Mondrian Predictive Systems for Censored Data

Henrik Boström[1] Henrik Linusson[1,2] Anders Vesterberg[3]

[1]KTH Royal Institute of Technology, Sweden

[2]Ekkono Solutions AB, Sweden

[3]Scania CV AB, Sweden

September 14, 2023

Outline

Conformal predictive systems

Time-to-event prediction

Handling censored data

Empirical investigation

Concluding remarks

Conformal predictive systems

Conformal predictive systems transform point predictions into cumulative distribution functions (conformal predictive distributions), which provide p-values for given target values and target values for given p-values.

Conformal predictive distributions come with a *validity* guarantee; the p-value for the true target is distributed uniformly on $[0, 1]$.

Conformal predictive systems

A *split conformal predictive system* can be constructed as follows:

1. randomly divide the training data into two disjoint subsets; the proper training set and the calibration set
2. train the underlying model h using the proper training set
3. calculate scores $\alpha_1, \ldots, \alpha_q$ for the calibration set, where

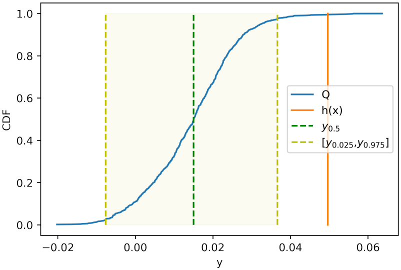
$$\alpha_i = \frac{y_i - h(\boldsymbol{x}_i)}{\sigma_i}$$

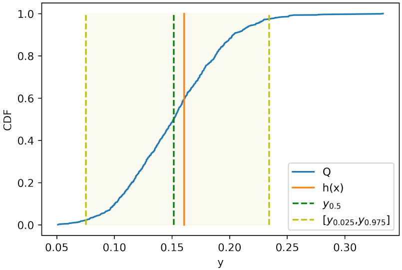
4. let $\alpha_{(1)}, \ldots, \alpha_{(q)}$ be the scores sorted in ascending order
5. for a test object $\boldsymbol{x}$ with difficulty σ:

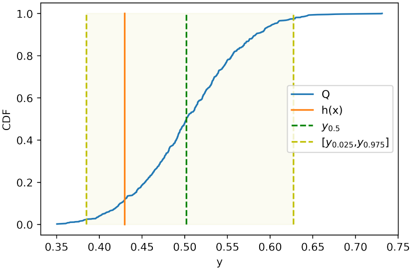
   let $C_{(i)} = h(\boldsymbol{x}) + \alpha_{(i)}\sigma$ for $i \in \{1, ..., q\}$
   let $C_{(0)} = -\infty$ and $C_{(q+1)} = \infty$
   output the conformal predictive distribution:

$$Q(y) = \begin{cases} \frac{n+\tau}{q+1} & \text{if } y \in \left(C_{(n)}, C_{(n+1)}\right) \text{ for } n \in \{0, ..., q\} \\ \frac{n'-1+(n''-n'+2)\tau}{q+1} & \text{if } y = C_{(n)} \text{ for } n \in \{1, ..., q\} \end{cases}$$

Mondrian predictive systems







Time-to-event prediction

- Well-calibrated probabilities are needed for effective decision-making, e.g., when balancing the maintenance cost against the cost of failure on the road.
- Events may not (yet) have been observed for all of the examples, which means that the time-to-event is not always known, e.g., breakdown may have occurred only for some but not all vehicles in a fleet.
- Removing examples for which the event has not occurred or adjusting such examples by imputing target values are not guaranteed to maintain validity of the conformal predictive systems.

Handling censored data

Given

- a set of objects $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q\}$
- a set of values $Y = \{y_1, \ldots, y_q\}$
- a set of binary (event) indicators $E = \{e_1, \ldots, e_q\}$, where $e_i = 1$ if and only if y_i is not censored[1]
- a test object x,

generate a conformal predictive distribution Q, such that $Q(y) \sim U(0, 1)$, where y is the true target for x

[1] the censoring value for the object $\boldsymbol{x}_i$ is greater than the true target

Idea: Use the Kaplan-Meier Estimator (KME)

Assuming that the censoring values are independent of the true targets:

$$\hat{S}(t) = \prod_{k=0}^{t} \left(1 - \frac{|\{y_j \in Y : e_j = 1 \wedge y_j = y_k\}|}{|\{y_j \in Y : y_j \geq y_k\}|}\right)$$

The above assumption does however not imply that non-conformity scores computed from censoring values are independent of non-conformity scores computed from the true targets.

To keep the independence, we in addition assume that:

$$h(\boldsymbol{x}_i) = 0$$

$$\sigma_i = 1$$

Hence, the non-conformity scores become:

$$\alpha_i = y_i$$

Mondrian predictive systems for censored data

- Split the calibration set into k Mondrian categories, e.g., by binning the predicted values of the underlying model
- Form a CDF using the KME for each Mondrian category
- For a test object, find the category and return the corresponding CDF

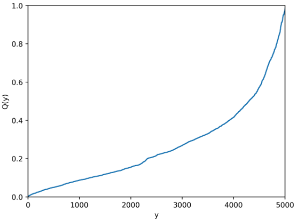
Experimental setup

- Dataset: serum free light chain (FLC)
  - 7874 instances
  - 8 features
- Target: time to death
  - synthetic censoring (50%)
  - actual censoring (72%)
- 75% used for training
  25% for testing

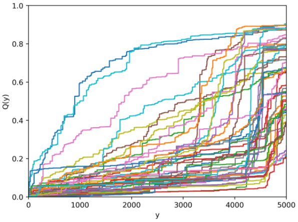
Employed methods

- The Kaplan-Meier estimator (KME); no hyperparameters
- Random Survival Forests (RSF); default settings for all hyperparameters except for `n_estimators = 500`
- Censored Quantile Regression Forests (QRF); default settings for the individual regression trees, `n_estimators = 500` and `k = 200` (no. of neighbors)
- Conformal Predictive Systems (CPS); using a `RandomForestRegressor` generated with default settings for all hyperparameters except for `n_estimators = 500`, and half of the training set for calibration, 25 bins

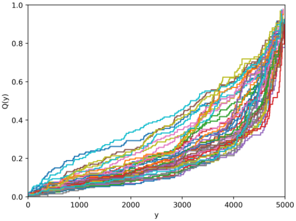
CDFs for 50 random test objects



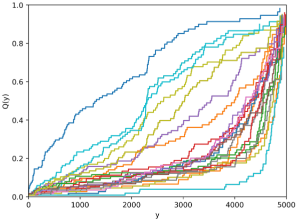
a) Kaplan-Meier Estimator



b) Random Survival Forest

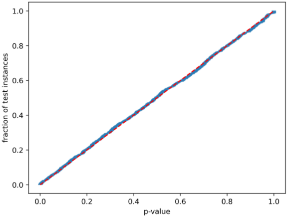


c) Quantile Regression Forest

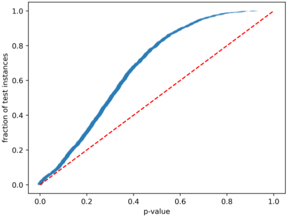


d) Conformal Predictive System

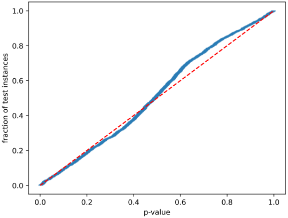
Distribution of p-values for true targets



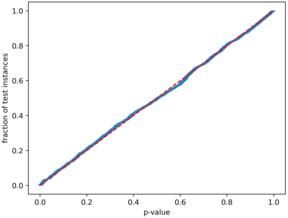
a) Kaplan-Meier Estimator



b) Random Survival Forest



c) Quantile Regression Forest



d) Conformal Predictive System

Runtimes and Kolmogorov-Smirnov test

	Training time	Testing time	KS-test
KME	0.00	0.02	7.70e-01
RSF	43.37	4.49	6.95e-174
QRF	1.39	0.82	2.33e-09
CPS	0.56	0.07	3.06e-01

Error rates and time points

	Error	Mean time-point	Median time-point
KME	0.0153	48.0	48
RSF	0.0390	402.0	186
QRF	0.0039	12.3	5
CPS	0.0094	78.6	14

a) 99% confidence

	Error	Mean time-point	Median time-point
KME	0.0523	487.0	487
RSF	0.0676	1344.6	1110
QRF	0.0444	510.5	461
CPS	0.0449	710.9	532

b) 95% confidence

	Error	Mean time-point	Median time-point
KME	0.1027	1195.0	1195
RSF	0.1362	2318.3	2165
QRF	0.0893	1352.8	1370
CPS	0.0977	1621.7	1764

c) 90% confidence

Non-synthetic censoring

	C-index	Mean TTE	Median TTE	Training time	Testing time
KME	0.5000	4998.0	4998	0.00	0.00
RSF	0.7234	4554.6	4998	36.58	2.83
QRF	0.5804	4655.4	4715	1.38	0.56
CPS	0.6379	4241.7	4486	0.58	0.05

Concluding remarks

- An approach to handling censored data using Mondrian predictive systems has been proposed
- The approach has been shown empirically to be valid, in contrast to random survival forests and censored quantile regression forests
- The approach is currently constrained to generate one CDF for each Mondrian category; future work includes relaxing this constraint