IE-506 Machine Learning - Principles and Techniques **Jan-May,2024**

Project End-term Report : Interval Temporal Random Forests with an application to Covid-19 Diagnosis

Team Name : PriAlgo Prowess Team Members : Priyansu

Abstract

This project report presents an investigation into the use of interval temporal random forests for diagnosing COVID-19 based on time series data derived from cough and breath recordings. The approach leverages interval temporal logic, offering a logical and interpretable framework for analyzing temporal data. Building on the interval temporal logic decision tree extraction algorithm, this project extends it to interval temporal random forests, which maintain logical interpretability while incorporating a functional component.

Due to limited access to the dataset used in the referenced paper, a time series dataset from a public repository is used for experimentation.The performance of the models is assessed in terms of accuracy and sensitivity.

1. Introduction

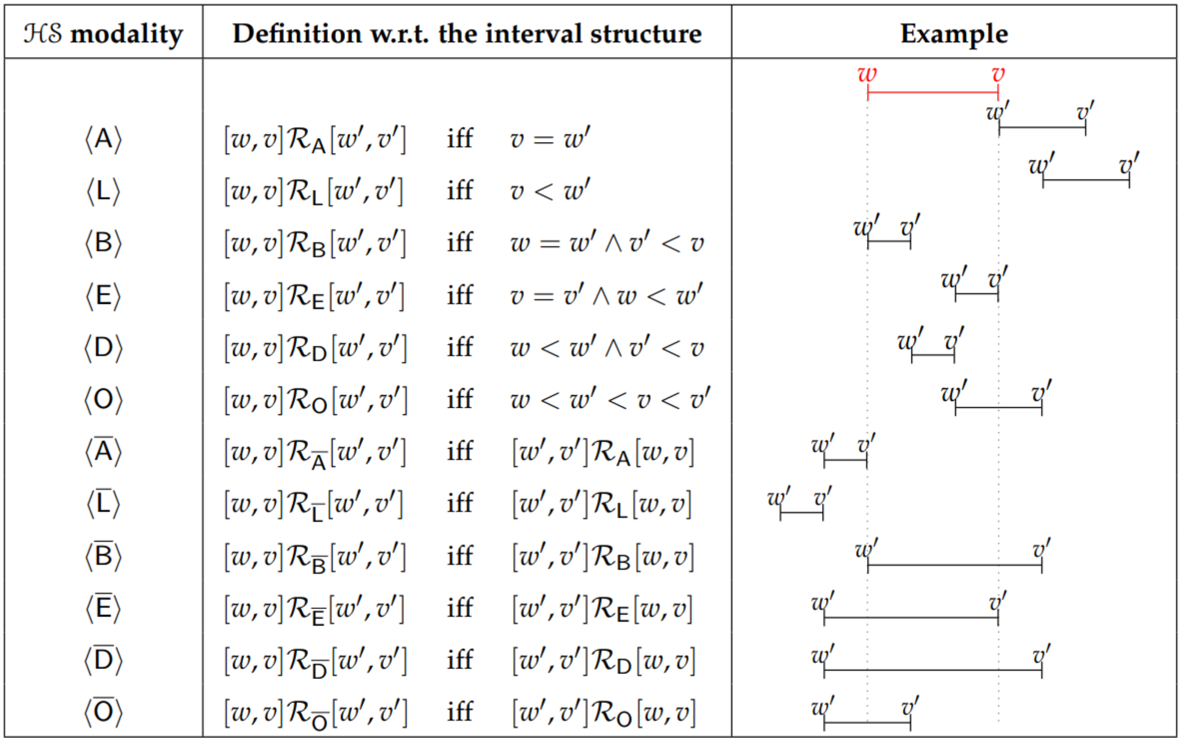
Machine Learning (ML) is a cornerstone of modern Artificial Intelligence, automating the extraction and expression of underlying theories from data for various applications. ML encompasses both functional and symbolic learning, with the former focusing on learning functions that represent underlying theories, and the latter on learning logical descriptions. While functional learning often provides high accuracy, symbolic learning offers interpretability, making it easier for humans to understand and explain the results.

Traditionally, symbolic learning was limited to propositional logic, which constrained its ability to handle temporal, spatial, and non-propositional data. However, recent advancements, such as interval temporal logic decision trees, have expanded the expressive power of symbolic methods, particularly in classifying multivariate time series data. These trees are part of a larger effort to develop modal symbolic learning, which uses propositional modal logic to enhance the capabilities of symbolic learning.

For these problems ,an approach to interval temporal random forests is presented, which extends the principles of interval temporal logic decision trees to a forest of trees. By building on the idea of sets of trees being more performant than single trees, interval temporal random forests offer a powerful method for multivariate time series classification.

2. Notation : Interval Temporal Logic

Several interval temporal logics have been proposed recently [18], but Halpern and Shoham's (HS) [19] is the most widely studied due to its natural representation of temporal intervals. In the context of finite domains, we consider a finite, initial subset of natural numbers, denoted by [N] = {1, 2, ..., N}, where N > 1. A strict interval over [N] is represented as [x, y], where x, y ∈ [N] and x < y. There are 12 different binary ordering relations between two strict intervals on a linear order, known as Allen's interval relations [1], including relations like adjacent to (RA), later than (RL), begins (RB), ends (RE), during (RD), and overlaps (RO).



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

FIGURE 4.8: Allen's interval relations and $\mathcal{HS}$ modalities.

Halpern and Shoham's interval temporal logic (HS) [19] is a multi-modal logic that includes a set AP of atomic propositions, propositional connectives $\vee$ and $\neg$, and a modality for each Allen's interval relation. Formulas in HS are generated by the grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle X\rangle\varphi,$$

where $p \in AP$ and $X \in X$. Other propositional connectives and constants (e.g., $\wedge$, $\rightarrow$, and $\top$) can be derived in the standard way. The strict semantics of HS is given in terms of interval models $T = \langle I([NT\ ]), V\rangle$, where $[NT\ ]$ is a finite linear order, $I([NT\ ])$ is the set of all (strict) intervals over $[NT\ ]$, and V is a valuation function that assigns to every atomic proposition $p \in AP$ the set of intervals $V(p)$ on which p holds. The truth of a formula φ on a given interval $[x, y]$ in an interval model T, denoted by $T, [x, y] \Vdash \varphi$, is defined by structural induction on the complexity of formulas.

In summary, HS provides a formalism for reasoning about temporal intervals in a finite domain, allowing for the expression of complex temporal relationships using a set of modalities corresponding to Allen's interval relations.

3. Methods and Approach

3.1 Temporal Decision Trees

In the context of temporal data sets described by n attributes $\{A_1,\dots,A_n\}$, a propositional alphabet AP is defined to represent temporal constraints. Each proposition in AP has an interval semantics, evaluated over time intervals, reflecting the continuous nature of time series data. For a time series T and a time point t, $A(t)$ denotes the value of attribute A at time t, and $dom(A)$ is the domain of A. Propositions in AP take the form $A \bowtie \sim \gamma a$, where A is an attribute, $\bowtie$ and $\sim$ are comparison operators, and $a \in dom(A)$. These propositions are evaluated over intervals of time, allowing for fuzzy constraints based on a fraction γ of values in the interval satisfying the constraint.

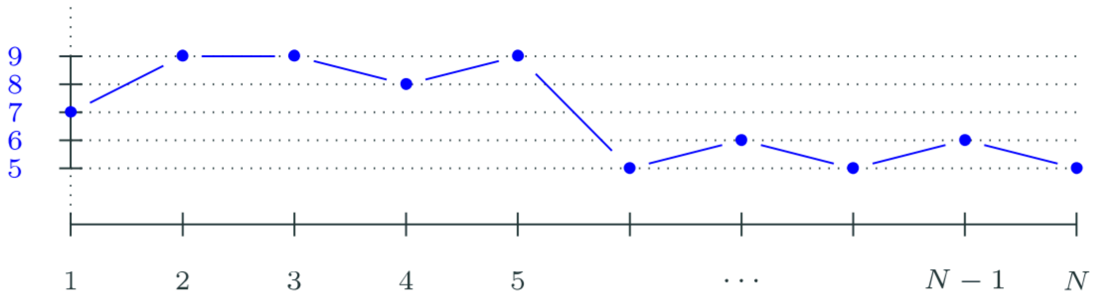
Temporal decision trees (TDTs) are rooted trees where each non-leaf node represents a decision based on a temporal or atemporal proposition, and each leaf node represents a class label. A TDT is defined by the grammar:

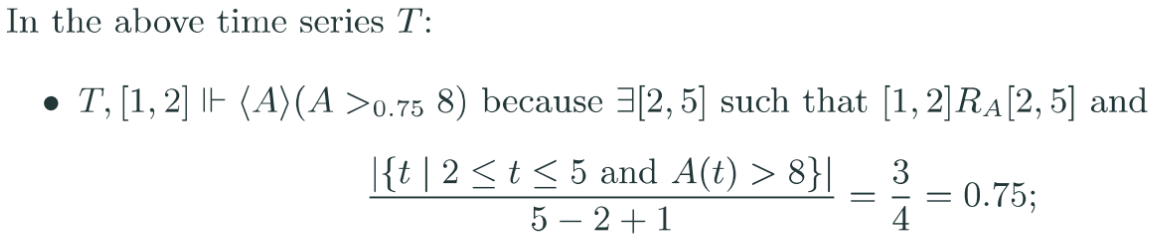
$$\tau ::= (S{=}\wedge\tau)\vee(\neg S{=}\wedge\tau)\vee(S\diamond\wedge\tau)\vee(\neg S\diamond\wedge\tau)\vee C$$

where $S{=}\in S{=}$ is an atemporal decision, $S\diamond\in S\diamond \in S$ is a temporal existential decision, and $C\in C$ is a class label.

To improve the overall experimental complexity in the TRF approach, we build a hash table keyed on the tuple (T, [x , y] ,X ,A , a , ~) for specific values of γ. This hash table returns the truth value of the decision ⟨X⟩(A ▹◃~γ a).

An Example of taking a Decision





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

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

Here we have one sample and one feature and the feature is a N point time series.

The above notation says that, there exists an interval related to the reference interval [1 ,2] with the relation ‘A’ (After) in which 75% of the values are greater than 8.

Unlike the static case, we do not ask if A ▹◃ a only in the current interval, but also if there exists an interval, related to the current one. Here ▹◃ ∈ { ≤ , > , = }.

Moreover, we may relax the requirement A ▷◁ a over a given interval [x , y] by asking that at least a certain fraction of the values of A in the interval [x , y] meet the condition denoted by A ▷◁$_{\gamma}$ a with γ ∈ (0,1] ⊂ R .

Information Based Learning

Information-based learning for decision trees has a rich history, with foundational work by Belson [4] and subsequent developments like the algorithm proposed in [32], which was the first implementation of a decision tree for classification.
It is used to measure the effectiveness of a particular decision in classifying the data.

Information-based learning is a general, greedy, sub-optimal approach to decision tree induction, as optimal decision tree induction is known to be NP-hard [22]. Entropy-based learning is a specific form of information-based learning and is the most common approach. It works by calculating the information or entropy of a dataset T with l distinct classes, denoted as Info(T), using the formula:

$$\mathrm{Info}(T) = -\textstyle\sum_{i=1}^{l} \pi_i \log \pi_i$$

Here, π_i represents the fraction of instances labeled with class C_i in dataset T. The entropy is a measure of the impurity or randomness in the dataset with respect to the class values.

In the context of binary trees, the main operation is splitting, which is done over a specific attribute A, a threshold value a ∈ dom(A), a value γ, and the operators ~ and ▷◁. Let S(A, a, γ, ~, ▷◁) represent the decision entailed by these parameters, and let (Te, Tu) be the partition of T entailed by this decision. The splitting information of S is calculated as:

$$\mathrm{InfoSplit}(T, S) = \tfrac{|Te|}{|T|}\mathrm{Info}(Te) + \tfrac{|Tu|}{|T|}\mathrm{Info}(Tu)$$

The entropy gain of a decision S is then defined as:

$$\mathrm{InfoGain}(T, S) = \mathrm{Info}(T) - \mathrm{InfoSplit}(T, S)$$

3.2 Temporal Random Forests

Temporal random forests (TRFs) extend the concept of random forests to the temporal domain. Similar to their propositional counterparts, TRFs are collections of decision trees trained on different subsets of the training data and attributes. Each tree in a TRF is a temporal decision tree, and the final classification decision for a given instance is determined by a voting aggregation function applied to the individual tree's predictions.

The key difference between a single decision tree and a random forest lies in the learning algorithm and the voting mechanism. While a single decision tree follows a deterministic path, a random forest introduces randomness in the selection of subsets of the training data and attributes, making it a hybrid symbolic-functional approach.

TRFs have been implemented using a generalized version of TCART, which allows for the use of a limited number of attributes and modal operators to find the best split at each step. This approach reduces the computational complexity compared to using the full set of attributes and operators. Additionally, TRFs utilize a preprocessing step to build a hash table that stores the truth values of decisions for specific parameter values, improving the efficiency of evaluating decisions during the learning process.

Overall, TRFs offer a powerful tool for temporal data analysis, leveraging the strengths of random forests in handling complex datasets while incorporating temporal semantics for improved decision-making.

4. Work Done Before Mid-term

Before the mid-semester, I acquired the "NATOPS'" time series dataset from GitHub and undertook preprocessing to derive two distinct datasets, Dataset1 and Dataset2.

Dataset1 comprises 120 samples, each with a single feature recorded 30 times and two target values. Employing the TCART algorithm for single decision trees on Dataset1 yielded promising results. The single decision tree achieved an accuracy of 97%, a precision of 0.94, sensitivity of 1, and specificity of 0.93. On the other hand, employing a Random Forest approach with 5 decision trees resulted in an accuracy of 93%, a precision of 1, sensitivity of 0.875, and specificity of 1.

Dataset2 consists of 120 samples, each containing 10 features recorded 30 times, with two target values. Again, utilizing the TCART algorithm for single decision trees on Dataset2 demonstrated excellent performance, achieving an accuracy of 100%, precision of 1, sensitivity of 1, and specificity of 1.

The successful application of the TCART algorithm for single decision trees and the TRF algorithm for random forests on these datasets showcases their effectiveness in handling time series data, promising avenues for further exploration and application in similar contexts.

5. Work Done After Mid-term

- On the temporal dataset used in mid-sem presentation, we assigned atemporal features using a probabilistic approach, where each class label was associated with specific probabilities for different atemporal features ('cough', 'breath', 'short_breath', 'headache', 'fever')
- Applied the TCART algorithm in two new time series data set
- Applied the TCART algorithm in different threshold values and different relations

6. Experiments And Results

6.1 Experiment 1

Original Dataset

Source: github
dataset id : “NATOPS”
No of samples = 360
No of features = 24
Time points = 51
Class labels = 6

Experimented Dataset

No of samples = 200
No of features = 8
Time points = 30
Class labels = 6

Accuracy =96.09%

6.2 Experiment 2

Assigned atemporal features using a probabilistic approach.
Atemporal features include cough, breath, short_breath, headache and fever.

Experimented Dataset
No of samples = 120
No of features = 2
Features used = 2 and 3 (and the atemporal features)
Time points = 30
Class labels = 3

Confusion Matrix: [[2 4 1]
[6 8 0]
[0 1 8]]

Here is a comparison of the results while including and excluding the atemporal features

Including Atemporal features

Covariance matrix

$$\begin{bmatrix} 2 & 4 & 1 \\ 6 & 8 & 0 \\ 0 & 1 & 8 \end{bmatrix}$$

Precision: [0.25 0.62 0.89]
Recall: [0.28 0.57 0.89]
F1 Score: [0.27 0.59 0.89]
Accuracy = 60%

Excluding Atemporal features

$$\begin{bmatrix} 2 & 4 & 1 \\ 9 & 5 & 0 \\ 0 & 1 & 8 \end{bmatrix}$$

Precision: [0.18 0.50 0.89]
Recall: [0.28 0.36 0.89]
F1 Score: [0.22 0.42 0.89]
Accuracy : 50%

6.3 Experiment 3

Original Dataset
Source: github
dataset id : “LSST”
No of samples = 4925
No of features = 6
Time points = 36
Class labels = 14

Experimented Dataset
No of samples = 70
No of features = 3
Time points = 26
Class labels = 3
Used threshold = 0.7

Accuracy =96.7%

6.4 Experiment 4 (tried less than threshold = 0.75)

Original Dataset

Source: github
dataset id : “Cricket”
No of samples = 180
No of features = 6
Time points = 1197
Class labels = 12

Experimented Dataset

No of samples = 60
No of features = 6
Time points = 40
Class labels = 4

Accuracy =93%

Confusion Matrix =

$$\begin{bmatrix} 0 & 0 & 2 & 0 \\ 0 & 6 & 0 & 0 \\ 2 & 0 & 4 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix}$$

```
Precision:  [0.      1.      0.67     1.      ]
Recall:     [0.      1.      0.67     1.      ]
f1 score:   [0.      1.      0.67     1.      ]
```

Here it can be seen that the second and the third classes were predicted accurately while the first class had zero precision and recall.

6.5 Experiment 5 (tried greater than threshold = 0.75)

Original Dataset

Source: github
dataset id : “Cricket”
No of samples = 180
No of features = 6
Time points = 1197
Class labels = 12

Experimented Dataset

No of samples = 60
No of features = 6
Time points = 40
Class labels = 4

Accuracy =97%

6. Conclusion

- 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.

6. 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 Publishing Company, 1984.

[6] Dataset "git+https://github.com/timeseriesAI/tsai.git"