Distribution Choice in Non-life Insurance Risk Model with Statistical Learning Method

XIAOSHAN SU

The AFIR-ERM Colloquium - Florence 2019

emlyon business school

Outline

Motivation

Traditional methods: Log-likelihood, AIC, BIC,...

Disadvantages:

- simple decision rule
- need fit each distribution to each dataset
- no feature information to explain why one distribution is preferred over another one

Difficulty: computation burden for the assignment with lots of alternative distributions and datasets

Motivation

In non-life insurance, the claim number is usually modeled with the following distributions:

- Poisson $N \sim Possion(\lambda)$ $E(N) = \lambda, Var(N) = \lambda, E(N) = Var(N)$
- Binomial $N \sim B(n, p)$ $E(N) = np, Var(N) = np(1 - p), E(N) > Var(N)$
- Negative binomial $N \sim NB(\mu, \phi)$ $E(N) = \mu, Var(N) = \mu + \frac{\mu^2}{\phi}, E(N) < Var(N)$

Then, we have a simple rule of selecting distribution:

$$\begin{cases} \text{Poisson} & \text{VMR}=1 \\ \text{Binomial} & \text{VMR} < 1 \\ \text{Negative binomial} & \text{VMR} > 1 \end{cases}$$

Motivation

Questions:

- Can we find such kind of rule for any distribution selection problem?
- The simple rule only makes use of first and second order moments information. Can we build rules that exploit more distribution information?

Solution

Solution: treat a distribution selection problem as a classification problem

- Machine learning classifiers are employed to identify and build rules.
- Features: descriptive statistic variables
- Data:
  a) simulation of distribution parameters
  b) simulate data from each distribution with simulated parameters
  c) each record is built with distribution label and summary statistics for each simulated dataset

Method Specification

Denote by $\{x_1, ..., x_p\}$ the descriptive statistic variables and $\{D_1, ..., D_m\}$ the set of m alternative distributions from which we choose most appropriate distribution for real data. Machine learning classifier can be regarded as a mapping $T : (x_1, ..., x_p) \rightarrow y, y \in \{D_1, ..., D_m\}$.

- Training sample:
  - simulate parameters: generate N pieces of parameter set for each distribution D_i
  - simulate data: generate a sample of size n for each distribution D_i with each piece of parameter set
  - records: compute $\{x_1, ..., x_p\}$ for each sample and generate mN records
- Train classifier using training sample: decision tree, k-nearest neighbour classifier, neural network, support vector classifier, bagging, boosting and random forest, etc
- $k-$fold cross-validation method to compare the performance of classifiers and select the best one to predict the most appropriate distribution for real data

Non-life Insurance Application

In collective risk model, the aggregate loss S is written as follows:

$$S = \sum_{i=1}^{N} X_i,$$

where N denotes claim frequency, X_i denotes claim severity and $N, X_1, X_2, \ldots$ are independent.

Claim frequency:

- Poisson $N \sim \text{Possion}(\lambda), \lambda > 0$
- Binomial $N \sim \text{B}(n, p), n \in \mathbb{N}, 0 \leqslant p \leqslant 1$
- Negative binomial $N \sim \text{NB}(\mu, \phi), \mu > 0, \phi > 0$

Claim severity:

- Exponential distribution $X_i \sim \text{Exp}(\lambda), \lambda > 0$
- Gamma distribution $X_i \sim \Gamma(\alpha, \beta), \alpha, \beta > 0$
- Pareto distribution $X_i \sim \text{Pa}(\alpha, \theta), \alpha, \theta > 0$
- ...

Distribution Selection for Claim Frequency

- We generate 2000 pieces of parameter sets for each distribution using the following Uniform distributions:

$$\begin{cases} \mu \sim \text{Uniform}(0.1, 20), & \text{for Poisson, Binomial and Negative binomial} \\ \phi \sim \text{Uniform}(0.1, 10), & \text{for Negative binomial} \\ p \sim \text{Uniform}(0.01, 1) & \text{for Binomial.} \end{cases}$$

- Each parameter set produces 10000 sample points for each distribution.
- We compute the VMR, mean, variance, standard deviation, skewness, kurtosis, percentage of zero, coefficient-of-variation, quantiles from 10% to 90% with 10% increments, quantiles 95%, 99%, the 10%, 20%, 30%, 40% inter-quantiles, the range for each sample and obtain a training sample of size 30000 with 24 input features and a output variable distribution label.
- We train decision tree with this sample.

Distribution Selection for Claim Frequency

- The decision tree identifies the most important feature VMR (simple rule):

$$\begin{cases} \text{Poisson} & 0.97 \leqslant \text{VMR} < 1 \\ \text{Binomial} & \text{VMR} < 0.97 \\ \text{Negative binomial} & \text{VMR} \geqslant 1 \end{cases}$$

- The decision tree can be employed to find rules in cases without simple rules.

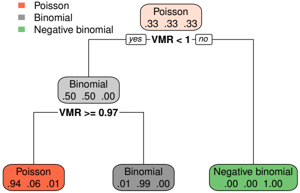


Figure: The decision tree

Distribution Selection for Claim Frequency

The training accuracy of decision tree reaches 97.32%.

Table: The confusion matrix

Actual	Predicted		
	Poisson	Binomial	NB
Poisson	1974	24	2
Binomial	123	1877	0
NB	12	0	1988

Distribution Selection for Claim Frequency

In real application, we only seek accurate classifier to help make the right selection, without need of knowing the rules. Thus, we can choose some other better classifiers. Then, 5-fold cross-validation method is used to compare seven classifiers.

Model	Training Error	Testing Error
classification tree	2.65%	2.83%
k-nearest neighbour	2.25%	5.58%
neural network	0.33%	4.58%
support vector classifier	18.7%	18.6%
bagging	2.52%	2.87%
boosting	2.68%	2.8%
random forest	0.13%	2.4%

Distribution Selection for Claim Frequency

The feature importance from random forest:

Feature	Mean-Decrease OOB Error rate	Feature	Mean-Decrease Gini
variance to mean ratio	47.19	variance to mean ratio	1344.67
standard deviation	29.06	skewness	594.23
variance	28.86	kurtosis	457.37
CV	26.27	CV	253.37
skewness	24.52	variance	253.21
mean	22.18	standard deviation	245.20
range	20.77	range	177.43
kurtosis	20.41	10% Inter-Quantile	112.93
Quantile 99%	16.25	20% Inter-Quantile	83.12
10% Inter-Quantile	15.89	zero percentage	74.61
zero percentage	15.54	30% Inter-Quantile	49.29
Quantile 95%	13.89	mean	48.73
Quantile 90%	13.18	Quantile 99	48.04
20% Inter-Quantile	12.62	Quantile 10	38.47
Quantile 20%	11.92	Quantile 20	29.99
Quantile 80%	11.42	40% Inter-Quantile	29.19
Quantile 10%	11.06	Quantile 95	25.75
Quantile 40%	10.98	Quantile 30	21.02
30% Inter-Quantile	10.48	Quantile 40	20.91
Quantile 30%	10.46	Quantile 90	19.21
Quantile 50%	9.02	Quantile 60	16.05
Quantile 60%	8.76	Quantile 50	14.35
Quantile 70%	8.64	Quantile 70	13.70
40% Inter-Quantile	5.98	Quantile 80	12.08

Distribution Selection for Claim Severity

We illustrate distribution choice for claim severity among exponential, Gamma and Pareto distributions. The 2000 pieces of sample, each with 10000 sample points, are generated from the simulation of each distribution. The parameters set of each sample is simulated from the following uniform distribution:

$$\begin{cases} \mu \sim \text{Uniform}(0.1, 10), & \text{for Exponential, Gamma and Pareto} \\ \alpha \sim \text{Uniform}(1, 3), & \text{for Gamma} \\ \alpha_2 \sim \text{Uniform}(1, 3), & \text{for Pareto.} \end{cases}$$

Distribution Selection for Claim Severity

We choose 27 descriptive statistic variables as input features, including mean, variance, standard deviation, coefficient of variation (CV), variance to mean ratio, skewness, kurtosis, quantiles from 10% to 90% with 10% increments, quantiles 95%, 99%, the 10%, 20%, 30%, 40% inter-quantiles, the range, $\frac{L(VaR_{0.99}^2)}{L(VaR_{0.99})^2}$, $\frac{E(X-VaR_{0.99}|X>VaR_{0.99})}{E(X-VaR_{0.95}|X>VaR_{0.95})}$, $E(X-VaR_{0.99}|X>VaR_{0.99})$, and $\frac{VaR_{0.99}}{VaR_{0.95}}$.

We observe the feature CV:

- Exponential Case: $X \sim Exp(\lambda), \lambda > 0$, $\mathrm{CV} = 1$
- Gamma Case: $X \sim \Gamma(\alpha, \beta), \alpha, \beta > 0$, $\mathrm{CV} = \frac{1}{\sqrt{\alpha}}$
- Pareto Case: $X \sim Pa(\alpha, \beta), \alpha, \beta > 0$, $\mathrm{CV} = \frac{1}{\sqrt{\alpha(\alpha-2)}}$

Distribution Selection for Claim Severity

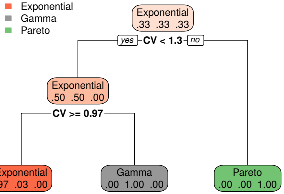


Figure: The decision tree

We find the following simple rule:

$$\begin{cases} \text{Exponential} & 0.97 \leqslant \text{CV} < 1.3 \\ \text{Gamma} & \text{CV} < 0.97 \\ \text{Pareto} & \text{CV} \geqslant 1.3 \end{cases}$$

Distribution Selection for Claim Severity

The training error is only 0.0112.

Table: The confusion matrix

Actual	Predicted		
	Exponential	Gamma	Pareto
Exponential	1996	4	0
Gamma	63	1937	0
Pareto	0	0	2000

Distribution Selection for Claim Severity

We use the 5-fold cross-validation method to compare the following seven classifiers.

Model	Training Error	Testing Error
decision tree	1.11%	1.13%
k-nearest neighbour	1.23%	1.9%
neural network	0.15%	1.98%
support vector classifier	2.97%	4.9%
bagging	1.11%	1.13%
boosting	0	1.22%
random forest	0	1.2%

Distribution Selection for Claim Severity

The feature importance from random forest:

Feature	Mean-Decrease OOB Error rate	Mean-Decrease Gini	
coefficient of variation	35.81	coefficient of variation	942.02
Quantile 10%	27.84	skewness	698.19
Quantile 20%	22.65	kurtosis	565.02
Quantile 30%	19.77	$\frac{E(X - VaR_{0.99} \mid X > VaR_{0.99})}{E(X - VaR_{0.95} \mid X > VaR_{0.95})}$	515.18
variance to mean ratio	18.76	$\frac{VaR_{0.99}}{VaR_{0.95}}$	490.04
Quantile 40%	16.64	range	148.68
Quantile 99%	16.31	variance to mean ratio	136.32
skewness	15.94	Quantile 10%	125.18
20% Inter-Quantile	15.88	$E(X - VaR_{0.99} \mid X > VaR_{0.99})$	78.01
Quantile 95%	15.38	Quantile 20%	65.90
standard deviation	15.24	variance	47.53
variance	14.90	standard deviation	40.22
10% Inter-Quantile	14.61	Quantile 30%	31.98
Quantile 70%	14.38	Quantile 40%	26.01
30% Inter-Quantile	14.36	Quantile 99%	15.09
Quantile 50%	14.12	Quantile 50%	10.29
40% Inter-Quantile	13.98	Quantile 60%	8.65
Quantile 90%	13.93	Quantile 70%	7.18
mean	13.93	Quantile 95%	6.91
$E(X - VaR_{0.99} \mid X > VaR_{0.99})$	13.47	10% Inter-Quantile	6.53
Quantile 60%	13.25	20% Inter-Quantile	5.98
Quantile 80%	12.78	30% Inter-Quantile	5.96
$\frac{E(X - VaR_{0.99} \mid X > VaR_{0.99})}{E(X - VaR_{0.95} \mid X > VaR_{0.95})}$	12.45	Quantile 90%	5.80
kurtosis	12.43	mean	5.70
$\frac{VaR_{0.99}}{VaR_{0.95}}$	12.25	Quantile 80%	5.67
range	10.25	40% Inter-Quantile	5.18
$\frac{L(VaR^2_{0.99})}{L(VaR_{0.99})^2}$	0	$\frac{L(VaR^2_{0.99})}{L(VaR_{0.99})^2}$	0.04

Practical Application

The dataset comprises six claims, including building and contents (BC) claims, inland marine (IM) claims, comprehensive new vehicles (PN) claims, comprehensive old vehicles (PO) claims, new vehicle collision (CN) claims, old vehicle collision (CO) claim. For each claim, both of claim frequency and severity are contained in the dataset.

	BC_Freq	BC_Sev	IM_Freq	IM_Sev	PN_Freq	PN_Sev	PO_Freq	PO_Sev	CN_Freq	CN_Sev	CO_Freq	CO_Sev
[1,]	0	0.00	0	0.00	0	0.000	0	0	0	0	2	3137.530
[2,]	0	0.00	0	0.00	2	5217.005	0	0	0	0	0	0.000
[3,]	0	0.00	0	0.00	0	0.000	0	0	0	0	4	23490.778
[4,]	0	0.00	0	0.00	0	0.000	0	0	0	0	2	2739.195
[5,]	1	6838.87	0	0.00	0	0.000	0	0	0	0	0	0.000
[6,]	0	0.00	2	10794.97	0	0.000	0	0	0	0	3	1402.213

Practical Application

We observe distribution choice of claim frequency for six claims.

	All Classifiers	Log-Likelihood	AIC	BIC
BC	NB	NB	NB	NB
IM	NB	NB	NB	NB
PN	NB	NB	NB	NB
PO	NB	NB	NB	NB
CN	NB	NB	NB	NB
CO	NB	NB	NB	NB

Practical Application

We observe distribution choice of claim severity among exponential, Gamma, Pareto distributions.

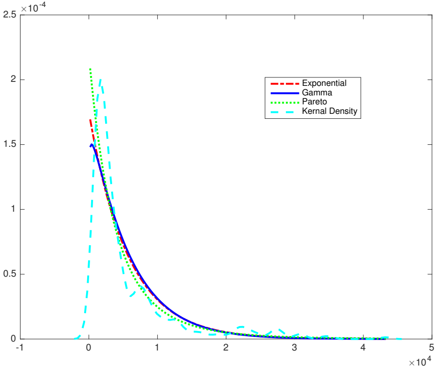
	Log-Likelihood	AIC	BIC	DT	KNN	NN
BC	Pareto	Pareto	Pareto	Pareto	Pareto	Gamma
IM	Pareto	Pareto	Pareto	Pareto	Exponential	Gamma
PN	Pareto	Pareto	Pareto	Pareto	Gamma	Gamma
PO	Pareto	Pareto	Pareto	Pareto	Pareto	Pareto
CN	Pareto	Pareto	Pareto	Exponential	Exponential	Gamma
CO	Pareto	Pareto	Pareto	Pareto	Exponential	Gamma

SVM	Bagging	Boosting	RF
Pareto	Pareto	Pareto	Pareto
Pareto	Pareto	Pareto	Pareto
Pareto	Pareto	Pareto	Pareto
Pareto	Pareto	Pareto	Pareto
Pareto	Exponential	Exponential	Exponential
Pareto	Pareto	Pareto	Pareto

The decision tree, bagging, boosting and random forest agree on the same results. Perhaps it is caused by choosing decision tree as weak learners in bagging, boosting and random forest. We recommend using decision tree, bagging, boosting and random forest for distribution selection.

Practical Application

	Log-Likelihood	AIC	BIC
Exponential	3589.9	7181.7	7185.6
Gamma	3589.3	7182.6	7190.4
Pareto	3582.2	7168.4	7176.2



Thank you for your attention.