Interval Temporal Random Forests with an application to Covid-19 Diagnosis

Presented as a part of end term project review for IE 506:

Team Name: PriAlgo Prowess
Members: Priyansu (23N0464) , Rubul Gogoi (23n0462)

Outline

Problem Overview

- The paper aims to develop a machine learning method for distinguishing COVID-19-positive individuals from negative cases based on cough/breath recordings

- Utilizes **temporal random forests** to analyze multivariate **time series data** obtained from cough/breath samples.

- Aims to create a screening tool for real-time diagnosis, offering a simpler alternative to complex computations.

- Focuses on producing interpretable results that can be visualized and even transformed into audible sounds for easier recognition of positive cases.

Motivation

- Medical Diagnosis Advancement
- Innovative Application of Temporal Symbolic Learning
- Contribution to Symbolic Learning Methods
- Interpretability and Visualization of Results

Approach

Time series classification model that leverages:

- **TCART algorithm** (Temporal Classification and Regression Trees)

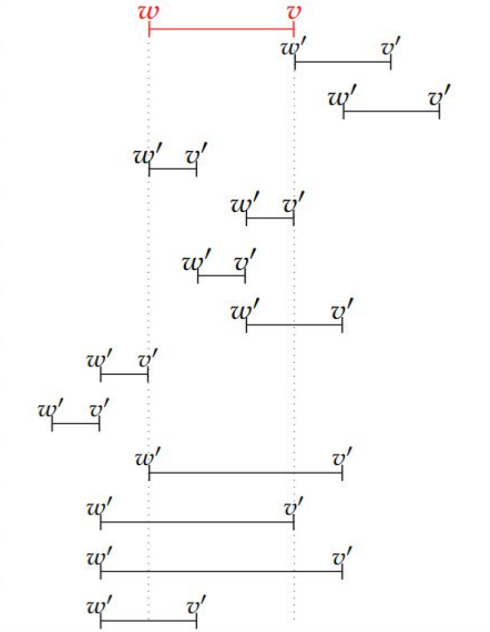
  A time series version of the CART algorithm

- **Allen's Temporal Relations** for structured temporal reasoning,

  A formal way to describe relationship between two time intervals

Allen's relations

$\mathcal{HS}$ modality	Definition w.r.t. the interval structure	Example
$\langle A \rangle$	$[w,v]\mathcal{R}_A[w',v']$ iff $v = w'$	
$\langle L \rangle$	$[w,v]\mathcal{R}_L[w',v']$ iff $v < w'$	
$\langle B \rangle$	$[w,v]\mathcal{R}_B[w',v']$ iff $w = w' \wedge v' < v$	
$\langle E \rangle$	$[w,v]\mathcal{R}_E[w',v']$ iff $v = v' \wedge w < w'$	
$\langle D \rangle$	$[w,v]\mathcal{R}_D[w',v']$ iff $w < w' \wedge v' < v$	
$\langle O \rangle$	$[w,v]\mathcal{R}_O[w',v']$ iff $w < w' < v < v'$	
$\langle \overline{A} \rangle$	$[w,v]\mathcal{R}_{\overline{A}}[w',v']$ iff $[w',v']\mathcal{R}_A[w,v]$	
$\langle \overline{L} \rangle$	$[w,v]\mathcal{R}_{\overline{L}}[w',v']$ iff $[w',v']\mathcal{R}_L[w,v]$	
$\langle \overline{B} \rangle$	$[w,v]\mathcal{R}_{\overline{B}}[w',v']$ iff $[w',v']\mathcal{R}_B[w,v]$	
$\langle \overline{E} \rangle$	$[w,v]\mathcal{R}_{\overline{E}}[w',v']$ iff $[w',v']\mathcal{R}_E[w,v]$	
$\langle \overline{D} \rangle$	$[w,v]\mathcal{R}_{\overline{D}}[w',v']$ iff $[w',v']\mathcal{R}_D[w,v]$	
$\langle \overline{O} \rangle$	$[w,v]\mathcal{R}_{\overline{O}}[w',v']$ iff $[w',v']\mathcal{R}_O[w,v]$	



CART algorithm

In the traditional CART algorithm, decisions involved in splitting a tree,

1. **Feature**
2. **Threshold** : to split the data into two subsets (e.g., $x \leq t$, $x > t$).
3. **Evaluate the quality of split** : Gini Impurity, Entropy & Information gain

TCART algorithm

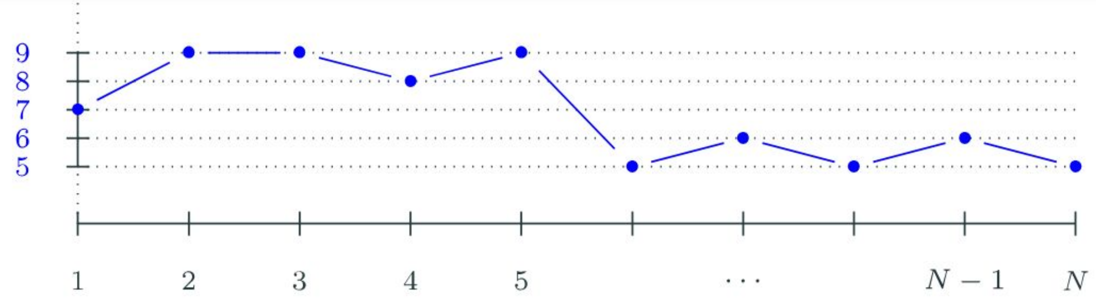
In the TCART algorithm, we have additional three more decisions to make,

1. **Best Reference Interval** : A specific time interval in the data to anchor the decision-making process.

2. **Best Relation** : One of Allen's temporal relations (e.g., *before*, *after*, *during*, etc.) that defines how other intervals relate to the reference interval.

3. **Gamma Value (γ \ gamma)** : A threshold percentage of values satisfying the condition (e.g., being greater than a threshold) in intervals that match the temporal relation.

TCART algorithm



In the above time series T:

- $T, [1,2] \Vdash \langle A \rangle (A >_{0.75} 8)$ because $\exists [2,5]$ such that $[1,2] R_A [2,5]$ and

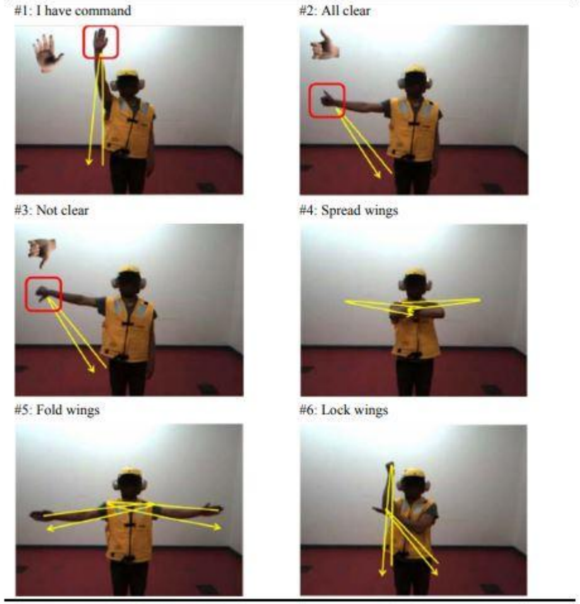
$$\frac{|\{t \mid 2 \leq t \leq 5 \text{ and } A(t) > 8\}|}{5-2+1} = \frac{3}{4} = 0.75;$$

- $T, [3,5] \nVdash \langle L \rangle (A >_{0.2} 7)$ that is $T, [3,5] \Vdash [L](A \leq_{0.2} 7)$;
- $T, [N-1, N] \nVdash \langle \overline{L} \rangle (A \leq_{1.0} 4)$ that is $T, [N-1, N] \Vdash [\overline{L}](A >_{1.0} 4)$.

Naatops Dataset

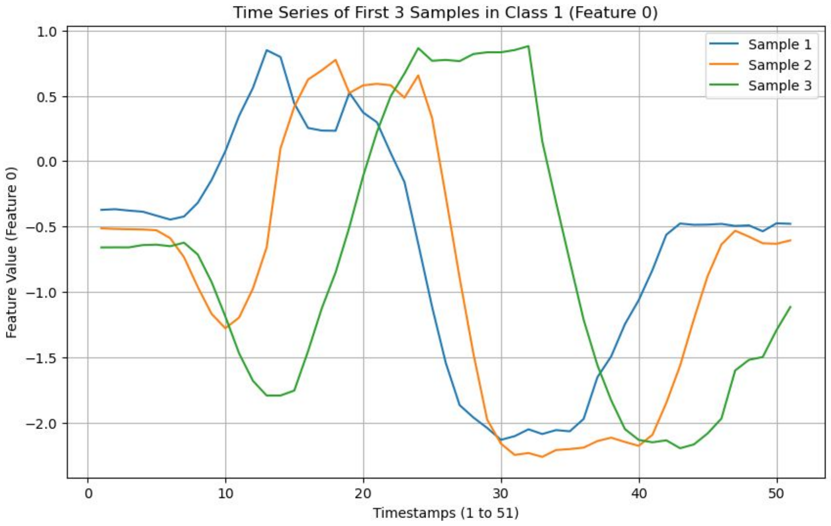
- The dataset has **360 samples**.
- Each sample can be classified in one of the following six categories :
  1. **I have command** 2. **All clear** 3. **Not clear**
  4. **Spread wings** 5. **Fold wings** 6. **Lock wings**
- The data is generated by sensors on the **hands**, **elbows**, **wrists** and **thumbs**.
- The data are the **x,y,z** coordinates for each of the **eight locations.**
- Each sample has **24 features**.
- Each feature was recorded at **51 timestamps** (from 1 to 51)

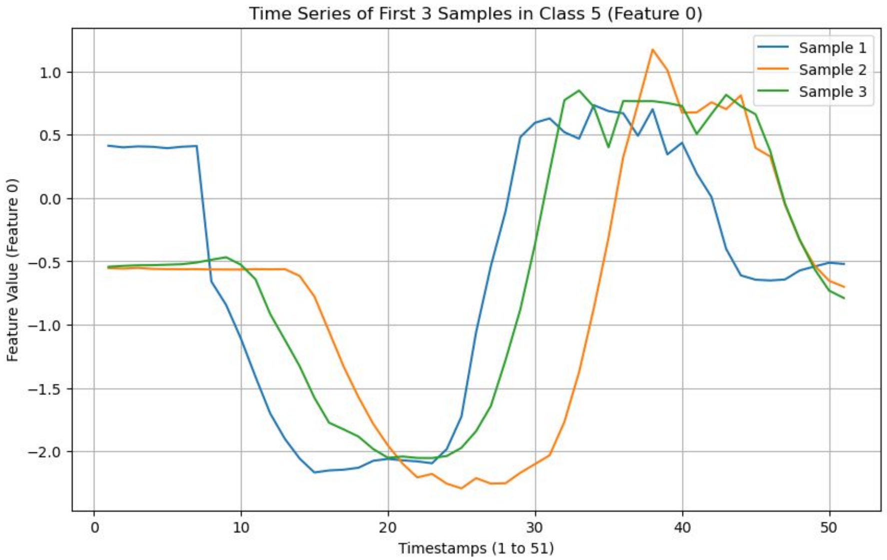
Table 1 The six selected gestures from the NATOPS database; figures have been extracted from (see online version for colours)



Source: Song et al. (2011)

Naatops Dataset





Example of a split

Here, Best feature = 1

Best threshold = 2.051978

Best reference interval = (10,15)

Best allen's relation = Later

```
(16, 2, 30) [1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1]
(39, 2, 30) [2, 2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 1, 2, 1, 2, 2, 2, 1, 1, 2, 2, 2, 2, 1, 1, 2, 2, 1, 1, 1, 2, 1, 1, 1]
1 (10, 15) L 2.051978
16 39
```

Experiments & Results

Excluding Atemporal relations

Actual \ Predicted	I have command	All clear	Not clear	Spread wings	Fold wings	Lock wings	Total
I have command	46	3	4	2	3	2	60
All clear	6	42	5	3	2	2	60
Not clear	5	7	41	3	2	2	60
Spread wings	3	4	6	43	3	1	60
Fold wings	3	2	4	6	42	3	60
Lock wings	3	3	3	4	5	42	60
Total	60	60	60	60	60	60	360

Accuracy: 71.1%

Macro Precision: 71.4%

Macro Recall: 71.1%

Experiments & Results

Including Atemporal relations : Assigned atemporal features using a probabilistic approach

Atemporal features = ('cough', 'breath', 'short_breath', 'headache', 'fever')

Actual \ Predicted	I have command	All clear	Not clear	Spread wings	Fold wings	Lock wings	Total
I have command	47	3	3	2	3	2	60
All clear	5	44	4	3	2	2	60
Not clear	4	6	43	3	2	2	60
Spread wings	3	3	5	45	3	1	60
Fold wings	3	2	3	5	44	3	60
Lock wings	3	2	3	3	4	45	60
Total	60	60	60	60	60	60	360

Accuracy: 73.4%%   Macro Precision: 74.6%   Macro Recall: 74.6%

Conclusions

- Highlights the significance of interpretability and explainability in machine learning, especially in medical applications.
- It introduces Interval Temporal Random Forests as a novel approach for diagnosing COVID-19 from cough/breath samples.
- Future research aims to generalize symbolic learning methods, enhance interpretation techniques, and explore multi-dimensional data analysis.
- The ultimate goal is to develop clinically useful rules for COVID/Non-COVID classification to combat the pandemic effectively.

Future Work

- Conduct clinical studies to validate the usefulness of the developed methodologies in real-world settings for COVID-19 diagnosis and potentially other medical applications.

- Develop techniques for enhancing the interpretability of temporal random forests without sacrificing performance.

References

[1] A. Liaw and M. Wiener. Classification and regression by RandomForest. *R News*, 2(3):18–22,2002.

[2] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[3] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*.Wadsworth PublishingCompany, 1984.

[6] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In Proc. of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 3474–3484, 2020.

[5] M. Kudo, J. Toyama, and M. Shimbo. Multidimensional curve classification using PassingThrough regions. Pattern Recognition Letters, 20(11):1103–1111, 1999. (https://dl.acm.org/doi/10.1145/3394486.3412865)

[6] Y. Yamada, E. Suzuki, H. Yokoi, and K. Takabayashi. Decision-tree induction from time-series data based on a standard- example split test. In Proc. of the 12th International Conference on Machine Learning, page 840–847. AAAI Press, 2003(https://www.researchgate.net/publication/221346278_Decision-tree_Induction_from_Time-series_Data_Based_on_a_Standard-example_Split_Test)

[7] CART Algorithm(https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/)