttt{f}_-\}$
- $\texttt{Dom}+(\Psi, \mathcal{S}, \texttt{f}_+) = \{p \mid \forall (s, \lambda_s) \in \mathcal{S},\ c_\epsilon(\Psi_p(s) \wedge \neg\lambda_s) \leq \texttt{f}_+\}$
-

Specification Learning Framework

Parameter Identification: We formulate the problem as finding parameters for an Increasing Parametric Pattern Predictor (IPPP).

- $\mathtt{Dom}{-}(\Psi, \mathcal{S}, \mathtt{f}_-) = \{p \mid \forall (s, \lambda_s) \in \mathcal{S},\ c_\epsilon(\neg \Psi_p(s) \wedge \lambda_s) \leq \mathtt{f}_-\}$
- $\mathtt{Dom}{+}(\Psi, \mathcal{S}, \mathtt{f}_+) = \{p \mid \forall (s, \lambda_s) \in \mathcal{S},\ c_\epsilon(\Psi_p(s) \wedge \neg \lambda_s) \leq \mathtt{f}_+\}$
-

$\mathtt{Dom}-$ and $\mathtt{Dom}+$ are *upset* and *downset* respectively.

Specification Learning Framework

Parameter Identification: We formulate the problem as finding parameters for an Increasing Parametric Pattern Predictor (IPPP).

- $\texttt{Dom}-(\Psi, \mathcal{S}, \texttt{f}_-) = \{p \mid \forall (s, \lambda_s) \in \mathcal{S},\ c_\epsilon(\neg \Psi_p(s) \wedge \lambda_s) \leq \texttt{f}_-\}$
- $\texttt{Dom}+(\Psi, \mathcal{S}, \texttt{f}_+) = \{p \mid \forall (s, \lambda_s) \in \mathcal{S},\ c_\epsilon(\Psi_p(s) \wedge \neg\lambda_s) \leq \texttt{f}_+\}$
- $\texttt{DomInter}(\Psi, \mathcal{S}, \texttt{f}_+, \texttt{f}_-) = \texttt{Dom}+(\Psi, \mathcal{S}, \texttt{f}_+) \cap \texttt{Dom}-(\Psi, \mathcal{S}, \texttt{f}_-)$

$\texttt{Dom}-$ and $\texttt{Dom}+$ are *upset* and *downset* respectively.

Specification Learning Framework

Parameter Identification: We formulate the problem as finding parameters for an Increasing Parametric Pattern Predictor (IPPP).

- $\mathtt{Dom}-(\Psi, \mathcal{S}, \mathtt{f}_-) = \{p \mid \forall(s, \lambda_s) \in \mathcal{S},\ c_\epsilon(\neg\Psi_p(s) \wedge \lambda_s) \leq \mathtt{f}_-\}$
- $\mathtt{Dom}+(\Psi, \mathcal{S}, \mathtt{f}_+) = \{p \mid \forall(s, \lambda_s) \in \mathcal{S},\ c_\epsilon(\Psi_p(s) \wedge \neg\lambda_s) \leq \mathtt{f}_+\}$
- $\mathtt{DomInter}(\Psi, \mathcal{S}, \mathtt{f}_+, \mathtt{f}_-) = \mathtt{Dom}+(\Psi, \mathcal{S}, \mathtt{f}_+) \cap \mathtt{Dom}-(\Psi, \mathcal{S}, \mathtt{f}_-)$

$\mathtt{Dom}-$ and $\mathtt{Dom}+$ are *upset* and *downset* respectively.
The set of values ($\mathtt{f}_-$, $\mathtt{f}_+$) for which a solution exists ($\mathcal{P}$) is also an *upset*.

$$\mathcal{P}(\Psi, \mathcal{S}) = \{(\mathtt{f}_+, \mathtt{f}_-) \mid \mathtt{DomInter}(\Psi, \mathcal{S}, \mathtt{f}_+, \mathtt{f}_-) \neq \emptyset\}$$

Upset, Downset and their intersection

A set $\overline{X}$ is an *upset* if for all $p, q \in \mathbb{R}^n$ such that $p \leq q$ if $p \in \overline{X}$ then $q \in \overline{X}$.
A set $\underline{X}$ is a *downset* if for all $p, q \in \mathbb{R}^n$ such that $q \leq p$ if $p \in \underline{X}$ then $q \in \underline{X}$.

- We need a tool that help us represent and compute upsets.
- (Not) Surprisingly, ParetoLib [1] (Bahirkin et al. in FORMATS19) does exactly this.

Upset, Downset and their intersection

A set $\overline{X}$ is an *upset* if for all $p, q \in \mathbb{R}^n$ such that $p \leq q$ if $p \in \overline{X}$ then $q \in \overline{X}$.
A set $\underline{X}$ is a *downset* if for all $p, q \in \mathbb{R}^n$ such that $q \leq p$ if $p \in \underline{X}$ then $q \in \underline{X}$.
`DomInter` = solution set = intersection of an upset and a downset

- We need a tool that help us represent and compute upsets.
- (Not) Surprisingly, ParetoLib [1] (Bahirkin et al. in FORMATS19) does exactly this.
- We extend the tool to add an algorithm that computes the intersection of an upset and a downset.

Algorithm to compute Intersection of Upset and Downset

The main ideas are as follows:

- Parameter space, solution set and undecided region. Everything is a union of boxes!
- We start with the whole parameter space (again a box) in the undecided region. We repeatedly search and divide boxes.
- Choose a box in undecided region, do an improved binary search on the diagonal and split the box.
- Stop when the the undecided region gets smaller than a user-defined bound (V_δ).

Searching on a Line (Diagonal)

- Use a modification of binary search to find the points where the boundaries intersect the line. Exploit monotonicity.

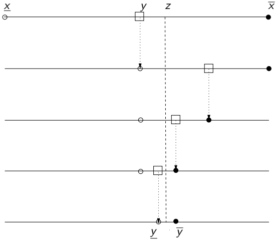
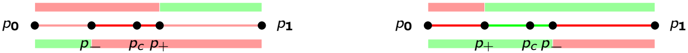


Figure 4: Binary search and the successive reduction.



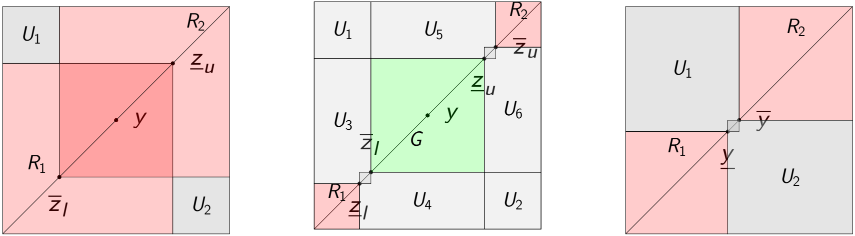
(a) Negative intersection.

(b) Positive intersection.

Figure 5: Intersection on a line.

Decomposing the box

- Deduce the green (positive intersection), red (no solution) and grey (undecided) regions by drawing hyper-rectangles (boxes).
- The number of boxes generated is at most quadratic in dimension of the parameter space.
- Add the grey boxes to the processing queue and repeat.



(a) Negative intersection

(b) Positive intersection

(c) No intersection found

Figure 6: Illustration of sub-boxes.

Experiments on ECG signals

- Electrocardiogram (ECG) signals capture information about electrical activity of the heart and can help detect anomalies in its functioning.
- We characterize several features (e.g. peaks, ditches and separation between them) as parametric specifications.
- Give solution set of parameters with the best possible trade-off between false positives and false negatives.

Characterization of ECG Pulses

- We use three ECGs (100, 123, 221) each containing between 1500 to 2500 labelled pulses taken from the MIT-BIH Arrhythmia Database of Physionet [3, 6]. Considering only the labels for normal pulses.
- Everything unlabelled is assumed not to be a normal pulse.
- $\Psi^{ch}_{(p_1,p_2,p_3)} := ((\mathrm{Max}_{[-\mathrm{c},-\mathrm{p}_1]}\, s - \mathrm{Min}_{[-\mathrm{c},-\mathrm{p}_1]}\, s) \leq p_2) \ \wedge ((\mathrm{Max}_{[-\mathrm{p}_1,\mathrm{p}_1]}\, s) \geq -p_3) \wedge ((\mathrm{Max}_{[\mathrm{p}_1,\mathrm{c}]}\, s - \mathrm{Min}_{[\mathrm{p}_1,\mathrm{c}]}\, s) \leq p_2)$
- For ECG-221, no false negatives (fn) and a single false positive (fp) [shown in Fig. 1].
- For ECG-123, it can match with fn=1 and fp=0.
- For ECG-100, neither the number of fp nor fn can go below 30.
- Reason: Expressed only the shape of the heart pulses not their rhythm.

Pareto Front Between $\mathtt{f}_-$ and $\mathtt{f}_+$

- We can modify the intersection algorithm to quickly query whether the solution set is empty for a given $(\mathtt{f}_-, \mathtt{f}_+)$.
- For ECG-100; Brown-Green corresponds to $V_\delta = 1\%$. Red-Brown corresponds to $V_\delta = 0.1\%$.

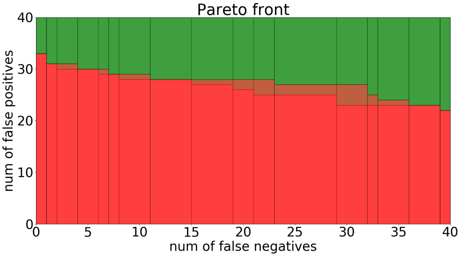


Figure 7: ECG 100, $V_\delta = 1\%$ vs $V_\delta = 0.1\%$

3D Intersection/Solution Set (ECG 221)

Once we have the Pareto front with adequate accuracy, we can explore the parameter space for different values of V_δ, $\texttt{f}_-$ and $\texttt{f}_+$.

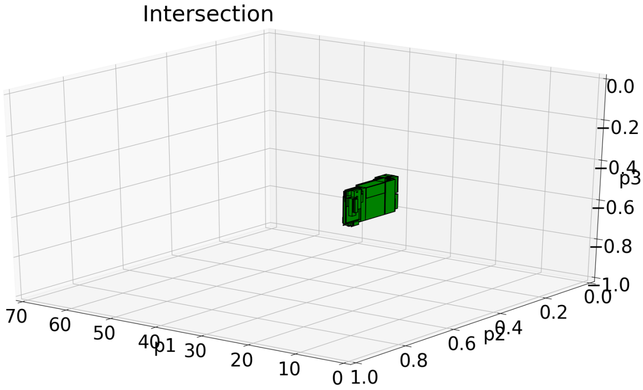


Figure 8: ECG 221,
$V_\delta = 0.01\%$, $\texttt{f}_- = 0$, $\texttt{f}_+ = 1$

3D Intersection/Solution Set (ECG 123)

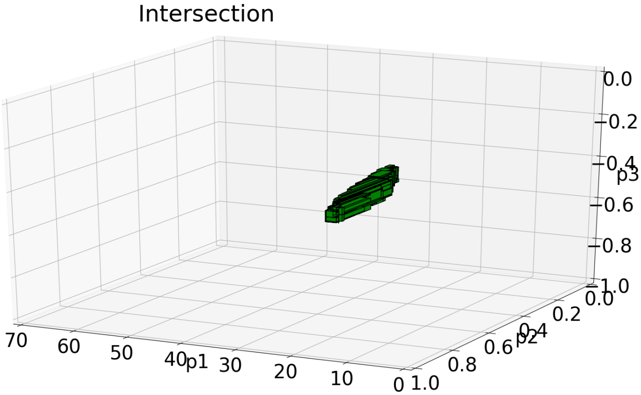


Figure 9: ECG 123,
$V_\delta = 0.01\%$, $\mathtt{f}_- = 1$, $\mathtt{f}_+ = 0$

Classification of ECG Pulses

ECGFiveDays dataset from the Time Series Classification Archive [2] of UCR (cousin of UC Berkeley).

- Two classes of ECGs taken 5 days apart from the same person.
- Find a classifier formula. Inspired by/Copied from [5] Mohammadinejad et al. ICCPS'20.

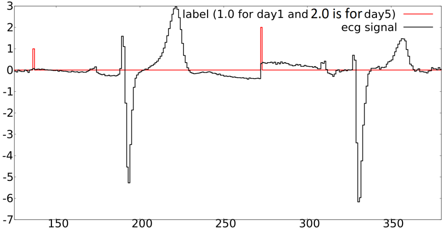


Figure 10: ECGs signals from $\mathtt{day}_1$ and $\mathtt{day}_5$

Visual Inspection, PSTL Features and Enumeration

We saw a ditch sandwitched between two peaks in Figure 10.

- Do visual inspection. Come up with PSTL feature specifications.
- Ranking ($m = |D_1 \triangle D_5| / |D_1 \cup D_5|$) and enumeration ($\varphi_i \vee \varphi_j$).
- $\Psi^{cl_1} := (\texttt{ditch} \wedge \mathrm{F}_{[p_1, 60-p_2]}\, \texttt{peak}) \vee (\ (\mathrm{Max}\ \ s\ U\ \texttt{ditch}) \geq p_3)$
- $\Psi^{cl_2} := (\texttt{ditch} \wedge \mathrm{F}_{[p_1, 60-p_2]}\, \texttt{peak}) \vee (\texttt{ditch} \wedge (\mathrm{Max}_{[0,60]}\ \ s) \leq -p_3)$

Table 1: Features and formulae.

Feature	Formula	D_1 for day$_1$	D_5 for day$_5$
Def. of peak	$(s \geq (\mathrm{Max}_{[-10,10]}\ \ s)) \wedge s \geq 1$	NA	NA
Def. of ditch	$(s \leq (\mathrm{Min}_{[-10,10]}\ \ s)) \wedge s \leq -1$	NA	NA
Depth of the ditch	$(\mathrm{Min}_{[0,136]}\ \ s) \leq p\ \{or \geq p\}$	(-6.12, -4.767)	(-6.51, -5.71)
Location of the ditch	$\mathrm{F}_{[\theta_1,\theta_2]}\, \texttt{ditch}$	(51.00, 58.99)	(51.00, 59.99)
Height of peak 1	$(\mathrm{Max}\ \ s\ \ \mathrm{U}\ \ \texttt{ditch}) \leq p\ \{or \geq p\}$	(1.01, 5.42)	(0.77, 3.81)
Location of peak 1	$\mathrm{F}_{[\theta_1,\theta_2]}\, \texttt{peak}$	(48.00, 56.99)	(0.00, 55.99)
Height of peak 2	$\texttt{ditch} \wedge ((\mathrm{Max}_{[0,60]}\ \ s) \leq p)\ \{or \geq p\}$	(1.25, 3.296)	(1.43, 2.58)
Location of peak 2	$\texttt{ditch} \wedge \mathrm{F}_{[\theta_1,\theta_2]}\, \texttt{peak}$	(25.00, 30.99)	(23.00, 26.99)

Classifiers Found and Their Accuracy

$\Psi^{cl_1}_{(28.3,11.0,4.0)}$ and $\Psi^{cl_2}_{(27.5,1.0,-1.3)}$ have error values 2/861 and 17/861 respectively on the original testing set.

Table 2: Accuracy and performance results

Configuration	time (s)			Testing error		Training error	
	$\delta = 10^{-1}$	$\delta = 10^{-2}$	$\delta = 5.10^{-3}$	Ψ^{cl_1}	Ψ^{cl_2}	Ψ^{cl_1}	Ψ^{cl_2}
Confg. 1 (23, 861)	2	184	787	2/861	17/861	0/23	0/23
Confg. 2 (100, 761)	1.5	6	10	2/761	17/761	0/100	0/100
Confg. 3 (300, 561)	2	3	5	2/561	NA	0/300	NA
Confg. 4 (861, 0)	13	79	153	NA	NA	NA	NA
Confg. 5 (861, 0)	5	8.5	12	0/0	NA	2/861	NA

NA: Not Applicable. Parameter search is unsuccessful.

Future Work

- Computation of the exact or approximate solution sets for non-monotonic parametric specifications.
- Trade-offs among parameters and also between tightness and robustness. Tightest parameters for the given training examples might not generalize well.
- $\mathrm{F}_{[\tau_1,\tau_2]}\,\varphi$ is monotonic but $\tau_1 \leq \tau_2$. Use polyhedra?
- New: Timed Regular Expressions (TRE) + STL ? TRE can use polyhedra for time zones.

References I

Alexey Bakhirkin, Nicolas Basset, Oded Maler, and José Ignacio Requeno.
ParetoLib: A python library for parameter synthesis.
In *Proceedings of the 17th International Conference on Formal Modeling and Analysis of Timed Systems*, volume 11750 of *Theoretical Computer Science and General Issues*, pages 114–120, Cham, 2019. Springer.

Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML.
The UCR time series classification archive, October 2018.
`https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`.

References II

Ary L. Goldberger, Luis A.N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley.
PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals.
Circulation, 101(23):e215–e220, 2000.

Andrey N. Kolmogorov and Vladimir M. Tikhomirov.
ε-entropy and ε-capacity of sets in function spaces.
Uspekhi Matematicheskikh Nauk, 14(2(86)):386, 1959.

Sara Mohammadinejad, Jyotirmoy V. Deshmukh, and Aniruddh G. Puranic.
Mining environment assumptions for cyber-physical system models.
In *Proceedings of the 11th ACM/IEEE International Conference on Cyber-Physical Systems (to appear)*, ICCPS'20. IEEE, 2020.

References III

George B Moody and Roger G Mark.
The impact of the MIT-BIH arrhythmia database.
IEEE Engineering in Medicine and Biology Magazine, 20(3):45–50, 2001.