Falling Rule List

CS 594, Provenance & Explanations,

Prof. Boris Glavic

Leonardo Borgioli

October 15, 2024

UNIVERSITY OF ILLINOIS CHICAGO

Index

- **Introduction:** Context, Falling Rule Lists, Paper Objectives, Proposed Method
- **Background:** Classic Discrete Distributions, Bayesian Inference, Point Estimators
- **Training Falling Rule List:** Model parameters, Likelihood, Prior, Mining Algorithm, Posterior, Summary
- **Experiments:** Predicting Hospital Readmission, Performance on Public Datasets
- **Conclusion:** Positive aspects of the paper, the negative aspects of the paper

Introduction

Context, Falling Rule Lists, Paper Objectives, Proposed Method

Context

Introduction

- In **healthcare**, patients and action need to be **prioritized** based on **risk**.
- Most **at-risk** patients should be handled **first**
- **Tradition** paradigm of predictive **models** does not contain such logic
- Often, models also lack **interpretability**.
- **Gap** between what we want to **achieve** with a model and what **can be achieved** with it

Falling Rule List

Introduction

- **Ordered** list of **if-then** rules, **sorted** by an importance criteria.
- **Estimated** probability of success **decreases monotonically** down the list

	Conditions		Probability	Supp.
IF	IrregularShape AND Age ≥ 60	THEN risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age ≥ 45	THEN risk is	78.13%	64
ELSE IF	IlDefinedMargin AND Age ≥ 60	THEN risk is	69.23%	39
ELSE IF	IrregularShape	THEN risk is	63.40%	153
ELSE IF	LobularShape AND Density ≥ 2	THEN risk is	39.68%	63
ELSE IF	RoundShape AND Age ≥ 60	THEN risk is	26.09%	46
ELSE		THEN risk is	10.38%	366

Objectives

The paper aims to achieve the following **objectives**:

Implementation

Propose an **algorithm** creating a **falling rule list** for patient diagnosis

Usage

Create a model with a **high level of interpretability** for the physicians (by looking at the list, they will understand the decision criterias).

Proposed Method

Alerts and repeats

- **Binary classification** model to estimate $p(Y|x)$, Y is the disease and x the patient features.
  - — **Y** indicates the **presence of a disease**
  - — **x** patients **features**
- Coditional distribution, **ordered list of IF THEN**. With the $p(Y = 1)$ decreasing after each rule
- **Bayesian Parametrization** to characterize the posterior falling rule list.

Background

Classic Discrete Distributions, Bayesian Inference, Point Estimators

Introduction

Background

- One of the **most challenging** parts of Bayesian parameterization is **choosing** the right **distribution** to represent the model. The commonly used distributions will be introduced.
- **Bayesian Inference** will be introduced as a **method**
- **Point estimators** as well

Common Discrete Distributions

Background

This paper uses the following distributions in its model:

- **Bernoulli:** distribution captures **binary cases**, $x \epsilon [0, 1]$: it's the coin toss distribution. It's parametrized by $p = P(X = 1) \epsilon [0, 1]$.
- **Poisson:** distribution describes a **rare event limit**: there are more and more $(n \to \infty)$ Bernoulli(p) random variables, but each has less and less of a chance of giving $1 (p = \frac{\lambda}{n} \to 0)$. It's parametrized by $\lambda > 0$.
- **Gamma:** distribution models the **waiting time until the occurrence** of k events in a Poisson process. It's parametrized by a shape parameter α (the number of events) and a rate parameter β.

Bayesian Inference
Background

Statistical method that **updates** the **probability** of a hypothesis as more **evidence or information** becomes available.

Discrete case

$$p_{\Theta|X}(\theta|x) = \frac{p_\Theta(\theta)p_{X|\Theta}(x|\theta)}{\sum_t p_\Theta(\theta)p_{X|\Theta}(x|t)}$$

- $p_\Theta(\theta)$ is the **Prior distribution**, our belief on the unknown truth Θ
- $p_{X|\Theta}(x|\theta)$is the **likelihood** representing the relation between the observation X and Θ
- $p_{\Theta|X}(\theta|x)$ is the **Posterior distribution** representing our belief in X after observing Θ

Point Estimator, MAP estimator

Background

$\hat{\theta}$ is an estimator that maps an observation x into a realistic θ, called a point estimator (used in a single observation).

Theorem

$$\hat{\theta}_{MAP} = argmax_{\theta} p_{\Theta|X}(\theta|x)$$

Training Falling Rule List

Model parameters, Likelihood, Prior,Mining Algorithm, Posterior, Summary

Plan

Material and Methods

- **Objective:** find the optimal Rule list
- We need therefore to **parameterize** the model (prior and likelihood).
  - — Enforce **monotonicity** over the **risk** score r_l associated with each IF cause
  - — Build the **prior specific**
- Find the **optimal Point Estimator**, that can build the optimal Rule list

Model Parameters

Material and Methods

	Conditions		Probability	Supp.
c_0 : IF	IrregularShape AND Age ≥ 60	THEN **r0** is	85.22%	230
c_1 : ELSE IF	SpiculatedMargin AND Age ≥ 45	THEN **r1** is	78.13%	64

- $L \in Z^+ \rightarrow$ size of the **list** (2 in this case)
- $c_l(.) \in B_x(.),\ for\ l = 0, ..., L-1 \rightarrow$ **IF** clauses
- $r_l \in R,\ for\ l = 0, ..., L \rightarrow$ **risk** score $s.t.\ r_{l+1} \leq r_l\ for\ l = 0, ..., L-1$
- r_l fed into a **logistic** function to produce **risk** probability
- $L + 1$ **nodes** and **risk** probabilities. +1 for default patients matching none of the L rules (ELSE case)

Model Parameters

Introduction

	Conditions		Probability	Supp.
IF	IrregularShape AND Age ≥ 60	THEN risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age ≥ 45	THEN risk is	78.13%	64
ELSE IF	IlDefinedMargin AND Age ≥ 60	THEN risk is	69.23%	39
ELSE IF	IrregularShape	THEN risk is	63.40%	153
ELSE IF	LobularShape AND Density ≥ 2	THEN risk is	39.68%	63
ELSE IF	RoundShape AND Age ≥ 60	THEN risk is	26.09%	46
ELSE		THEN risk is	10.38%	366

Prior - Plan

Material and Methods

- **Reparametrization** to enforce **monotonicity** on r_l
- Build the **prior specific**
- The **prior specific** is **exposed** only to the outputs of a **Mining algorithm** to help with computations.

Prior

Reparametrization

$$r_l = \log(v_l) \quad \text{for} \quad l = 0, \ldots, L$$

$$v_l = K \prod_{l'=l}^{L-1} y_{l'} \quad \text{for} \quad l = 0, \ldots, L-1$$

Constraints:

$$v_L = K$$

$$y_l \geq 1$$

$$K \geq 0$$

Therefore r_L (risk of default rule) is **equal** to $\log(K)$.
After parametrization, we obtain the following:

$$\theta = \{L, \{c_l(.)\}_{l=0}^{L-1}, \{\gamma_l\}_{l=0}^{L-1}, K\}$$

Mining Algorithm

Material and Methods

- Place **positive prior** probability of $\{c_l\}_{l=0}^{L-1}$ only over a **list of booleans B**
- **B** is a the result of a **mining algorithm** (FPGrowth is used in this case)
- **Input** is a **binary** dataset , where x is a boolean vector and the output is a set of subset of the features of the dataset

Input

Binary dataset , where x is a boolean vector

Output

Set of **subsets** of the **features** of the dataset

Prior Specific

Initialize hyperparameter

$H = \{B, \lambda, \{\alpha_l\}_{l=0}^{|B|-1}, \alpha_K, \beta_K, {w_l}_{l=0}^{|B|-1}\}$

Initialize $\Theta \leftarrow \{\}$

$L \sim Poisson(\lambda)$

$For\ l = 0, .., L-1$

$$c_l(\cdot) \sim p_{c(\cdot)}\left(\cdot \mid \Theta; B, \{w_l\}_{l=0}^{|B|-1}\right)$$

$$p_{c(\cdot)}\left(c(\cdot) = c_j(\cdot) \mid \Theta; B, \{w_l\}_{l=0}^{|B|-1}\right)$$

$\propto w_j$ if $c_j(\cdot) \notin \Theta$ and 0 otherwise.

Update $\Theta \leftarrow \Theta \cup \{c_l(\cdot)\cdot\}$

$For\ l = 0, .., L-1$ draw

$\gamma_l \sim Gamma_1(\alpha_l, \beta_l)$,

Draw $K \sim Gamma(\alpha_k, \beta_k)$

- $L \sim Poisson(\lambda)$,where λ is the **prior decision length** decided by the user.
- **l-thrule** with prob. $\propto$ to a user designed weight w_l.
- **K** models the **risk of patients** not satisfying any rules

Posterior Probability

Material and Methods

- **Objective:** Finding the decision list with the **maximum posterior probability**.

$$p_{post}\left(L, c_{0,\ldots,L-1(.)}, K, \gamma_{0,\ldots,L-1} \middle| y_{1,\ldots,N}; c_{1,\ldots,N}\right)$$

- The posterior does **not** have a **simple** solution. It can be computationally **expensive** to even calculate the posterior distribution.
- **Monte Carlo** sampling from the posterior distribution over the decision parameter:

$$\theta = \{L, \{c_l(.)\}_{l=0}^{L-1}, \{\gamma_l\}_{l=0}^{L-1}, K\}$$

Obtaining the MAP

Material and Methods

$\theta^* = \{L^*, c_{0,\ldots,L^*-1}(\cdot)^*, K^*, \gamma_{0,\ldots,L^*-1}\}$, where

$$L^*, c^*_{0,\ldots,L^*-1}(\cdot), K^*, \gamma^*_{0,\ldots,L^*-1} \in \text{argmax}_{L,c_0,\ldots,L-1}(\cdot), K, \gamma_{0,\ldots,L-1}\mathcal{L}$$

where $\mathcal{L} = log(p_{post})$. This optimization problem is equivalent to finding:

$$L^*, c_{0,\ldots,L^*-1}(\cdot)^* \quad \in \text{argmax}_{L,\{c_l(\cdot)\}_{l=0}^{L-1}} \mathcal{L}\left(L, \{c_l(\cdot)\}_{l=0}^{L-1}, K^*, \gamma^*_{0,\ldots,L-1}\right)$$

where

$$K^*, \gamma^*_{0,\ldots,L-1} \quad \in \text{argmax}_{K,\gamma_{0,\ldots,L-1}} \mathcal{L}\left(L, \{c_l(\cdot)\}_{l=0}^{L-1}, K, \gamma_{0,\ldots,L-1}\right)$$

Note that K^* and $\gamma^*_{0,\ldots,L-1}$ depend on $L, \{c_l(\cdot)\}_{l=0}^{L-1}$.

Summary

Material and Methods

- **FRL** takes the **mined rules** and attempts to **build** a sequential **list** of rules (decision list).
- Each rule is evaluated based on its **ability** to explain the **positive** and **negative** samples (i.e., X_pos and X_neg). Rules that best separate positive from negative samples are prioritized.
- **Bayesian parameterization** to characterize the **posterior** falling rule list.

Experiments

Predicting Hospital Readmission, Performance on Public Datasets

Predicting Hostpital Readmission

Material and Methods

- **Falling Rule Lists** to preliminary **readmission** data to predict whether a patient will be **readmitted** to the hospital within **30 days**.
- Pre-operative and Post-operative data for **8000 patients**.
- Other **30 features**.

Results

Falling Rule List

No parameters were **tuned** in the Falling Rule List.

The prior **condition on L**, each rule had an **equal chance** of being in the rule list.

$\lambda = 8$, **simulated annealing** search for 5000 steps.

Measured out-if-sample performance using the **AUROC from 5-fold CV**, where the **MAP decision** list was used to predict each fold test.

Method	Mean AUROC
FRL	.80 (.02)
NF_FRL	.75 (.02)
NF_GRD	.75 (.02)
RF	.79 (.03)
SVM	.62 (.06)
Logreg	.82 (.02)
Cart	.52 (.01)

Resulting FRL

	Conditions		Probability	Support
IF	BedSores AND Noshow	THEN read. risk is:	33.25%	770
ELSE IF	PoorPrognosis AND MaxCare	THEN read. risk is:	28.42%	278
ELSE IF	PoorCondition AND Noshow	THEN read. risk is:	24.63%	337
ELSE IF	BedSores	THEN read. risk is:	19.81%	308
ELSE IF	NegativeIdeation AND Noshow	THEN read. risk is:	18.21%	291
ELSE IF	MaxCare	THEN read. risk is:	13.84%	477
ELSE IF	Noshow	THEN read. risk is:	6.00%	1127
ELSE IF	MoodProblems	THEN read. risk is:	4.45%	1325
ELSE		Read. risk is:	0.88%	3031

Performance on Public dataset

Performance on several UCI datasets:
Columns of the Mamm: BI-RADS assessment, Age, Shape, Margin, Density, Severity

Method	Spam	Mamm	Breast	Cars
FRL	.91(.01)	.82(.02)	.95(.04)	.89(.08)
NF_FRL	.90(.03)	.67(.03)	.70(.11)	.60(.21)
NF_GRD	.91(.03)	.72(.04)	.82(.12)	.62(.20)
SVM	.97(.03)	.83(.01)	.99(.01)	.94(.08)
Logreg	.97(.03)	.85(.02)	.99(.01)	.92(.09)
CART	.88(.05)	.82(.02)	.93(.04)	.72(.17)
RF	.97(.03)	.83(.01)	.98(.01)	.92(.05)

Conclusion & Comments

Conclusion, Positive aspect of the paper, negative aspect of the paper

Conclusion

- **New class** of **interpretive** predictive model.
- No loss in accuracy for using the FRL.
- ”An interpretable model that is actually **used is bette**r that one that is more accurate that sits **on a shelf**”. Director U.S. National Institute of Justice.

Comments

Positive Aspects

Novel model explained in detailed.

One of the **highest interpretable** models existing in the domain.

Tested on different **datasets** and with **different setups**.

8000 patients for hospitalization dataset is and **consequent amount** of samples.

Negative Aspects

Comparison should include **boosting models** like XGBoost or L-GBM.

SVM performed terribly, when it is a widely used model in this context.

30 features for a hospitalization dataset is **very small**

Thank you!

Annex : PMF gamma Distribution

Theorem

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(k)} \, for \; x > 0.$$

Annex : Mining Algorithm

X_{pos}, X_{neg}=mine_antecedents(data): mines rules from the training data using the FP-Growth algorithm, separately for positive and negative samples, and then forms binary representations of the data points that satisfy each rule, returning the sets of positive and negative examples, rule lengths, and the list of mined antecedents. LINK

Annex : AUROC metric

Area Under the Receiver Operating Characteristic curve measures the out-of-sample performance of a binary classifier by evaluating its ability to distinguish between classes, with a score of 1 indicating perfect classification and 0.5 representing random guessing.