



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

Logistics Capital & Strategy – Posted by LogCapStrat –  Anywhere

2016



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

Logistics Capital & Strategy – Posted by LogCapStrat –  Anywhere

2016



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

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



The image shows the top section of the Kaggle website. At the top is a dark navigation bar with the Kaggle logo, a search bar, and links for Competitions, Datasets, Kernels, Discussion, Jobs, and a Sign In button. Below this is a large banner for the 'Research Prediction Competition' titled 'Web Traffic Time Series Forecasting'. The banner features a background image of a highway at night with light trails. Text on the banner includes 'Forecast future traffic to Wikipedia pages', '\$25,000 Prize Money', and 'Google · 377 teams · 2 months to go'. Below the banner are tabs for Overview, Data, Kernels, Discussion, Leaderboard, and Rules.

kaggle Search kaggle Q Competitions Datasets Kernels Discussion Jobs ... Sign In

Research Prediction Competition

Web Traffic Time Series Forecasting

Forecast future traffic to Wikipedia pages

\$25,000 Prize Money

Google · 377 teams · 2 months to go

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#)

Overview

Description

Evaluation

Prizes

Timeline

This competition focuses on the problem of forecasting the future values of multiple time series, as it has always been one of the most challenging problems in the field. More specifically, we aim the competition at testing state-of-the-art methods designed by the participants, on the problem of forecasting future web traffic for approximately 145,000 Wikipedia articles.

Sequential or temporal observations emerge in many key real-world problems, ranging from biological data, financial markets, weather forecasting, to audio and video processing. The field of time series encapsulates many different problems, ranging from analysis and inference to classification and forecast. What can you do to help predict future views?

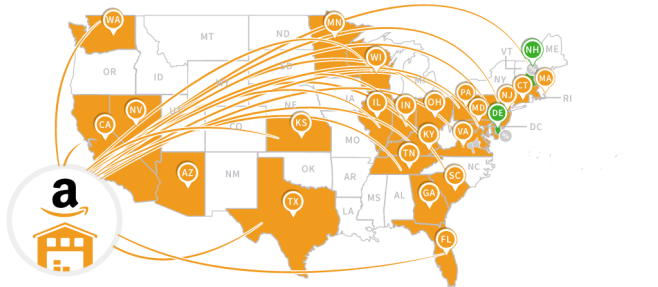


Large collections of time series



Amazon's Warehouse States

The states on this map have warehouses that store and ship inventory for Amazon FBA Sellers.



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 of 15 ETS models that are appropriate to the data

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

auto.arima algorithm

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

ets() and auto.arima() in R

ets algorithm

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

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

auto.arima algorithm

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

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

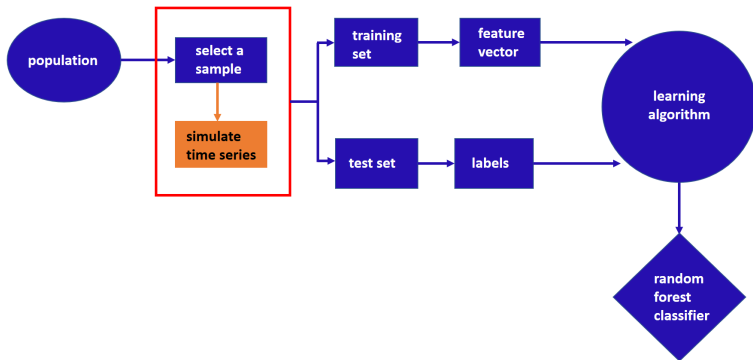
Time series features

Cognostics: Computer-aided diagnostics

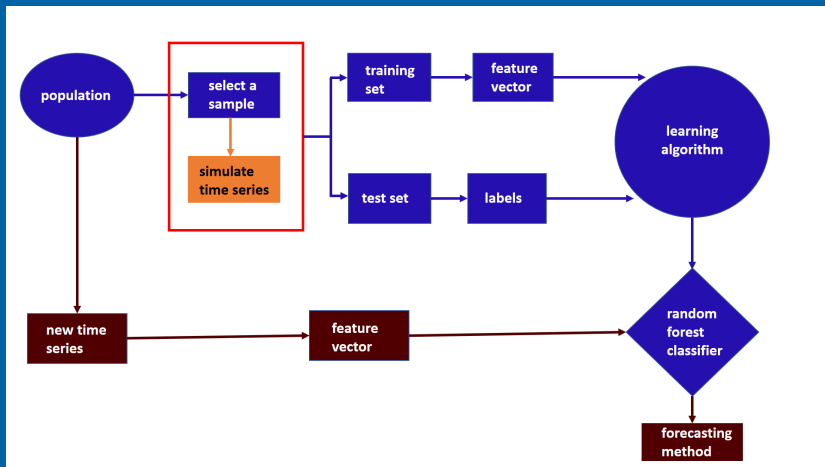
(John W. Tukey, 1985)

- 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

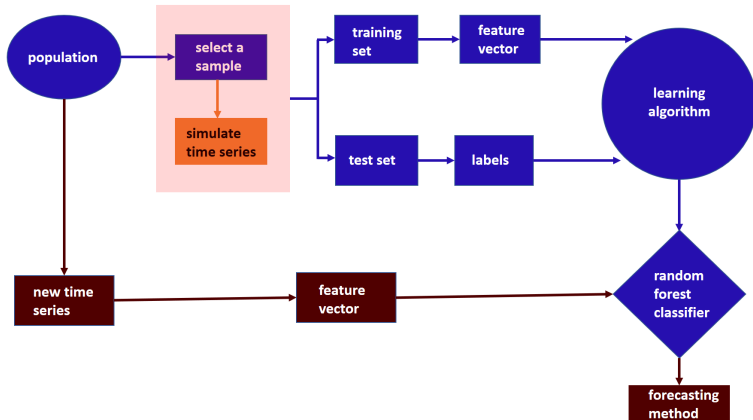
Methodology: “offline” part of the algorithm



Methodology: “online” part of the algorithm

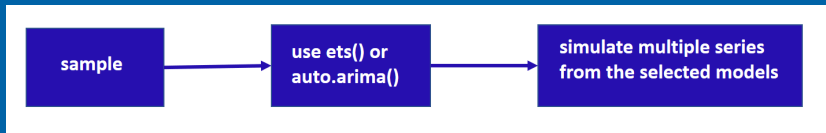


Methodology: reference set

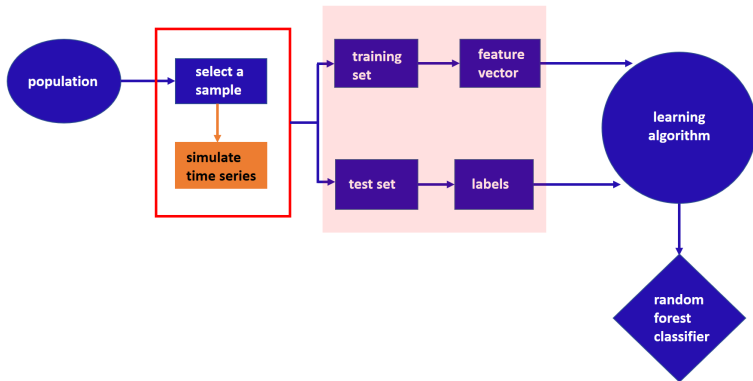


Augmenting the reference set with simulated series

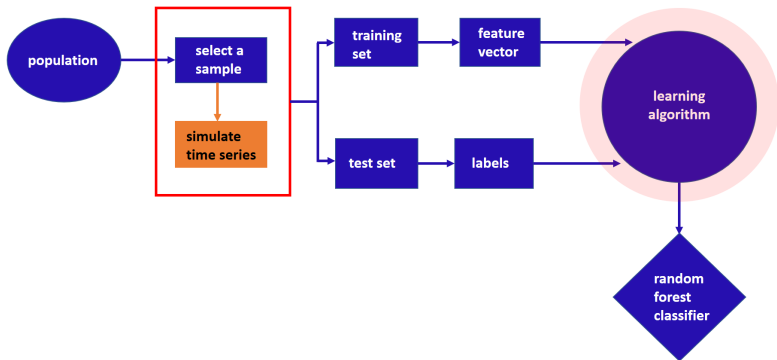
- when our sample is too small to build a reliable classifier
- when we wish to add more of some types of time series to the training set in order to get a more balanced sample
- How?



Methodology: features and class labels

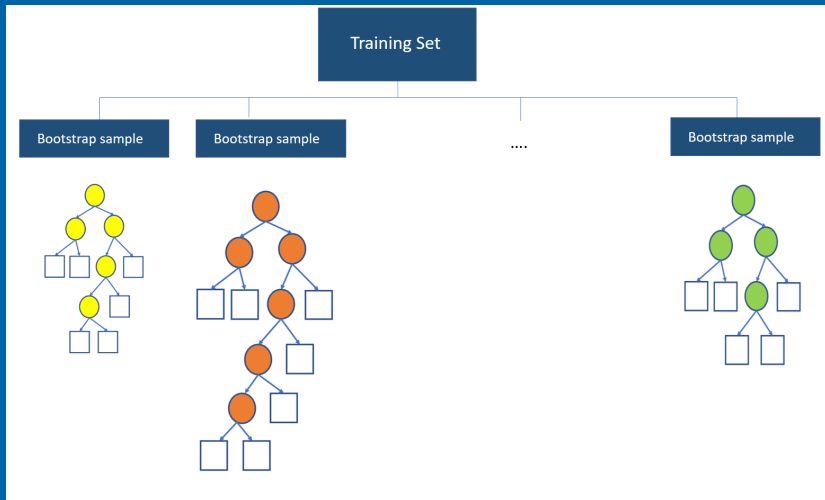


Methodology: random forest





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

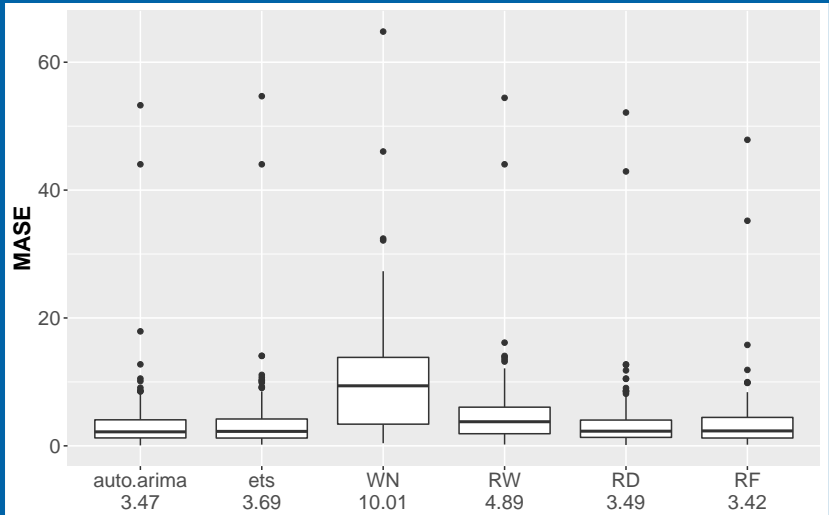
Preliminary study

- 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/thiyanagt/YSC-2017>

thiyanga.talagala@monash.edu