Time series forecasting method based on frequent pattern mining

Jintao Qu1,2, Xiaohui Cui1,2*, Xiuyang Meng1, Senyang Zheng1

1 School of Information Science and Technology, Beijing Forestry University, Beijing, 100083, China
2 Engineering Research Center for Forestry-orientend Intelligent Information Processing, National Forestry and Grassland Administration, Beijing, 100083, China

*Corresponding author’s e-mail: cuixiaohui@bjfu.edu.cn

Abstract. Pattern Sequence-based Forecasting (PSF) is an effective method for time series prediction. However, the accuracy of this method depends on the selection of parameters such as the length of the pattern sequence and the number of clusters. In diverse time series data sets, these parameters are often priori unknown. This paper innovatively introduces a pattern mining method before the PSF pattern clustering to guide the clustering process and realize the automation of initial parameter selection. Experimental results show that the method proposed in this paper effectively eliminates the uncertainty of PSF caused by the selection of initial parameters. Compared with the original model, it improves the efficiency while ensuring the advantage of prediction accuracy.

1. Introduction
The PSF proposed in 2011 is a model sequence-based forecasting method, which has good robustness and interpretability. However, the limitation of this model is that the quality of its prediction results is extremely dependent on its previous pattern clustering process. This unsupervised clustering process often requires human intervention because the cluster center is difficult to determine.

This paper takes time series pattern mining as the main method to improve the efficiency and accuracy of PSF prediction. Time series pattern mining is an NP-hard problem. Since the concept of time series pattern was formed in 2002, dozens of researchers have applied it to medicine[1], biology[2], telemedicine[3], etc. field. Lin et al. [4] and Chi et al. [5] both use data discretization, segmented approximation and other means to pre-process the sequence to reduce the time complexity of the algorithm. The precise algorithm proposed by Mueen and Keogh[6] uses a triangle inequality with sliding window pruning to detect the pattern sequence in the insect's probe potential map. These methods can theoretically achieve a low time complexity, but because they require considerable preprocessing, the constant factor of the algorithm is very large.

In order to solve the time complexity problem of time series pattern mining, this paper is designed from the perspective of intelligent optimization, inspired by co-evolution strategies, and based on the computational intelligence framework from problem encoding, population generation strategies, evolution strategies, topological structures, etc. Intelligent mode mining method with strong robustness and adaptability. Through the pattern mining method, the sequence frequent pattern in the data set is used as the clustering center, which realizes the automatic selection of initial parameters and improves the prediction accuracy.
2. The Proposed Methodology
The algorithm flow consists of two parts: sequential pattern mining and PSF. The pattern mining method is the basis of PSF. Sequence pattern mining adopts heuristic methods to adapt to the pattern mining problem of different characteristic time series. In the PSF prediction process, we ignore details such as data normalization and pre-processing, because these processes require specific analysis based on specific issues.

![Figure 1. The Proposed Methodology](image)

2.1. The PSF Algorithm
The principle of the PSF algorithm is simple, but it is very effective in many time series forecasting problems. PSF is a prediction method based on the similarity of pattern sequences [7]. Differently from the earliest proposed PSF method, in this article, we calculate the predicted value of future unit time by weighted average based on the similarity between the current sequence and frequent patterns. It is worth mentioning that in the PSF method of this paper, the sequence fragments participating in the clustering include all the continuous substrings of length $w$ in the whole timeseries data, instead of simple sequence cutting, which effectively shields the interference of the noise sequence to the prediction result and enhances the interpretability of the model.

![Figure 2. Principle of PSF](image)

2.2. The Frequent Pattern Mining Algorithm
Pattern mining before pattern clustering is the key to improving prediction performance. We transform the time series pattern mining problem into a three-dimensional optimization problem, and solve this problem through a particle swarm optimization algorithm, which can obtain a relatively optimal solution within a controllable time.
2.2.1 Problem Encoding
The design of the encode method is an extremely critical element in the use of intelligent optimization algorithms to solve practical problems. In the problem of time series pattern mining, for a time series \( t \) of length \( n \).

\[
t = [t_1, t_2, t_3, \ldots, t_n]
\]

We want to find the \( k \) most similar pattern sequences in \( t \) as pattern clustering centers. We use a three-dimensional vector \([a, b, w]\) to encode a pair of subsequences of length \( w \):

\[
([t_{a}, t_{a+1}, t_{a+2}, \ldots, t_{a+w+1}], [t_{b}, t_{b+1}, t_{b+2}, t_{b+w+1}])
\]

It is easy to prove that this encoding method can be mapped to all points in the solution space, and a code also uniquely corresponds to a certain potential solution in the problem space, which satisfies the completeness and reliability of the coding of the intelligent optimization algorithm.

2.2.2 Population generation strategy
The dispersion of the initial population in the entire solution space plays a vital role in the effect of pattern mining. Integrating the performance of multiple generation strategies on benchmarks, we innovatively generate random initial populations through Logistic mapping. Compared with the traditional pseudo-random number generation of the initial solution, the chaotic method can make the initial population cover the solution space more evenly and achieve a more comprehensive search for the solution space.

2.2.3 Evolution strategy
The traditional evolution strategy is only based on the evolution of the individual's own fitness, and does not consider the influence of the evolved environment and the complex connection between the individual on the evolution of the individual. In the application, defects such as immature convergence and slow convergence are shown.

The pattern mining method proposed in this article adopts a gradual evolution strategy. When a certain number of times of evolution and no better solution is produced, the step length of evolution will decrease exponentially. This reduce unnecessary calculations and speed up the convergence rate.

2.2.4 Fitness calculation
In the execution of the optimization algorithm, the sequence similarity, as the fitness of the individual, has a vital influence on the efficiency and accuracy of the entire algorithm. Euclidean distance and dynamic time warping distance are commonly used indicators for calculating similarity of time series. Dynamic time warping is more flexible and has stronger fault tolerance and adaptability. However, the time complexity of its calculation process is too high, and it can easily become the performance bottleneck of the entire optimization algorithm. In addition, the paper by Mueen et al[7], pointed out that the optimization effect of dynamic time warping distance in the pattern mining problem is not obvious. Comprehensive consideration, we use Euclidean distance as a measure of sequence similarity.

3. Experiments and results

3.1. Data source and experimental environment
The performance of the pattern mining method proposed in this paper on the benchmark function is better than the traditional particle swarm algorithm in terms of accuracy and convergence rate. The experiment uses Sugon I840-G30 server as the experimental hardware environment.

3.2. Experimental results
We spliced time series data sets such as EGG, EOG, CARCOUNT, dowjones, LSFS_10, etc. into a long time series with a length of 1.2 million, input them into the pattern mining method, and discovered the following frequent patterns.
On the basis of the frequent pattern mining algorithm, with reference to related work, the model is used to forecast the trend of the New York electricity price market for the whole year of 2005 month by month. After averaging multiple forecasts, the annual MER rate is 8.33%, which is slightly better than the 8.41% obtained by the single PSF method. That is, our new method ensures the accuracy of prediction while realizing automated modeling.

4. Conclusion
This paper designs a time series forecasting method based on pattern mining, analyzes the basic principles of the algorithm, and gives the technical realization points of several core optimization links. This method can effectively solve the time series forecasting problems in certain fields, and can realize the automatic selection of initial key parameters. This paper mainly uses the pattern mining method to optimize prediction. The heuristic pattern mining method proposed in this paper performs well in the pure pattern mining scenario, and has certain application potential in related fields.

Acknowledgments
This work was supported by National College Students' innovation and entrepreneurship training program (No. G201910022063).

References
[1] Abe, H., Ohsaki, M., Yokoi, H., & Yamaguchi, T. (2005, June). Implementing an integrated time-series data mining environment based on temporal pattern extraction methods: a case study of an interferon therapy risk mining for chronic hepatitis. In Annual Conference of the Japanese Society for Artificial Intelligence (pp. 425-435). Springer, Berlin, Heidelberg.
[2] Androulakis, I. P. (2005). Selecting maximally informative genes. Computers & chemical engineering, 29(3), 535-546.
[3] Arita, D., Yoshimatsu, H., & Taniguchi, R. I. (2005, June). Frequent motion pattern extraction for motion recognition in real-time human proxy. In Proc. of JSAI Workshop on Conversational Informatics (pp. 25-30).
[4] Lonardi, J. L. E. K. S., & Patel, P. (2002). Finding motifs in time series. In Proc. of the 2nd Workshop on Temporal Data Mining (pp. 53-68).
[5] Chiu, B., Keogh, E., & Lonardi, S. (2003, August). Probabilistic discovery of time series motifs. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 493-498).
[6] Mueen, A., Keogh, E., Zhu, Q., Cash, S., & Westover, B. (2009, April). Exact discovery of time series motifs. In Proceedings of the 2009 SIAM international conference on data mining (pp. 473-484). Society for Industrial and Applied Mathematics.
[7] Alvarez, F. M., Troncoso, A., Riquelme, J. C., & Ruiz, J. S. A. (2010). Energy time series forecasting based on pattern sequence similarity. IEEE Transactions on Knowledge and Data Engineering, 23(8), 1230-1243.