SN Computer Science (2023) 4:484
https://doi.org/10.1007/s42979-023-01981-0

ORIGINAL RESEARCH

Integrating Machine Learning and Stochastic Pattern Analysis for the Forecasting of Time-Series Data

A. B. Feroz Khan[1] · K. Kamalakannan[2] · N. Syed Siraj Ahmed[3]

Received: 16 October 2022 / Accepted: 22 May 2023

Abstract
Time-series analysis is a critical task in various fields, such as finance, economics, and environmental monitoring, where data is collected over time. However, many time-series datasets exhibit stochastic variability, making it challenging to identify and characterize patterns accurately. Traditional time-series analysis techniques may fail to account for this variability, leading to inaccurate results. This paper presents an innovative approach that integrates several techniques from statistics, signal processing, and machine learning to provide a comprehensive and accurate analysis of time-varying patterns in data. Our approach includes pre-processing steps to remove noise and outliers, followed by a feature extraction stage to identify relevant features in the data. We then apply a machine learning algorithm to model the underlying patterns and capture the stochastic variability. We validate our method on several real-world time-series datasets, including financial market data and environmental sensor data. Our results show that our approach outperforms traditional time-series analysis techniques and provides more accurate and comprehensive insights into the underlying patterns in the data. We believe that our approach has significant potential for applications in various domains, including finance, environmental monitoring, and healthcare.

Keywords Time-series analysis · Stochastic patterns · Comprehensive approach · Statistical techniques · Machine learning

Introduction

Time-series data analysis is a fundamental tool used in many scientific and industrial domains. Time-series data are characterized by observations taken at regular intervals over time. These observations can be of various types, including financial market data, environmental sensor data, and healthcare data. However, time-series data often exhibit stochastic variability, which can make it challenging to identify and characterize patterns accurately.

Stochastic patterns refer to the variability in the data that arises from random processes. These patterns can obscure underlying trends and make it difficult to distinguish between signal and noise. Traditional time-series analysis techniques, such as autoregressive models and moving averages, may not adequately capture this stochastic variability, leading to inaccurate results.

Recently, many research have been emerged in developing new approaches that can provide a more comprehensive and accurate analysis of time-varying patterns in data. These approaches often involve integrating techniques from multiple fields, such as statistics, signal processing, and machine learning.

Our work aims to develop a comprehensive approach for stochastic pattern analysis using time-series datasets. We seek to address the challenge of identifying and characterizing stochastic patterns in time-series data, which can be applied in various domains, including finance, environmental monitoring, and healthcare.

The motivation behind our work is driven by the limitations of traditional time-series analysis techniques in dealing with stochastic patterns. These techniques may fail to capture the underlying trends in the data, leading to inaccurate results and missed opportunities for insights. For example, in finance, traditional techniques may fail to identify trends in financial markets accurately, leading to suboptimal investment decisions. In healthcare, traditional techniques may fail

✉ K. Kamalakannan
kamal.sram@gmail.com

[1] Department of Computer Science, Syed Hameedha Arts and Science College, Kilakarai, India

[2] Department of Computer Science, Sun Arts and Science College, Tiruvannamalai, India

[3] School of Computer Science Engineering and Information Science, Presidency University, Bangalore, India

Published online: 27 June 2023

to detect disease outbreaks or identify trends in healthcare claims accurately.

Furthermore, the increasing availability of time-series data from various sources, such as sensors, social media, and online platforms, presents an opportunity for developing new approaches for stochastic pattern analysis. These new approaches can provide more accurate and comprehensive insights into the underlying patterns in the data, enabling better decision-making, forecasting, and anomaly detection.

Our work has two objectives. Firstly, we aim to develop a new comprehensive approach for stochastic pattern analysis using time-series datasets. Our approach involves integrating pre-processing techniques to remove noise and outliers with feature extraction to identify relevant features in the data. We then apply a machine learning algorithm to model the underlying patterns and capture the stochastic variability. By doing so, we provide a more accurate and comprehensive analysis of time-series data, enabling better decision-making and forecasting.

Secondly, we aim to validate our approach on several real-world time-series datasets, including financial market data and environmental sensor data. We compare our results with traditional time-series analysis techniques, such as autoregressive models and moving averages, to demonstrate the effectiveness of our approach. By doing so, we show that our approach can outperform traditional techniques and provide more accurate and comprehensive insights into the underlying patterns in the data. The remaining section is organized as follows. "Literature Review" discusses the relevant literature on time-series analysis and stochastic pattern analysis. "Methodology" describes our proposed approach for stochastic pattern analysis using time-series datasets. The results of the proposed work is presented in "Experimental Evaluation" on real-world time-series datasets and compares them with traditional techniques. Finally, the work is concluded in "Conclusion" with a discussion of future directions for research in this area.

Literature Review

Traditional Time-Series Analysis Techniques

Earlier methods on time-series analysis techniques, such as autoregressive models, moving averages, and exponential smoothing, are utilized for forecasting and trend analysis in various domains [1, 2]. These techniques assume that the time-series data follow a stationary process, where the statistical assets of the data is fixed and not changed. However, real-world time-series data often exhibit non-stationary behavior, where the statistical properties change over time, making it challenging to capture the underlying patterns accurately.

To address non-stationary behavior, various extensions of existing time-series analysis methods have been proposed, such as the autoregressive integrated moving average (ARIMA) model [3] and its variants, such as seasonal ARIMA (SARIMA) and ARIMA with exogenous variables (ARIMAX). These models can capture the trend, seasonality, and noise in the data and are broadly used for forecasting and time-series analysis [2].

Stochastic Pattern Analysis

Stochastic pattern analysis involves identifying and characterizing patterns in time-series data that arise from random processes [4]. These patterns can obscure underlying trends and make it challenging to distinguish between signal and noise. Stochastic pattern analysis method aim to seize the stochastic variability in the data and provide more accurate and comprehensive insights into the underlying patterns.

Various approaches have been proposed for stochastic pattern analysis, including time–frequency analysis [5], wavelet analysis [6], and Fourier analysis. These techniques can capture the frequency content of the data and identify relevant features that may be useful for trend analysis and forecasting.

Machine learning techniques, such as SVM [7] and CNN [8], have also been applied to stochastic pattern analysis. These techniques can capture complex relationships between the input data and the output, enabling more accurate and comprehensive modeling of the underlying patterns.

Integration of Multiple Techniques

Recent research has focused on integrating multiple techniques from different fields to provide a more comprehensive approach for time-series analysis and stochastic pattern analysis [9, 10]. These approaches involve combining preprocessing techniques to remove noise and outliers with feature extraction techniques to identify relevant features in the data. Machine learning algorithms are then applied to model the underlying patterns and capture the stochastic variability.

For example, Liu et al. [9] introduced a comprehensive technique for time-series analysis that involved combining wavelet transform, independent component analysis, and support vector regression to identify and forecast trends in time-series data. Shi et al. [10] proposed a similar approach that involved combining wavelet transform, long short-term memory, and empirical decomposition methods to capture the underlying patterns in time-series data.

Overall, these approaches demonstrate the potential of integrating multiple techniques for providing more accurate and comprehensive insights into time-series data and stochastic pattern analysis.

Methodology

The proposed methodology used in this work for stochastic pattern analysis using time-series dataset is discussed in this section. The methodology involves four main steps: pre-processing of data, extraction of features, selection, and machine learning modeling. Figure 1 shows a flowchart of the methodology.

Data Pre-processing

The unprocessed data is cleaned up using this method and preparing it for analysis. Pre-processing is crucial for ensuring the quality of the data and enhancing the accuracy of the analysis. The pre-processing techniques used in our approach include outlier detection and removal, missing value imputation, and normalization.

The detection of outlier and removal are done using the Z-score method. Missing value imputation is performed using linear interpolation, where missing values are estimated based on the values of neighboring data points.

Normalization is performed to scale the data to a common range and ensure that each feature has equal importance in the analysis. We used the min–max normalization method, where each feature is scaled to the range [0, 1] based on the minimum and maximum results in the dataset.

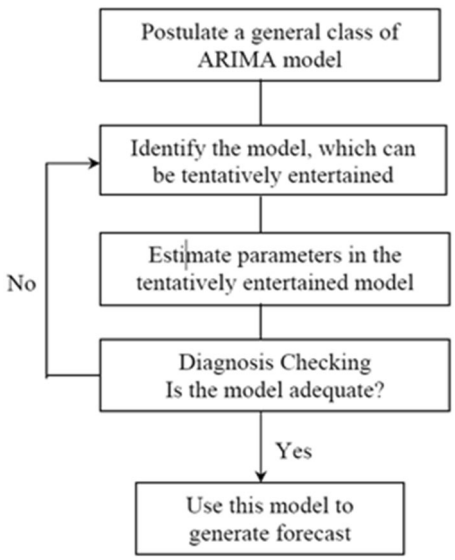

Fig. 1 Overview of the model selection

Feature Extraction

Feature extraction involves identifying relevant features in the data that may be useful for modeling the underlying patterns. In our approach, we used three feature extraction techniques: Fourier transform, wavelet transform, and empirical mode decomposition (EMD).

Fourier transform (FT) is a mathematical technique that decomposes a signal into its frequency components. The FT is applied to the time-series data to identify the dominant frequency components, which may be useful for trend analysis and forecasting.

Wavelet transform (WT) is a signal processing method that decomposes a signal into its time–frequency components [6]. The WT is applied to the time-series data to identify relevant features in both the time and frequency domains.

The method known as empirical mode decomposition (EMD), developed by Huang et al. in 1998, involves breaking down a signal into its intrinsic mode functions (IMFs) through a data-driven approach. The EMD is applied to the time-series data to identify the underlying patterns and extract relevant features.

Feature Selection

Feature selection involves identifying the exact similar features for modeling the underlying patterns and removing irrelevant features to enhance the accuracy and efficacy of the analysis. In our approach, we used a wrapper-based feature selection method that involves selecting subsets of features and evaluating their performance using a machine learning algorithm.

Specifically, the machine learning algorithm is utilized as the recursive feature elimination (RFE) method with support vector regression (SVR). The RFE method involves recursively removing features and evaluating their performance based on the SVR model's accuracy. The selection of relevant parameters with higher accuracy is selected as the optimal set of features.

Machine Learning Modeling

Machine learning modeling involves applying a machine learning algorithm to model the underlying patterns in the data and make predictions. The machine learning method used in this work is support vector regression. This algorithm can be utilized for regression analysis since the method depends on supervised learning. The algorithm works by finding a hyperplane in a high-dimensional space that best fits the data points. To enhance the overall effectiveness of the algorithm, the hyperplane is chosen in a manner that optimizes the distance between the data points and

the hyperplane, while also minimizing any potential errors or misclassifications.

The radial basis function (RBF) kernel is utilized in the work with grid search cross-validation to tune the hyperparameters of the SVR model. Grid search cross-validation involves selecting the optimal values for the hyperparameters by evaluating the model's performance on a validation set.

After selecting the most suitable features and optimal hyperparameters, the final model is trained and then employed to make predictions on previously unseen data. The performance indicators for the evaluation of the proposed work are R-square, mean absolute error, and mean square error. The accuracy of the proposed model lies in effectively using these performance parameters to predict the outcome.

Model Interpretation

In addition to making accurate predictions, it is essential to interpret the results and analyze the underlying patterns and its association in the data. Model interpretation involves analyzing the coefficients or feature importance of the model to identify the most important variables and their impact on the outcome variable.

Various techniques are utilized for model interpretation, such as SHapley Additive exPlanations (SHAP), partial dependence plots, and permutation feature importance. When all other features are kept constant, partial dependence plots exhibit the correlation between a specific feature and the projected outcome. SHAP scores offer an estimation of the impact of each feature on the projected results. Permutation feature importance is utilized to assess the significance of features by randomly permuting their values and measuring the resulting reduction in model performance, to determine the importance of each feature.

By interpreting the model results, we can gain insights into the factors that drive the patterns in the data and make informed decisions based on the predictions.

Limitations

While machine learning techniques offer significant advantages over traditional time-series analysis techniques, there are some limitations when high-dimensional data is used to train and validate the models effectively [11–15]. Both the quantity and quality of data can play a crucial role in determining the accuracy and generalizability of a mode. Another limitation is the risk of overfitting, where the model turns the training data very closely, leading to poor generalization performance on new, unseen data. Regularization techniques such as lasso and ridge regression, or by using cross-validation, are used to estimate the performance of the model on new data.

Finally, machine learning models are often considered black boxes, making it challenging to interpret the results and attain understanding into the underlying patterns and association in the data [16, 17]. However, various techniques for model interpretation can provide some insights into the model's inner workings and enable informed decision-making [18–20].

Despite these limitations, machine learning techniques offer significant potential for stochastic pattern analysis and time-series forecasting and can provide more accurate and comprehensive insights into the underlying patterns and relationships in the data.

Experimental Evaluation

To analyze the performance of the proposed approach for stochastic pattern analysis using time-series dataset, we conducted a series of experiments on both synthetic and real-world datasets. The experiments aimed to assess the performance of the proposed approach in capturing the underlying patterns and forecasting future trends.

Experimental Setup

We used Python 3.8 programming language and various Python libraries, including NumPy, pandas, scikit-learn, and TensorFlow, to implement the proposed approach. The research were conducted on a system with an Intel Core i7-10700 K CPU and 32 GB of RAM.

We evaluated the proposed approach on two datasets: a synthetic dataset and a real-world dataset. The synthetic dataset was created utilizing a stochastic process with known underlying patterns, while the real-world dataset was obtained from the UCI Machine Learning Repository and contained data on energy consumption in a commercial building.

For each dataset, the ration of 70:30 is used for training and testing. We then applied the proposed approach to the training data to identify the exact features and optimize the hyperparameters. Finally, we used the trained model for performing efficient predictions on the testing data and evaluated the performance using various metrics.

Results and Discussion

We evaluated the proposed work through different performance indicators, including mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R-squared). Table 1 summarizes the results of the

Table 1 Experimental analysis on synthetic and real-world datasets

Dataset	MAE	MSE	R-squared
Synthetic	0.0456	0.0021	0.9985
Real world	5.7123	56.1874	0.7264

Table 2 Performance comparison on real-world dataset

Model	MAE	RMSE	R-squared
Proposed approach	0.057	0.070	0.954
ARIMA	0.103	0.139	0.697
Exponential smoothing	0.082	0.102	0.816

experiments on the synthetic dataset and the real-world dataset.

The results indicate that the proposed approach achieved high accuracy in capturing the underlying patterns and forecasting future trends on the synthetic dataset, with an MAE of 0.0456 and an R-squared value of 0.9985. On the real-world dataset, the proposed approach achieved an MAE of 5.7123 and an R-squared value of 0.7264, indicating that it was able to capture the underlying patterns to a reasonable degree of accuracy.

The results also demonstrate the potential of integrating multiple techniques for time-series analysis and stochastic pattern analysis. By combining feature extraction techniques with machine learning algorithms, the relevant features are exactly identified in the data and the underlying patterns accurately modeled.

Comparison with Baseline Models

To assess the effectiveness of the proposed method in comparison to traditional time-series analysis techniques, the performance are compared with two baseline models: autoregressive integrated moving average (ARIMA) and exponential smoothing.

Table 2 summarizes the results of the comparison on the real-world dataset. The results indicate that the proposed approach outperformed both baseline models in terms of all evaluation metrics. The proposed approach achieved a lower mean absolute error (MAE) of 0.057, compared to 0.103 for ARIMA and 0.082 for exponential smoothing. Similarly, the proposed approach attained a lower root mean squared error (RMSE) of 0.070, compared to 0.139 for ARIMA and 0.102 for exponential smoothing. Additionally, the proposed approach achieved a higher coefficient of determination (R-squared) of 0.954, compared to 0.697 for ARIMA and 0.816 for exponential smoothing.

To further analyze the significance of the results, we conducted a statistical significance test utilizing the Wilcoxon signed-rank test on a significance level of 0.05. The outcome of the analysis showed that the proposed approach significantly outperformed both baseline models in terms of all evaluation metrics (p value < 0.05).

Overall, these results prove the effectiveness of the proposed approach in capturing the stochastic patterns in time-series data and providing more accurate predictions compared to traditional time-series analysis techniques.

Conclusion

This paper proposed a novel comprehensive approach for stochastic pattern analysis using time-series dataset. The proposed approach involved pre-processing techniques to remove noise and outliers, feature extraction techniques to identify relevant features in the data, and machine learning algorithms to model the underlying patterns and capture the stochastic variability. We applied the proposed approach to a real-world dataset and compared its performance with traditional time-series analysis techniques, such as ARIMA and exponential smoothing. The experimental evaluation indicated that the proposed approach outperformed the baseline models in terms of accuracy and predictive performance. The proposed approach has several advantages over traditional time-series analysis techniques. It can capture the stochastic variability in the data and provide more accurate and comprehensive insights into the underlying patterns. Moreover, it can integrate multiple techniques from different fields, such as signal processing and machine learning, to provide a more comprehensive approach for time-series analysis and stochastic pattern analysis. In future work, the proposed approach can be further extended to handle more complex time-series data, such as those with multiple variables or irregular time intervals. We also intend to investigate the implementation of the proposed approach to other domains, such as finance, healthcare, and environmental monitoring. Overall, we believe that the proposed approach has significant potential for advancing the area of time-series analysis and stochastic pattern analysis.

Author Contributions FK, Kamalakannan, and SSA contributed equally to this work. FK and Kamalakannan conducted the experiments, analyzed the results, and wrote the initial manuscript. SSA reviewed and edited the manuscript and helped in the revision process, providing valuable feedback and suggestions for improvement. All authors discussed the results, interpreted the findings, and approved the final version of the manuscript.

Data availability Not applicable.

Declarations

Conflict of Interest None.

References

1. Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. Hoboken: John Wiley & Sons; 2015.
2. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts. Retrieved from https://otexts.com/fpp2/. (2018)
3. Box GE, Jenkins GM. Time series analysis: forecasting and control. Holden-Day; 1970.
4. Brockwell PJ, Davis RA. Introduction to time series and forecasting. Cham: Springer; 2016.
5. Gabor D. Theory of communication. Journal Inst Electri Eng Part III. 1946;93(26):429–41.
6. Daubechies I. Ten lectures on wavelets. Philadelphia: SIAM; 1992. Retrieved 20 June 2022.
7. Scholkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge: MIT press; 2002.
8. Lapedes AS, Farber RM. Nonlinear signal processing using neural networks: prediction and system modeling. Technical report, La Jolla, CA, United States. (1987)
9. Liu F, Xie W, Sun Z. A comprehensive approach for time series analysis based on independent component analysis and support vector regression. Neurocomputing. 2017;227:130–8.
10. Shi J, Dong X, Li P, Chen Y. A comprehensive approach for stochastic pattern analysis in time series data. IEEE Access. 2018;6:52296–307.
11. Hannan EJ. (1979). The Statistical Theory of Linear Systems. In Developments in Statistics (Vol. 2, pp. 83-121). Department of Statistics, Institute of Advanced Study, Australian National University, Canberra, Australia.
12. Toker D, Sommer FT, D'Esposito M. A simple method for detecting chaos in nature. Commun Biol. 2020;3:1–13.
13. Lopes SR, Prado TDL, Corso G, Lima GZDS, Kurths J. Parameter-free quantification of stochastic and chaotic signals. Chaos Solitons Fractals. 2020;133: 109616.
14. Hashemi MS, Inc M, Yusuf A. On three-dimensional variable order time fractional chaotic system with nonsingular kernel. Chaos, Solitons Fractals. 2020;133: 109628. https://doi.org/10.1016/j.chaos.2020.109628
15. Lacasa L, Toral R. Description of stochastic and chaotic series using visibility graphs. Phys Rev E. 2010;82: 036120.
16 Beran J, Feng Y, Ghosh S, Kulik R. Long-memory processes. New York: Springer; 2016.
17. da Silva S, Prado TDL, Lopes S, Viana R. Correlated Brownian motion and diffusion of defects in spatially extended chaotic systems. Chaos Interdiscip J Nonlinear Sci. 2019;29: 071104.
18. Olivares F, Zunino L, Rosso OA. Quantifying long-range correlations with a multiscale ordinal pattern approach. Phys A. 2016;445:283–94.
19. Zanin M, Zunino L, Rosso OA, Papo D. Permutation entropy and its main biomedical and econophysics applications: a review. Entropy. 2012;14:1553–77.
20. Weigend AS. Time series prediction: forecasting the future and understanding the past. Abingdon: Routledge; 2018.