MONASH University

MONASH BUSINESS SCHOOL

Feature-based Model Selection for Time Series Forecasting

Thiyanga Talagala

Rob J Hyndman

George Athanasopoulos

Large collections of time series

Freelance **Forecasting Multiple time series** 8 Mar 2016

Logistics Capital & Strategy – Posted by LogCapStrat – Anywhere

Logistics CAPITAL & STRATEGY

Job Description

Logistics Capital & Strategy is looking for a Data Scientist with expertise in Parallel computing to assist in code optimization and parallel processing of an under development forecasting model in R.

This is a contract position and we are expecting the project to completed over a period of 2 weeks.

Skills Required:
R Programming/Python/Scala (for code development)
MSSQL for data extraction into programming environment
Apache Spark or related big data processing frameworks to allow for high speed data processing

Project Scope:
The current forecasting model build on R needs to be scaled, and optimized to allow forecasting of millions of individual time series, ideally in a span of few hours.

Related

Statistician & R programmer March 16, 2017 Similar post	Quantitative Research Associate March 6, 2017 Similar post	R Shiny Developer March 14, 2017 Similar post

How to Apply

Large collections of time series

Freelance **Forecasting Multiple time series** 8 Mar 2016

Logistics Capital & Strategy – Posted by LogCapStrat – Anywhere

Logistics CAPITAL & STRATEGY

Job Description

Logistics Capital & Strategy is looking for a Data Scientist with expertise in Parallel computing to assist in code optimization and parallel processing of an under development forecasting model in R.

This is a contract position and we are expecting the project to completed over a period of 2 weeks.

Skills Required:
R Programming/Python/Scala (for code development)
MSSQL for data extraction into programming environment
Apache Spark or related big data processing frameworks to allow for high speed data processing

Project Scope:
The current forecasting model build on R needs to be scaled, and optimized to allow forecasting of millions of individual time series, ideally in a span of few hours.

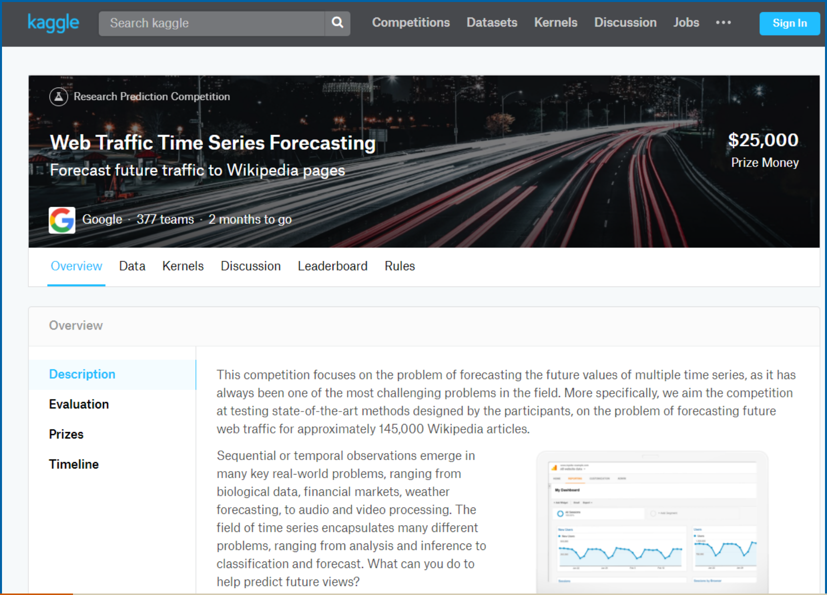
Related

forecasting of millions of individual time series

Statistician & R programmer	Quantitative Research Associate	R Shiny Developer
March 16, 2017	March 6, 2017	March 14, 2017
Similar post	Similar post	Similar post

How to Apply

Large collections of time series



Large collections of time series

Forecasting multiple time series

- Aggregate selection rule

Forecasting multiple time series

- Aggregate selection rule
    - Develop a single method which provides better forecasts across all time series.

Forecasting multiple time series

- Aggregate selection rule
  - Develop a single method which provides better forecasts across all time series.
  - No free lunch!

Forecasting multiple time series

- Aggregate selection rule
  - Develop a single method which provides better forecasts across all time series.
  - No free lunch!
- Individual model building or combined forecasts

Automatic time series forecasting

- ets algorithm
- auto.arima algorithm

ets() and auto.arima() in R

ets algorithm

- Apply each 15 ETS models that are appropriate to the data

auto.arima algorithm

- Use stepwise search to traverse model space, starting with a simple model

- For each model, optimize parameters using MLE
- Select best method using AICc

ets() and auto.arima() in R

ets algorithm

- Apply each 15 ETS models that are appropriate to the data

auto.arima algorithm

- Use stepwise search to traverse model space, starting with a simple model

- For each model, optimize parameters using MLE
- Select best method using AICc

Motivation

Reid(1972) pointed out that the performance of various forecasting methods changes according to the nature of data and if the reasons for these variations are explored they may be useful in selecting the most appropriate model.

Objective

Develop a framework that automates the selection of the most appropriate forecasting model for a given time series by using a large array of features computed from the time series.

Time series features

Cognostics: Computer-aided diagnostics
(John W. Tukey, 1985)

- Characteristics of time series

Time series features

Cognostics: Computer-aided diagnostics
(John W. Tukey, 1985)

- Characteristics of time series
- Depending on the research goals and domains, a variety of features have been introduced

Time series features

Cognostics: Computer-aided diagnostics
(John W. Tukey, 1985)

- Characteristics of time series
- Depending on the research goals and domains, a variety of features have been introduced
- Examples for time series features

Time series features

Cognostics: Computer-aided diagnostics
(John W. Tukey, 1985)

- Characteristics of time series
- Depending on the research goals and domains, a variety of features have been introduced
- Examples for time series features
  - strength of trend

Time series features

Cognostics: Computer-aided diagnostics
(John W. Tukey, 1985)

- Characteristics of time series
- Depending on the research goals and domains, a variety of features have been introduced
- Examples for time series features
  - strength of trend
  - strength of seasonality

Time series features

Cognostics: Computer-aided diagnostics
(John W. Tukey, 1985)

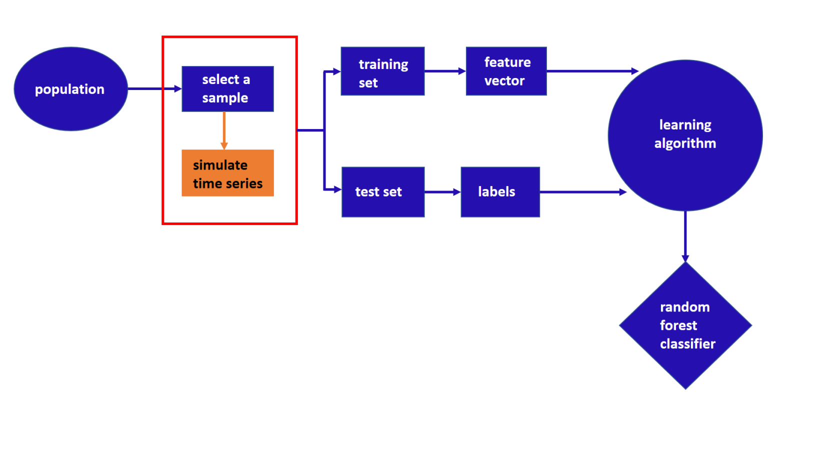
- Characteristics of time series
- Depending on the research goals and domains, a variety of features have been introduced
- Examples for time series features
  - strength of trend
  - strength of seasonality
  - lag correlation

Time series features

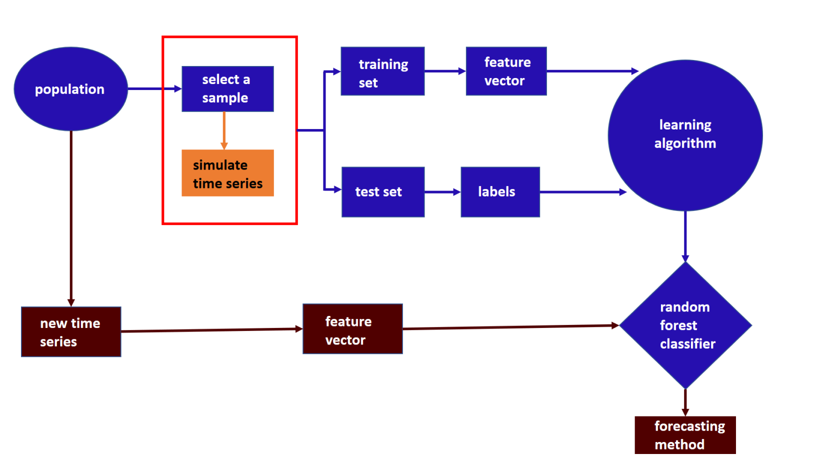
Cognostics: Computer-aided diagnostics
(John W. Tukey, 1985)

- Characteristics of time series
- Depending on the research goals and domains, a variety of features have been introduced
- Examples for time series features
    - strength of trend
    - strength of seasonality
    - lag correlation
    - spectral entropy

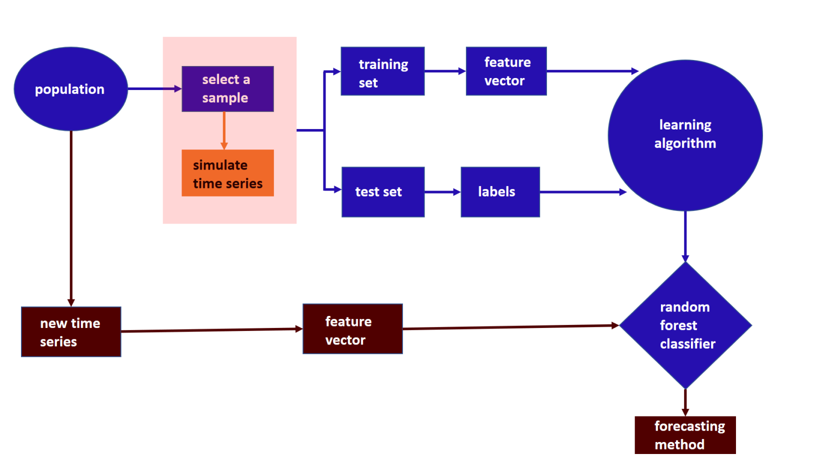
Methodology: “offline” part of the algorithm



Methodology: “online” part of the algorithm

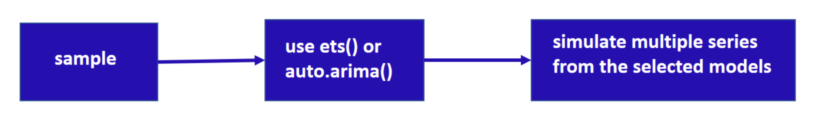


Methodology: reference set

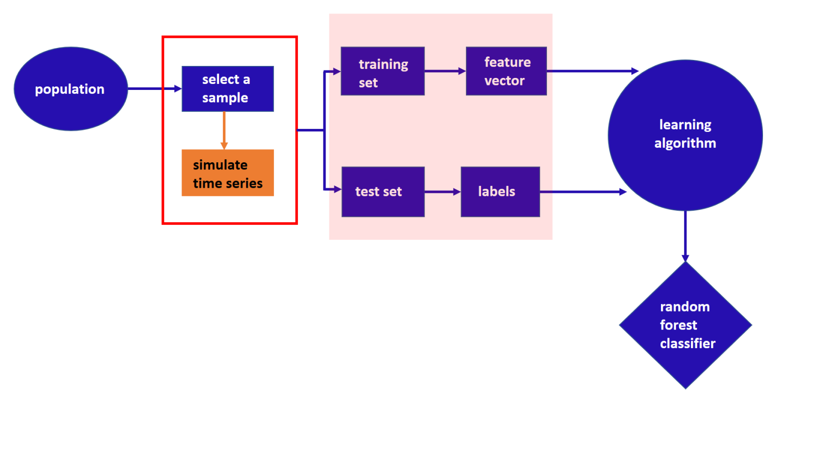


Augmenting the reference set with simulated series

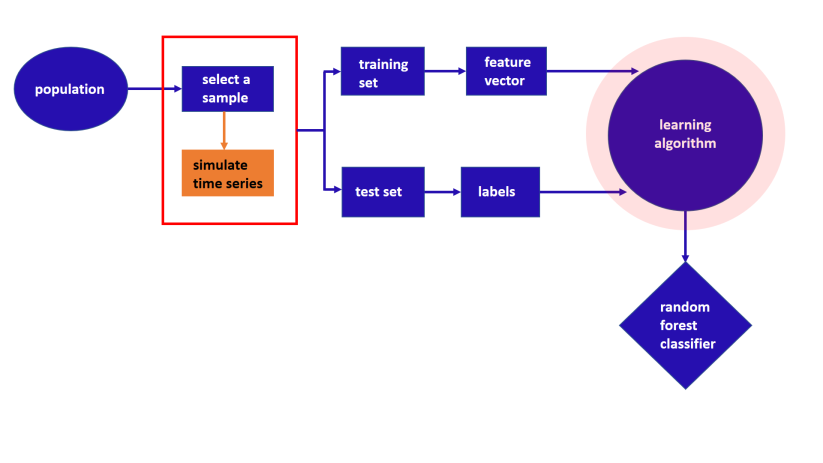
- when our sample is too small to build a reliable classifier
- when we wish to add more of some types of time series to the training set in order to get a more balanced sample
- How?



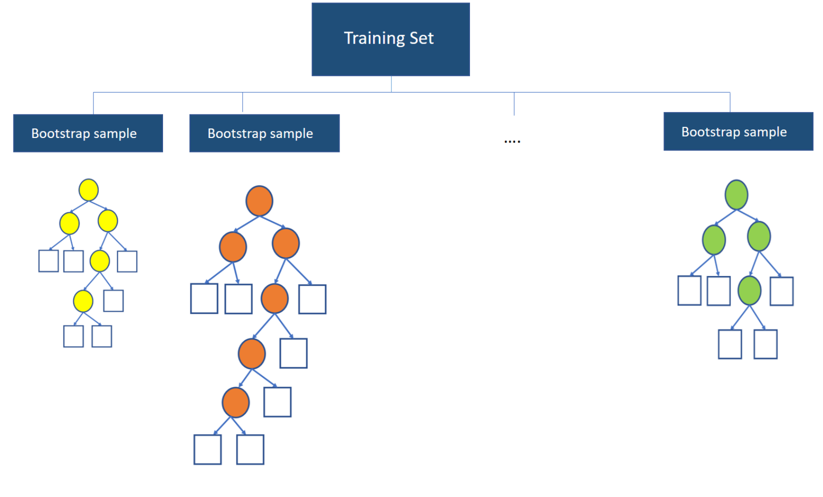
Methodology: features and class labels



Methodology: random forest



Random forest



The random forest algorithm for classification

- Let N be the number of trees to build.

The random forest algorithm for classification

- Let N be the number of trees to build.
- At each iteration,

The random forest algorithm for classification

- Let N be the number of trees to build.
- At each iteration,
  - Select a new bootstrap sample from the training set.

The random forest algorithm for classification

- Let N be the number of trees to build.
- At each iteration,
    - Select a new bootstrap sample from the training set.
    - Grow a random-forest tree to the bootstrapped data.

The random forest algorithm for classification

- Let N be the number of trees to build.
- At each iteration,
    - Select a new bootstrap sample from the training set.
    - Grow a random-forest tree to the bootstrapped data.
    - At each node, select **m** variables at random from the **p** variables.

The random forest algorithm for classification

- Let N be the number of trees to build.
- At each iteration,
  - Select a new bootstrap sample from the training set.
  - Grow a random-forest tree to the bootstrapped data.
  - At each node, select **m** variables at random from the **p** variables.
  - Select the best split-point among the **m**.

The random forest algorithm for classification

- Let N be the number of trees to build.
- At each iteration,
  - Select a new bootstrap sample from the training set.
  - Grow a random-forest tree to the bootstrapped data.
  - At each node, select **m** variables at random from the **p** variables.
  - Select the best split-point among the **m**.
- Overall prediction: Majority vote from all individually built trees.

Preliminary study

- We consider non-seasonal time series

Preliminary study

- We consider non-seasonal time series
- Data: Yearly time series of M1 and M3 competitions

Preliminary study

- We consider non-seasonal time series
- Data: Yearly time series of M1 and M3 competitions
    - Classification algorithm - yearly series of M3 competition

Preliminary study

- We consider non-seasonal time series
- Data: Yearly time series of M1 and M3 competitions
  - Classification algorithm - yearly series of M3 competition
  - Evaluation - yearly series of M1 competition

Preliminary study

- We consider non-seasonal time series
- Data: Yearly time series of M1 and M3 competitions
  - Classification algorithm - yearly series of M3 competition
  - Evaluation - yearly series of M1 competition
- Class labels

Preliminary study

- We consider non-seasonal time series
- Data: Yearly time series of M1 and M3 competitions
  - Classification algorithm - yearly series of M3 competition
  - Evaluation - yearly series of M1 competition
- Class labels
  - We consider random walks, white noise, ARIMA processes and ETS processes

Preliminary study

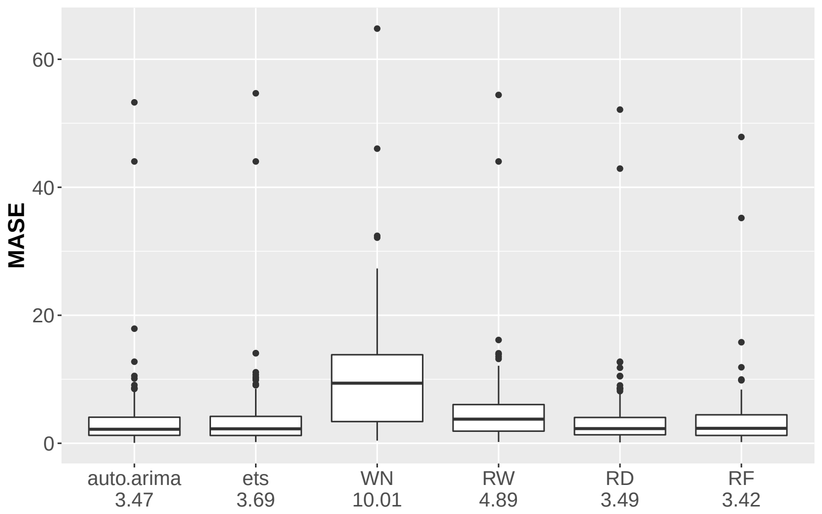
- We consider non-seasonal time series
- Data: Yearly time series of M1 and M3 competitions
  - Classification algorithm - yearly series of M3 competition
  - Evaluation - yearly series of M1 competition
- Class labels
  - We consider random walks, white noise, ARIMA processes and ETS processes
  - The model with the smallest MASE

Time series feature

- Strength of trend
- Spectral entropy
- Hurst exponent
- Lyapunov exponent
- Parameter estimates of Holt linear trend model

- Length
- Coefficient of determination of the linear trend model
- ACF and PACF based features - calculated on both the raw and differenced series

Results: Distribution of MASE



What next?

- Develop a more comprehensive set of features that are useful in identifying different data generating processes.

What next?

- Develop a more comprehensive set of features that are useful in identifying different data generating processes.
- Extend the time series collection to non-seasonal data.

What next?

- Develop a more comprehensive set of features that are useful in identifying different data generating processes.
- Extend the time series collection to non-seasonal data.
- Test for several large scale real time series data sets.

What next?

- Develop a more comprehensive set of features that are useful in identifying different data generating processes.
- Extend the time series collection to non-seasonal data.
- Test for several large scale real time series data sets.
- Consider other classification methods.

Acknowledgement

The Victorian Branch of the Statistical Society of Australia Inc. (SSA Vic)

Slides shared online at:

https://github.com/thiyangt/YSC-2017

thiyanga.talagala@monash.edu