Mining temporal data

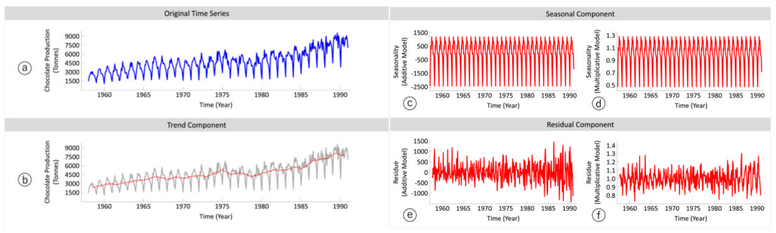
Samuel Gazda, Dominik Macko, Šárka Ščavnická and Katarína Švecová
Faculty of Informatics, Masaryk University, Brno

1. Abstract

Data can be represented in various formats, but the storage of temporal data provided researchers with the possibility of its chronological organisation. Mining the temporal data can bring us an important insight into real-life phenomena. By using machine learning models, we can find patterns in data that would be non-trivial to identify otherwise. Information provided by this can help us classify the data, find interesting properties of sequences, or even carry out predictions. Time series prediction is one of the most interesting, as reducing the future uncertainty is valuable in areas such as stock exchange or infrastructure load.[1] In this review, we explore the possibilities in temporal data exploration. We describe this type of data, explore several parametric and non-parametric models and last but not least we introduce libraries used in the time series mining.

2. Data Characteristics

Temporal data, often called time series, represents chronologically ordered observations of a variable. When there is more than one variable, we consider it to be multivariate. Time series can contain many patterns, like **trend** – long-term decrease or increase, **seasonality** – cyclic changes in constant intervals, and **residue** – short-term fluctuations.

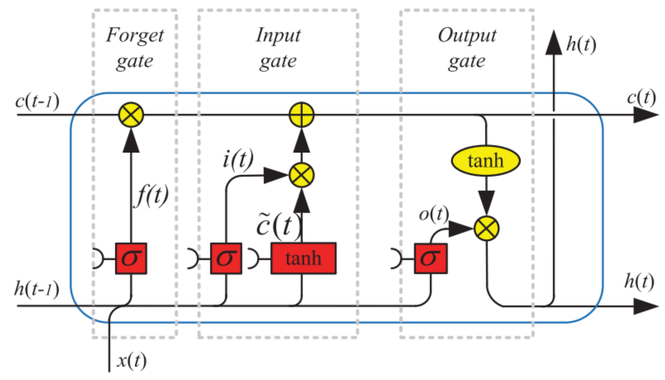


3. Libraries

Because temporal data is very specific, there are many Python libraries intended to be used for it. **Tsfresh** is a library which can be used to automatically extract various features. Moreover, statistical models and tests for time series are available in **statsmodels**. There are also libraries with machine learning models, like **sktime** and **darts**, which have a syntax similar to scikit-learn. Furthermore, libraries such as **PyTorch Forecasting** and **tsai**, which is built on top of PyTorch and fastai, contain various deep learning models useful for time series. AutoML can also be leveraged with temporal data by using specific libraries like **AutoTS**, **AtsPy**, and **PyCaret**. Additionally, **Kats**, a library by Facebook, intends to be a lightweight framework for a complete solution of time series analysis such as forecasting, detection, feature engineering, and even utilities like time series simulation.

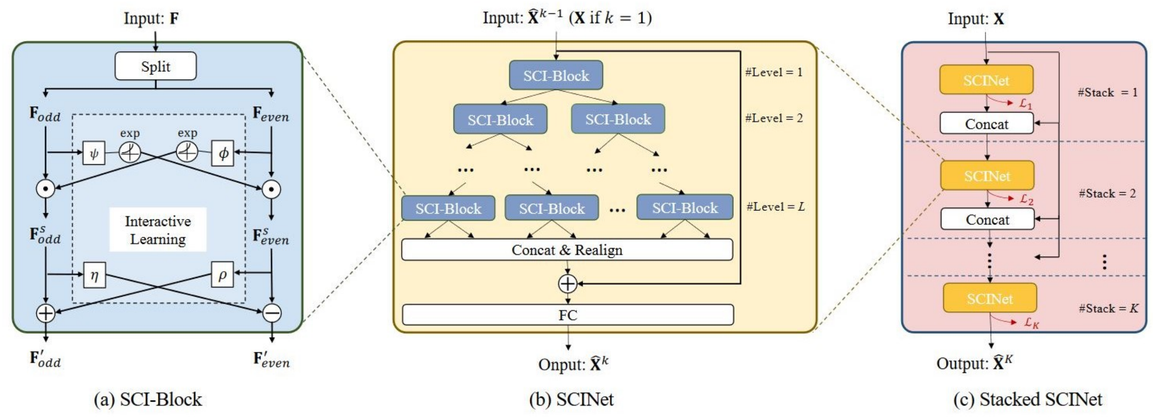
4. LSTM

First introduced in 1997, **Long short-term memory** is a deep-learning recurrent neural network used to learn and predict sequential data. Each LSTM cell has an input, an output gate, and a forget gate to include current and previous states in calculations. Having these extra gates compared to a classic recurrent network helps to solve the vanishing gradient problem. For effective calculation with matrices, several transformer models are used specifically for time series computation, such as probabilistic Informer or deterministic Query Selector. Although LSTM is often outperformed by other models on smaller datasets, to this day, it is used as one of the most common solutions for time series, as the whole model or a part of it. [2, 3]



5. SCI-Net

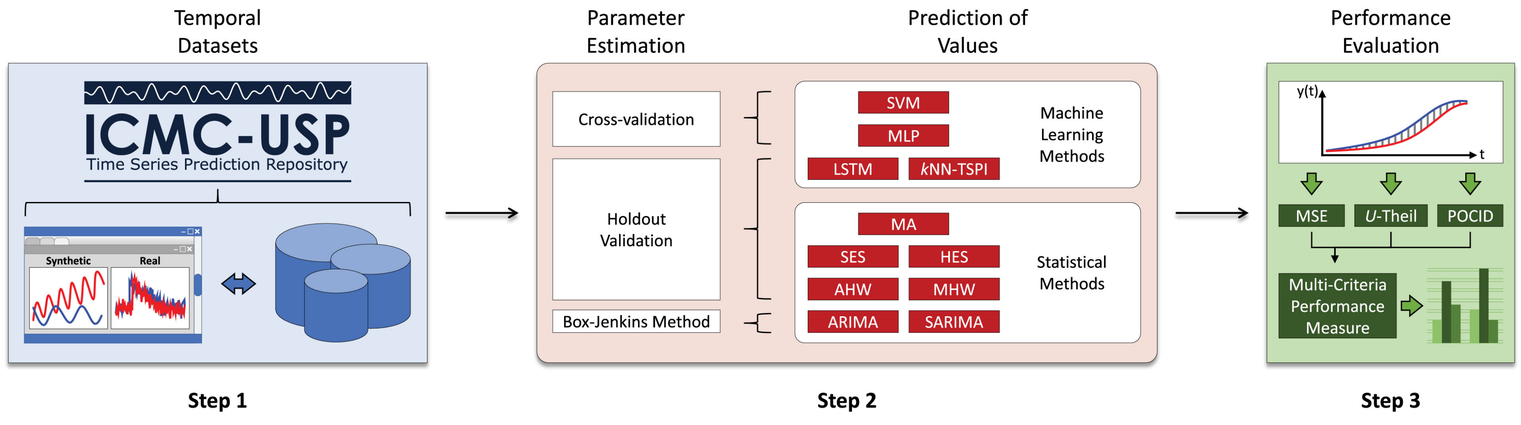
Sample convolutional and interaction network (SCInet) is a special neural network architecture for time series forecasting. Its hierarchical structure iteratively extracts and exchanges information at different temporal resolutions and learns an effective representation for the predictability. SCInet is composed of SCIblocks, which downsamples input data into two sub-sequences using distinct convolutional filters and then perform interactive learning between the two convolutional features. Experiments on real-world datasets show that the SCInet achieved on average more than 40% relative improvement compared to the contemporary state of the art approaches. [4]



(a) SCI-Block (b) SCINet (c) Stacked SCINet

6. SARIMA

Seasonal Autoregressive Integrated Moving Average (SARIMA) is an extension of ARIMA that supports univariate time series data with the seasonal component. The results show that SARIMA is the only statistical method able to outperform (without a statistical difference) these machine learning algorithms: ANN, SVM, and kNN-TSPI. [5]
In the time series prediction is very important to search for the best parameter setting to fit a model according to a dataset. The main parameter estimation methods are **Holdout validation, Cross-validation and Box-Jenkins methods**. Parameters for SARIMA can be defined using the Box-Jenkins method, which is also minimizing the Akaike Information Criterion (AIC). [5]



7. References

[1] Piccialli Mancuso and Sudoso. A machine learning approach for forecasting hierarchical time series. *Expert Systems with Applications*, 182:115102, November 2021.

[2] Cummins Gers, Schmidhuber. Learning to forget: continual prediction with lstm. 2:850–855 vol.2, 1999.

[3] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, 31(7):1235 – 1270, 2019.

[4] Liu Et Al. Time series is a special sequence: Forecasting with sample convolution and interaction. 2021.

[5] Antonio Rafael Sabino Parmezan, Vinicius M.A. Souza, and Gustavo Batista. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Information Sciences*, 484:302–337, 2019.