

# OPR-Miner: Order-preserving rule mining for time series

Youxi Wu<sup>a,b</sup>, Xiaoqian Zhao<sup>a</sup>, Yan Li<sup>c,\*</sup>, Lei Guo<sup>d</sup>, Xingquan Zhu<sup>e</sup>, Philippe Fournier-Viger<sup>f</sup>, Xindong Wu<sup>g</sup>

<sup>a</sup>School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China

<sup>b</sup>Hebei Key Laboratory of Big Data Computing, Tianjin 300401, China

<sup>c</sup>School of Economics and Management, Hebei University of Technology, Tianjin 300401, China

<sup>d</sup>State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, Tianjin 300401, China

<sup>e</sup>The Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University, FL 33431, USA

<sup>f</sup>Shenzhen University, Shenzhen, China

<sup>g</sup>Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education), Hefei University of Technology, Hefei 230009, China

---

## Abstract

Discovering frequent trends in time series is a critical task in data mining. Recently, order-preserving matching was proposed to find all occurrences of a pattern in a time series, where the pattern is a relative order (regarded as a trend) and an occurrence is a sub-time series whose relative order coincides with the pattern. Inspired by the order-preserving matching, the existing order-preserving pattern (OPP) mining algorithm employs order-preserving matching to calculate the support, which leads to low efficiency. To address this deficiency, this paper proposes an algorithm called efficient frequent OPP miner (EFO-Miner) to find all frequent OPPs. EFO-Miner is composed of four parts: a pattern fusion strategy to generate candidate patterns, a matching process for the results of sub-patterns to calculate the support of super-patterns, a screening strategy to dynamically reduce the size of prefix and suffix arrays, and a pruning strategy to further dynamically prune candidate patterns. Moreover, this paper explores the order-preserving rule (OPR) mining and proposes an algorithm called OPR-Miner to discover strong rules from all frequent OPPs using EFO-Miner. Experimental results verify that OPR-Miner gives better performance than other competitive algorithms. More importantly, clustering and classification experiments further validate that OPR-Miner achieves good performance.

*Keywords:* pattern mining, rule mining, time series, order-preserving, frequent trend

---

## 1. Introduction

A time series is a continuous numerical series of data or a group of real values that is commonly used in many fields, such as brain EEG clustering [1], stock prediction [2], and weather forecasting [3]. Many studies have been investigated. For example, Wu and Keogh [4] focused on time series anomaly detection. Rezvani et al. [5] studied a new pattern representation method for time series data to effectively detect the change point. Sequential pattern mining method, as a commonly used method, can also be used to discover patterns of interest to users in time series [6] after discretizing the time series into symbols. Note that although in episode mining, an event sequence has a set of consecutive time stamps [7, 8], it is far different from time series, since an event sequence is a group of discrete events, while time series is a group of continuous numerical values. Therefore, users can directly apply the episode mining methods on event sequences [9, 10], while users have to adopt some discretization methods at first, and then apply some sequential pattern mining methods on time sequence.

However, the existing discretizing methods pay too much attention to the values, such as piecewise linear approximation (PAA) [11] and symbolic aggregate approximation (SAX) [12]. Therefore, it is difficult to discover the frequent trends using sequential pattern mining methods. To address this deficiency, several methods have been investigated to find subsequences with the same trend, such as (delta, gamma) approximate matching [13, 14], weak gap

---

\*Corresponding author

Email address: lywuc@163.com (Yan Li)

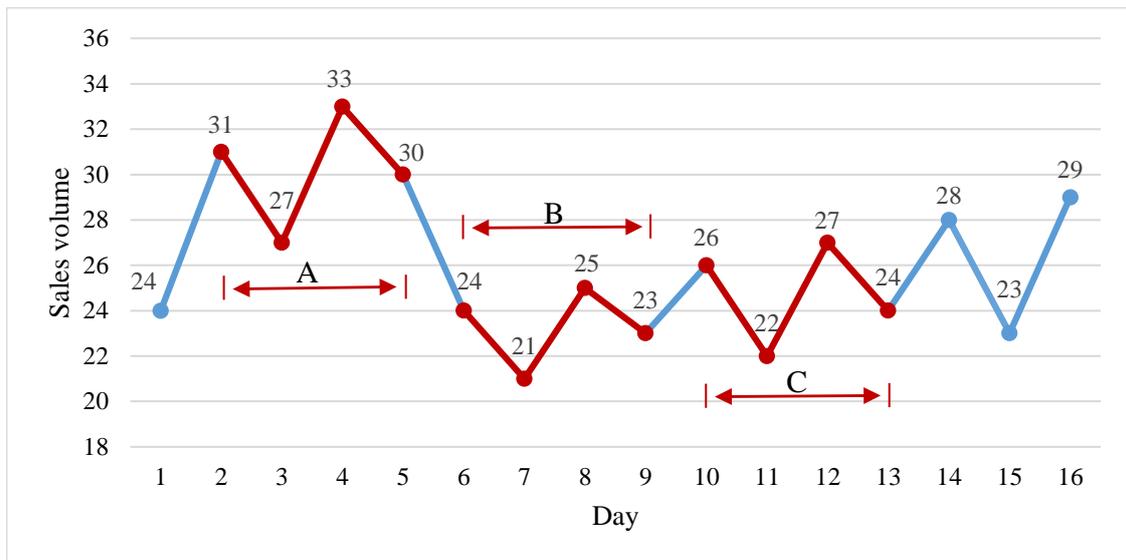


Figure 1: Sales volume of goods over 16 days. The relative order of sub-time series  $(t_2, t_3, t_4, t_5)=(31,27,33,30)$  is  $(3,1,4,2)$ , since 31 is the third smallest, 27 is the smallest, and so on. It can be seen that the trends in the sub-time series marked in red are exactly the same. If  $minsup = 3$ , then  $(3,1,4,2)$  is a frequent OPP, and OPP mining can discover similar frequently occurring trends.

strong pattern mining [15], and tri-way pattern mining [16, 17]. These methods need to set the parameters manually, which may cause the loss of important information in the process and destroy the continuity of the time series.

Recently, order-preserving matching [18, 19] (or called consecutive permutation pattern matching [20]) has been proposed, which does not need to discretize real numbers into symbols. Order preserving matching can find all occurrences of a pattern in a time series, where the pattern is a relative order (regarded as a trend) and an occurrence is a sub-time series whose relative order coincides with the pattern. Inspired by order-preserving matching, our previous work proposed the order-preserving pattern mining (OPP-Miner) algorithm [21], which used the relative order of real values to express a pattern called an order-preserving pattern (OPP). By mining OPPs, we can find frequent trends in a time series. An illustrative example is shown in Fig. 1. In the figure, regions A, B, and C have different means and variances, and the means of A, B, and C are 30.25, 23.25, and 24.75, respectively. The variances of A, B, and C are 4.69, 2.19, and 3.69, respectively. Finding patterns from such non-stationary data is challenging, because of the changing mean and variance. On the other hand, patterns may continually repeat themselves but with different mean values. For example, over years, the stock index has increased many times (showing an increased mean value), whereas the market patterns are rather similar. By observing order of patterns within a local region, OPP mining can find repetitive patterns with different mean values.

However, there are two problems with OPP mining: (i) OPP-Miner [21] adopts a pattern matching method to calculate pattern support. Although the space complexity of OPP-Miner is low, its efficiency is also low, since it does not use the calculation results of the sub-patterns. Hence, the efficiency of OPP-Miner needs to be improved. (ii) More importantly, although all OPPs can be discovered, how to further apply these mining patterns has not been deeply explored.

To improve the performance of OPP-Miner, we propose an algorithm called efficient frequent order-preserving pattern miner (EFO-Miner). Moreover, to utilize these OPPs effectively, we develop order-preserving rule (OPR) mining and propose the OPR-Miner algorithm, which can mine the implicit relationships between OPPs. The main contributions of the paper are as follows.

- 1) To efficiently mine frequent OPPs, we propose an EFO-Miner algorithm, which employs four strategies: pattern fusion, support-based pattern fusion, screening, and pruning.
- 2) To mine the implicit relationships between OPPs, we propose the OPR-Miner algorithm based on EFO-Miner to discover strong rules.
- 3) Experimental results verify that OPR-Miner yields better performance than other competitive algorithms. More-

over, clustering and classification experiments validate that OPR-Miner can be used to realize feature extraction and achieve good performance.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 provides a definition of the problem. Section 4 proposes the OPR-Miner algorithm and presents an analysis of its time and space complexities. In Section 5, we validate the performance of OPR-Miner. Section 6 concludes this paper.

## 2. RELATED WORK

Sequential pattern mining [22] is an important topic in the field of data mining, whose aim is to mine the subsequences from a sequential dataset that users are interested in and to help people understand the data and make decisions by analyzing the potential patterns [23]. To solve different types of problems, sequential pattern mining has been extended to include a variety of mining methods, such as sequential pattern mining with gap constraints (or repetitive sequential pattern mining) [24], negative sequential pattern mining [25, 26], high utility pattern mining [27], high average-utility pattern mining [28, 29], episode mining [30, 31], and OPP mining for time series [21].

Various sequential pattern mining methods have been applied in many fields, such as disease prediction [32], virus sequence analysis [33], and network clickstream analysis [34]. For example, Duan et al. [35] used outlying sequence pattern mining to analyze the outliers in sequence data. Wu et al. [36] developed top- $k$  contrast pattern mining to realize the feature extraction of sequence classification. Smedt et al. [37] discovered patterns for sequence classification using behavioral constraint templates. Wu et al. [38] used a Nettee to calculate the support of a pattern under nonoverlapping conditions. Zhang et al. [39] proposed a sequential pattern mining method based on periodic gap constraints.

However, frequent pattern mining may ignore the implicit relationships within the transaction, and sequential rule mining [40] was proposed to address this problem. For example, Pham et al. [41] proposed an efficient method of mining sequential rules by constructing a prefix tree structure, which generated a large number of redundant rules in the process. Moreover, Fournier-Viger et al. [42] proposed a partially-ordered sequential rule mining to improve prediction accuracy.

Although the works described above have achieved good mining results, these studies mainly focused on the mining of discrete sequences, such as DNA or protein sequences. Due to the high continuity of time series, it is difficult to apply this approach to time series composed of ordered and continuous values. A classical way is that users employ the symbolization methods to discretize the original real values into symbols, and then apply the sequential pattern mining methods to find the interesting patterns. Typical symbolization algorithms include segmentation notation, represented by PAA [11], and symbolic representation, represented by SAX [12]. The main advantage of the time series symbolization method is that the time series is converted into a sequence of symbols through certain transformation rules, thus allowing traditional symbol sequence mining methods to be applied. However, various kinds of noise are inevitably introduced, due to the setting of various hard intervals in the process of converting time series into symbol series. In addition, these methods also ignore the original characteristics of the sequence, making it difficult to find the trends in the data.

To overcome the drawbacks of the symbolization methods, our previous work proposed the OPP mining method which does not need to symbolize the time series [21]. To effectively discover the frequent OPPs, OPP-Miner was proposed and employed an OPP matching method to calculate the supports. In terms of OPP matching, Kim et al. [18] employed the KMP algorithm to find subsequences with the same trend in a sequence. However, their approach did not consider the case of equal values, and Cho et al. [19] therefore designed a new algorithm to determine whether two time series were in the same order, even if some elements were equal. To further improve the matching efficiency, Chhabra and Tarhio [43] proposed a filtration method to find all occurrences.

However, OPP-Miner [21] has two drawbacks. Firstly, the efficiency of OPP-Miner can be further improved, since OPP-Miner adopts a pattern matching method to calculate pattern support, which does not use the calculation results of the sub-patterns. Secondly, OPP-Miner discovers all OPPs. Nevertheless, the implicit relationships between OPPs are not discovered. To overcome the drawbacks of OPP-Miner, this paper proposes the EFO-Miner algorithm, which utilizes the results from sub-patterns to calculate the support of super-patterns, in order to effectively avoid redundant calculations and improve the mining efficiency. More importantly, this paper further proposes the OPR-Miner algorithm based on the EFO-Miner algorithm to find strong rules which can discover the implicit relationships between OPPs, and can be used to extract time series features for clustering and classification.

### 3. Problem Definition

**Definition 1.** A time series is a numerical series of the same statistical indicator that is arranged in the order of its occurrence time, and is denoted as  $\mathbf{t} = (t_1, \dots, t_i, \dots, t_n)$ , where  $1 \leq i \leq n$ .

**Definition 2.** The rank of an element  $p_i$  in pattern  $\mathbf{p} = (p_1, \dots, p_i, \dots, p_m)$  ( $1 \leq i \leq m$ ) is denoted as  $\text{rank}_p(p_i)$ . A pattern represented by the relative order of the elements is called an OPP, and can be expressed as  $R(\mathbf{p}) = (\text{rank}_p(p_1), \text{rank}_p(p_2), \dots, \text{rank}_p(p_m))$ .

**Example 1.** Suppose we have a pattern  $\mathbf{p} = (31, 27, 33, 30)$ . We know that 31 is the third smallest value in  $\mathbf{p}$ , i.e.,  $\text{rank}(31) = 3$ . Similarly,  $\text{rank}(27) = 1$ . Thus, the OPP of  $\mathbf{p}$  is  $R(\mathbf{p}) = (3, 1, 4, 2)$ .

**Definition 3.** Suppose we have a pattern  $\mathbf{p} = (p_1, p_2, \dots, p_m)$  and a time series  $\mathbf{t} = (t_1, \dots, t_i, \dots, t_n)$ . If there exists a sub-time series  $\mathbf{t}' = (t_i, t_{i+1}, \dots, t_{i+m-1})$  ( $1 \leq i$  and  $i+m-1 \leq n$ ) which satisfies  $R(\mathbf{t}') = R(\mathbf{p})$ , then  $\mathbf{t}'$  is an occurrence of pattern  $\mathbf{p}$  in time series  $\mathbf{t}$ , and we use  $\langle i+m-1 \rangle$  to represent the occurrence. The support of  $\mathbf{p}$  in  $\mathbf{t}$  is the number of occurrences, denoted by  $\text{sup}(\mathbf{p}, \mathbf{t})$ .

**Definition 4.** Given a minimum support threshold  $\text{minsup}$ , if the support of  $\mathbf{p}$  in  $\mathbf{t}$  is no less than  $\text{minsup}$ , i.e.,  $\text{sup}(\mathbf{p}, \mathbf{t}) \geq \text{minsup}$ , then pattern  $\mathbf{p}$  is called a frequent OPP.

**Example 2.** Suppose we have a sequence  $\mathbf{t} = (24, 31, 27, 33, 30, 24, 21, 25, 23, 26, 22, 27, 24, 28, 23, 29)$ , as shown in Fig. 1, and a sub-time series  $(t_2, t_3, t_4, t_5) = (31, 27, 33, 30)$ . We know that  $\mathbf{p} = R(t_2, t_3, t_4, t_5) = (3, 1, 4, 2)$ . Similarly,  $R(t_6, t_7, t_8, t_9) = R(t_{10}, t_{11}, t_{12}, t_{13}) = (3, 1, 4, 2)$ . There are therefore three occurrences of pattern  $(3, 1, 4, 2)$  in  $\mathbf{t}$ , i.e.,  $\text{sup}(\mathbf{p}, \mathbf{t}) = 3$ . If  $\text{minsup} = 3$ , then pattern  $\mathbf{p}$  is a frequent OPP. In this way, we can get all frequent OPPs  $F = \{(1, 2), (2, 1), (1, 3, 2), (2, 1, 3), (1, 3, 2, 4), (3, 1, 4, 2)\}$ .

**Definition 5.** Given a pattern  $\mathbf{p} = (p_1, p_2, \dots, p_m)$ , the sub-time series  $\mathbf{e} = R(p_1, p_2, \dots, p_{m-1})$  is called the prefix OPP of  $\mathbf{p}$ , and is denoted as  $\mathbf{e} = \text{prefix}(\mathbf{p})$ . Sub-time series  $\mathbf{k} = R(p_2, p_3, \dots, p_m)$  is called the suffix OPP of  $\mathbf{p}$ , and is denoted as  $\mathbf{k} = \text{suffix}(\mathbf{p})$ , where  $\mathbf{e}$  and  $\mathbf{k}$  are the order-preserving sub-patterns of  $\mathbf{p}$ , and  $\mathbf{p}$  is the order-preserving super-pattern of  $\mathbf{e}$  and  $\mathbf{k}$ .

**Definition 6.** Suppose  $\mathbf{x}$  and  $\mathbf{y}$  are frequent OPPs. If  $\mathbf{x}$  is the prefix OPP of  $\mathbf{y}$ , then the implication  $\mathbf{x} \rightarrow \mathbf{y}$  is called an order-preserving rule, where  $\mathbf{x}$  is the antecedent of the rule, and  $\mathbf{y}$  is the consequent of the rule.

**Definition 7.** The confidence rate of  $\mathbf{x} \rightarrow \mathbf{y}$ , denoted as  $\text{conf}(\mathbf{x} \rightarrow \mathbf{y})$ , is the ratio of the support of  $\mathbf{y}$  to that of  $\mathbf{x}$ , i.e.,  $\text{conf}(\mathbf{x} \rightarrow \mathbf{y}) = \text{sup}(\mathbf{y}, \mathbf{t}) / \text{sup}(\mathbf{x}, \mathbf{t})$ .

**Definition 8.** Given a minimum confidence rate threshold  $\text{minconf}$ , if  $\text{conf}(\mathbf{x} \rightarrow \mathbf{y}) \geq \text{minconf}$ , then  $\mathbf{x} \rightarrow \mathbf{y}$  is called a strong OPR.

**Definition 9.** Our aim is to discover all strong OPRs in frequent OPPs according to  $\text{minconf}$ .

**Example 3.** In Example 2,  $(2, 1, 3)$  is the prefix OPP of  $(3, 1, 4, 2)$ . Both  $(2, 1, 3)$  and  $(3, 1, 4, 2)$  are frequent OPPs, and their supports are 4 and 3, respectively. Hence,  $\text{conf}((2, 1, 3) \rightarrow (3, 1, 4, 2)) = 3/4 = 0.75$ . If  $\text{minconf} = 0.7$ , then  $(2, 1, 3) \rightarrow (3, 1, 4, 2)$  is a strong OPR. Since  $\text{conf}((2, 1) \rightarrow (2, 1, 3)) = 4/8 = 0.5$ , which is less than  $\text{minconf}$ , it is not a strong OPR. The strong OPRs in Example 2 are  $R = \{(1, 2) \rightarrow (1, 3, 2), (2, 1, 3) \rightarrow (3, 1, 4, 2)\}$ .

### 4. Proposed algorithms

In OPR mining, the key issue is to discover frequent OPPs. In Section 4.1, we introduce the principle of pattern fusion to generate candidate patterns. We propose the methods of support calculation based on pattern fusion (SPF) and SPF-Pro in Sections 4.2 and 4.3, respectively. Section 4.4 illustrates the pruning strategy that is applied to further prune candidate patterns based on SPF-Pro. Section 4.5 presents EFO-Miner, which is used to mine frequent OPPs. Finally, Section 4.6 proposes OPR-Miner to discover strong rules.

#### 4.1. Generating candidate patterns

To reduce the number of candidate patterns, we adopt a pattern fusion method proposed in [21] to generate candidate patterns.

For  $\mathbf{p} = (p_1, p_2, \dots, p_m)$  and  $\mathbf{q} = (q_1, q_2, \dots, q_m)$ , where  $m$  is the length of the patterns, if  $R(\text{suffix}(\mathbf{p})) = R(\text{prefix}(\mathbf{q}))$ , then  $\mathbf{p}$  and  $\mathbf{q}$  can generate a super-pattern with length  $m+1$ . Two cases are given below:

**Case 1:** If  $p_1 \neq q_m$ , then  $\mathbf{p}$  and  $\mathbf{q}$  can generate one pattern  $\mathbf{r} = (r_1, r_2, \dots, r_{m+1})$ , denoted as  $\mathbf{r} = \mathbf{p} \oplus \mathbf{q}$ .

1. If  $p_1 < q_m$ , then  $r_1 = p_1$ . Moreover, if  $q_i < p_1$ , then  $r_{i+1} = q_i$ . Otherwise,  $r_{i+1} = q_i + 1 (1 < i \leq m)$ .

2. If  $p_1 > q_m$ , then  $r_1 = p_1 + 1$ . Moreover, if  $q_i < p_1$ , then  $r_{i+1} = q_i$ . Otherwise,  $r_{i+1} = q_i + 1 (1 < i \leq m)$ .

**Case 2:** If  $p_1 = q_m$ , then  $\mathbf{p}$  and  $\mathbf{q}$  can generate two patterns  $\mathbf{r} = (r_1, r_2, \dots, r_{m+1})$  and  $\mathbf{h} = (h_1, h_2, \dots, h_{m+1})$ , denoted as  $\mathbf{r}, \mathbf{h} = \mathbf{p} \oplus \mathbf{q}$ . For pattern  $\mathbf{r}$ ,  $r_1 = p_1$  and  $r_{m+1} = p_1 + 1$ . Moreover, if  $q_i < p_1$ , then  $r_{i+1} = q_i$ . If  $q_i > p_1$ , then  $r_{i+1} = q_i + 1 (1 < i < m)$ .

For pattern  $\mathbf{h}$ ,  $h_1 = p_1 + 1$  and  $h_{m+1} = p_1$ . Moreover, if  $q_i < p_1$ , then  $h_{i+1} = q_i$ . If  $q_i > p_1$ , then  $h_{i+1} = q_i + 1 (1 < i < m)$ .

Example 4 illustrates the principle.

**Example 4.** Suppose there are only two frequent patterns with length-3,  $\mathbf{p} = (2, 1, 3)$  and  $\mathbf{q} = (1, 3, 2)$ . Based on the two patterns, we show that different strategies can generate different number of candidate patterns with length-4. Table 1 shows the sets of candidate patterns generated by enumeration and pattern fusion. If the enumeration method is adopted, there are four cases for each pattern, i.e., we can insert 1, 2, 3, or 4 at the end, while maintaining the relative order of the pattern (2, 1, 3). Thus, we get four candidate patterns (3, 2, 4, 1), (3, 1, 4, 2), (2, 1, 4, 3), and (2, 1, 3, 4), respectively. There are therefore eight candidate patterns using the enumeration strategy, since there are two length-3 patterns.

However, there are three candidate patterns using the pattern fusion strategy. We take  $(2, 1, 3) \oplus (1, 3, 2)$  as an example. Since  $p_1 = q_3 = 2$ , according to Case 2, pattern fusion generates two candidate patterns,  $\mathbf{r}$  and  $\mathbf{h}$ . For pattern  $\mathbf{r}$ ,  $r_1 = p_1 = 2$  and  $r_4 = p_1 + 1 = 3$ . Since  $q_1 = 1 < 2$ ,  $r_2 = q_1 = 1$ , and since  $q_2 = 3 > 2$ ,  $r_3 = q_2 + 1 = 4$ . Hence, pattern  $\mathbf{r}$  is (2, 1, 4, 3). Similarly, pattern  $\mathbf{h}$  is (3, 1, 4, 2). Table 1 shows a comparison of candidate patterns for these two different strategies.

Table 1: Comparison of candidate patterns

Frequent pattern	Enumeration	Patterns	Pattern fusion
(2, 1, 3)	(3, 2, 4, 1), (3, 1, 4, 2) (2, 1, 4, 3), (2, 1, 3, 4)	$(2, 1, 3) \oplus (1, 3, 2)$	(2, 1, 4, 3), (3, 1, 4, 2)
(1, 3, 2)	(2, 4, 3, 1), (1, 4, 3, 2) (1, 4, 2, 3), (1, 3, 2, 4)	$(1, 3, 2) \oplus (2, 1, 3)$	(1, 3, 2, 4)

From Table 1, we can see that the pattern fusion strategy outperforms the enumeration strategy, since the pattern fusion strategy can prune many useless candidate patterns, thus improving the mining efficiency.

Although the pattern fusion strategy was proposed in [21], the correctness and completeness were not given in that paper. Now, we show the correctness and completeness as follows.

**Theorem 1.** Each candidate pattern is generated exact once and all candidate patterns can be generated, i.e., the pattern fusion strategy is correct and complete.

*Proof.* Firstly, we show that the OPP mining satisfies the anti-monotonicity, which means that support of super-pattern  $\mathbf{r}$  is less than that of its prefix pattern  $\mathbf{p}$  or suffix pattern  $\mathbf{q}$ . Suppose  $\langle a \rangle$  is an occurrence of super-pattern  $\mathbf{r}$ . We can safely say that  $\langle a - 1 \rangle$  is an occurrence of pattern  $\mathbf{p}$ , and  $\langle a \rangle$  is an occurrence of pattern  $\mathbf{q}$ . Therefore,  $\text{sup}(\mathbf{r}, \mathbf{t}) \leq \text{sup}(\mathbf{p}, \mathbf{t})$  and  $\text{sup}(\mathbf{r}, \mathbf{t}) \leq \text{sup}(\mathbf{q}, \mathbf{t})$ . Hence, the OPP mining satisfies the anti-monotonicity.

Secondly, we show that each candidate pattern can be generated only once. Proof by contradiction. Suppose super-pattern  $\mathbf{r}$  can be generated twice, and suppose  $\mathbf{r}$  is generated by two different prefix patterns. Suppose  $\mathbf{r} = (r_1, r_2, \dots, r_m, r_{m+1})$ . Thus, its prefix pattern is  $(r_1, r_2, \dots, r_m)$ . According to Definition 2, we know that the relative order of  $(r_1, r_2, \dots, r_m)$  is only one, i.e., the result of  $R(r_1, r_2, \dots, r_m)$  is an OPP, rather than two OPPs. This contradicts

the assumption that  $\mathbf{r}$  is generated by two different prefix patterns. Hence, each candidate pattern is generated exact once.

Finally, we show that all candidate patterns can be generated. Suppose super-pattern  $\mathbf{r} = (r_1, r_2, \dots, r_m, r_{m+1})$  is not generated, the prefix and suffix patterns of  $\mathbf{r}$  are  $\mathbf{p} = R(r_1, r_2, \dots, r_m)$  and  $\mathbf{q} = R(r_2, \dots, r_m, r_{m+1})$ , respectively. There are two cases: (1) pattern  $\mathbf{p}$  or  $\mathbf{q}$  is infrequent; (2) patterns  $\mathbf{p}$  and  $\mathbf{q}$  are frequent, but super-pattern  $\mathbf{r}$  cannot be generated by the pattern fusion strategy.

Case 1: Suppose pattern  $\mathbf{p}$  is infrequent, i.e.,  $sup(\mathbf{p}, \mathbf{t}) < minsup$ . Then, according to the anti-monotonicity,  $sup(\mathbf{r}, \mathbf{t}) < minsup$ . Thus, pattern  $\mathbf{r}$  is also infrequent. Hence, in this case, it is not necessary to generate super-pattern  $\mathbf{r}$ . Similarly, if pattern  $\mathbf{q}$  is infrequent, then it is not necessary to generate super-pattern  $\mathbf{r}$ , either.

Case 2: Proof by contradiction. Suppose super-pattern  $\mathbf{r} = (r_1, r_2, \dots, r_m)$  cannot be generated by  $\mathbf{p} \oplus \mathbf{q}$ . We know that  $R(suffix(\mathbf{p})) = R(prefix(\mathbf{q})) = R(r_2, \dots, r_m)$ . Therefore, we can generate super-pattern  $\mathbf{r} = \mathbf{p} \oplus \mathbf{q}$  according to the pattern fusion strategy, which contradicts the assumption that super-pattern  $\mathbf{r} = (r_1, r_2, \dots, r_m, r_{m+1})$  cannot be generated. Hence, all candidate patterns can be generated.  $\square$

For example, in Table 1, although patterns (3,2,4,1) and (2,1,3,4) cannot be generated by  $(2,1,3) \oplus (1,3,2)$ , they can be generated by  $(2,1,3) \oplus (2,3,1)$  and  $(2,1,3) \oplus (1,2,3)$ , respectively. This example shows that all patterns can be generated by using the pattern fusion strategy.

#### 4.2. SPF for support calculation

OPP-Miner adopts a pattern matching method to calculate pattern support, which does not use the calculation results of the sub-patterns [21]. If we can use the occurrences of subpatterns to generate the occurrences of super-patterns, then the new method can improve the efficiency, and is feasible. The reason is shown as follows. Suppose pattern  $\mathbf{r}$  is generated by patterns  $\mathbf{p}$  and  $\mathbf{q}$ , i.e.,  $\mathbf{r} = \mathbf{p} \oplus \mathbf{q}$ , and  $\langle x \rangle$  is an occurrence of pattern  $\mathbf{r}$ . We can safely say that  $\langle x - 1 \rangle$  and  $\langle x \rangle$  are occurrences of patterns  $\mathbf{p}$  and  $\mathbf{q}$ , respectively. Similarly, we know that if  $\langle x - 1 \rangle$  is not an occurrence of pattern  $\mathbf{p}$  or  $\langle x \rangle$  is not an occurrence of pattern  $\mathbf{q}$ , then  $\langle x \rangle$  is not an occurrence of pattern  $\mathbf{r}$ . An illustrative example is shown as follows.

For example, in Fig. 1, we know that the relative order of sub-time series  $(t_2, t_3, t_4, t_5)$  is (3,1,4,2), i.e.,  $\langle 5 \rangle$  is an occurrence of pattern (3,1,4,2). Therefore, the relative orders of sub-time series  $(t_2, t_3, t_4)$  and  $(t_3, t_4, t_5)$  are (2,1,3) and (1,3,2), respectively. Moreover, the relative orders of sub-time series  $(t_{12}, t_{13}, t_{14})$  is (2,1,3), but that of  $(t_{13}, t_{14}, t_{15})$  is not (1,3,2). Therefore,  $\langle 15 \rangle$  is not an occurrence of pattern (3,1,4,2). Hence, we propose an algorithm called SPF to calculate the support based on pattern fusion, which can use the occurrences of sub-patterns to generate the occurrences of super-patterns. The details are shown as follows.

From Section 4.1, super-patterns  $\mathbf{r}$  and  $\mathbf{h}$  are generated by  $\mathbf{p} \oplus \mathbf{q}$  which can be seen as the prefix and suffix patterns of the super-patterns, respectively. Suppose  $\langle lp_i \rangle$  and  $\langle lq_j \rangle$  are the occurrences of  $\mathbf{p}$  and  $\mathbf{q}$ , respectively. All occurrences of  $\mathbf{p}$  and  $\mathbf{q}$  are stored in a prefix array  $\mathcal{P}_p$  and a suffix array  $\mathcal{S}_q$ , respectively, i.e.,  $\langle lp_i \rangle \in \mathcal{P}_p$  and  $\langle lq_j \rangle \in \mathcal{S}_q$ . The matching results of super-patterns  $\mathbf{r}$  and  $\mathbf{h}$  are stored in  $L_r$  and  $L_h$ , respectively. This method is demonstrated as follows.

**Rule 1.** If  $p_1 \neq q_m$ , then  $\mathbf{r} = \mathbf{p} \oplus \mathbf{q}$ :

As shown in Fig. 2, if and only if  $lq_j = lp_i + 1$ , then  $\langle lq_j \rangle$  is an occurrence of  $\mathbf{r}$ , i.e.,  $lq_j \in L_r$ .

**Rule 2.** If  $p_1 = q_m$ , then  $\mathbf{r}, \mathbf{h} = \mathbf{p} \oplus \mathbf{q}$ :

As shown in Figure 3, if  $lq_j = lp_i + 1$ , then  $\langle lq_j \rangle$  may be an occurrence of  $\mathbf{r}$  or  $\mathbf{h}$ . It is necessary to determine  $t_{begin}$  and  $t_{end}$  in  $\mathbf{t}$ , where  $begin = lq_j - m$  and  $end = lq_j$ . There are three cases:

**Case 1:** If  $t_{begin} < t_{end}$ , then  $\langle lq_j \rangle$  is an occurrence of  $\mathbf{r}$ , i.e.,  $lq_j \in L_r$ .

**Case 2:** If  $t_{begin} > t_{end}$ , then  $\langle lq_j \rangle$  is an occurrence of  $\mathbf{h}$ , i.e.,  $lq_j \in L_h$ .

**Case 3:** If  $t_{begin} = t_{end}$ , then  $\langle lq_j \rangle$  is an occurrence of neither  $\mathbf{r}$  nor  $\mathbf{h}$ .

Finally, the size of sets  $L_r$  and  $L_h$  are the supports of the super-patterns  $\mathbf{r}$  and  $\mathbf{h}$ , respectively, i.e.,  $sup(\mathbf{r}) = |L_r|$  and  $sup(\mathbf{h}) = |L_h|$ . An illustration is given in Example 5.

**Example 5.** Suppose we have a time series  $\mathbf{t}$ , as shown in Table 2. The matching sets of length-2 patterns  $\mathbf{p} = (1,2)$  and  $\mathbf{q} = (2,1)$  are  $L_p = \{\langle 2 \rangle, \langle 4 \rangle, \langle 8 \rangle, \langle 10 \rangle, \langle 12 \rangle, \langle 14 \rangle, \langle 16 \rangle\}$  and  $L_q = \{\langle 3 \rangle, \langle 5 \rangle, \langle 6 \rangle, \langle 7 \rangle, \langle 9 \rangle, \langle 11 \rangle, \langle 13 \rangle, \langle 15 \rangle\}$ , respectively.

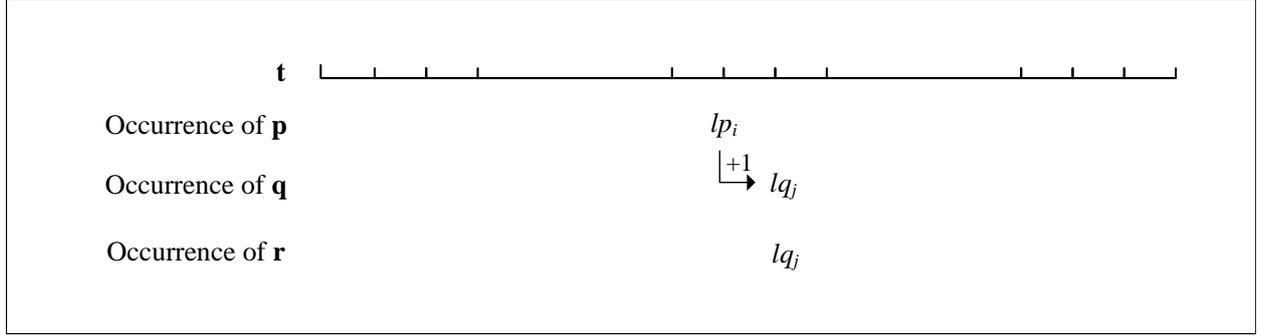


Figure 2: Occurrence of  $\mathbf{r}$  in sequence  $\mathbf{t}$

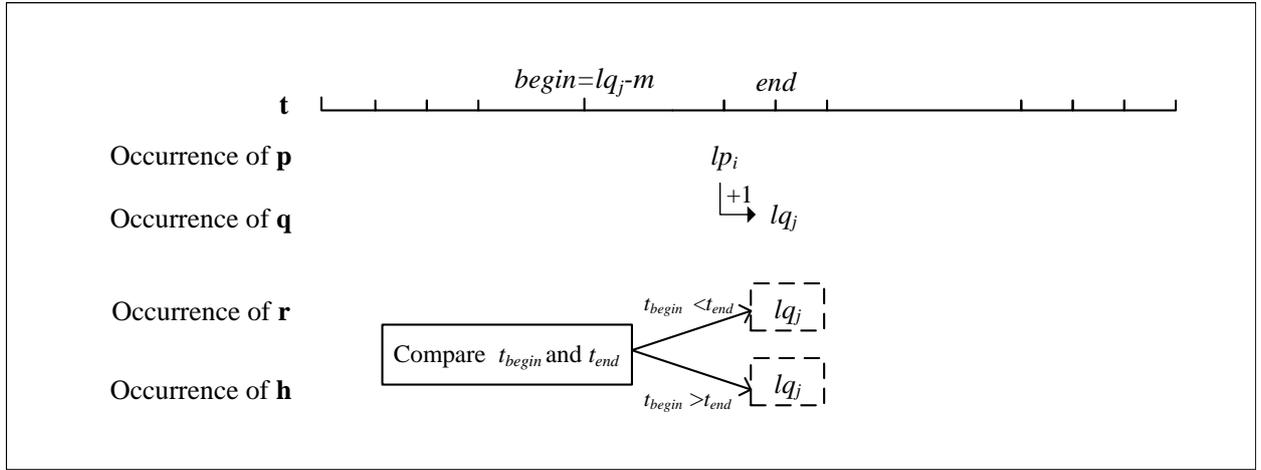


Figure 3: Occurrence of  $\mathbf{r}$  and  $\mathbf{h}$  in sequence  $\mathbf{t}$

Table 2: Time series element index

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$t$	24	31	27	33	30	24	21	25	23	26	22	27	24	28	23	29

$\mathbf{p} \oplus \mathbf{q}$  can generate two super-patterns,  $\mathbf{r} = (1,3,2)$  and  $\mathbf{h} = (2,3,1)$ . We know that  $\mathcal{P}_{\mathbf{p}} = L_{\mathbf{p}}$ , since  $\mathbf{p}$  is a prefix pattern. Similarly,  $\mathcal{S}_{\mathbf{q}} = L_{\mathbf{q}}$ . Moreover,  $2 \in \mathcal{P}_{\mathbf{p}}$  and  $2+1 = 3 \in \mathcal{S}_{\mathbf{q}}$ . Hence, according to Rule 2,  $\langle 3 \rangle$  may be an occurrence of  $\mathbf{r}$  or  $\mathbf{h}$ .  $begin = 3-2 = 1$  and  $end = 3$ . Thus,  $\langle 3 \rangle$  is one occurrence of  $\mathbf{r}$ , that is,  $\langle 3 \rangle \in L_{\mathbf{r}}$ , since  $t_1 = 24 < t_3 = 27$ . Similarly, we know that the matching set of  $\mathbf{r}$  is  $L_{\mathbf{r}} = \{\langle 3 \rangle, \langle 5 \rangle, \langle 9 \rangle, \langle 11 \rangle, \langle 13 \rangle, \langle 15 \rangle\}$  and  $sup(\mathbf{r}) = 6$ . The matching set of  $\mathbf{h}$  is  $L_{\mathbf{h}} = \{\langle 11 \rangle, \langle 15 \rangle\}$  and  $sup(\mathbf{h}) = 2$ .

#### 4.3. SPF-Pro for support calculation

In the SPF algorithm,  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$  are fixed. To further improve the efficiency of SPF, we propose a more efficient approach called SPF-Pro, in which  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$  are reduced dynamically, where the initial values of  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$  are  $L_{\mathbf{p}}$  and  $L_{\mathbf{q}}$ , respectively, i.e.,  $\mathcal{P}_{\mathbf{p}} = L_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}} = L_{\mathbf{q}}$ . This method is called a screening strategy.

**Screening strategy.** In Rules 1 and 2, if  $lp_i$  in  $\mathcal{P}_{\mathbf{p}}$  and  $lq_j$  in  $\mathcal{S}_{\mathbf{q}}$  are used to generate an occurrence of  $\mathbf{r}$  or  $\mathbf{h}$ , then  $lp_i$  and  $lq_j$  can be pruned. The new corresponding rules are shown below as Rules 3 and 4, respectively.

**Rule 3.** If  $p_1 \neq q_m$ , then  $\mathbf{r} = \mathbf{p} \oplus \mathbf{q}$ :

If and only if  $lq_j = lp_i + 1$ , then  $\langle lq_j \rangle$  is an occurrence of  $\mathbf{r}$ , i.e.,  $lq_j \in L_{\mathbf{r}}$ , and  $lp_i$  and  $lq_j$  are pruned from  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$ , respectively.

**Rule 4.** If  $p_1 = q_m$ , then  $\mathbf{r}, \mathbf{h} = \mathbf{p} \oplus \mathbf{q}$ :

If and only if  $lq_j = lp_i + 1$ ,  $\langle lq_j \rangle$  may be an occurrence of  $\mathbf{r}$  or  $\mathbf{h}$ . There are then three cases:

**Case 1:** If  $t_{begin} < t_{end}$ , then  $\langle lq_j \rangle$  is an occurrence of  $\mathbf{r}$ , i.e.,  $lq_j \in L_{\mathbf{r}}$ , and  $lp_i$  and  $lq_j$  are pruned from  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$ , respectively.

**Case 2:** If  $t_{begin} > t_{end}$ , then  $\langle lq_j \rangle$  is an occurrence of  $\mathbf{h}$ , i.e.,  $lq_j \in L_{\mathbf{h}}$ , and  $lp_i$  and  $lq_j$  are pruned from  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$ , respectively.

**Case 3:** If  $t_{begin} = t_{end}$ , then  $\langle lq_j \rangle$  is an occurrence of neither  $\mathbf{r}$  nor  $\mathbf{h}$ .

To prove the correctness of this screening strategy, we initially prove two theorems.

**Theorem 2.** Suppose  $\mathbf{p}$  can fuse with  $\mathbf{q}_1$  and  $\mathbf{q}_2$ , i.e.,  $\mathbf{r}_1, \mathbf{h}_1 = \mathbf{p} \oplus \mathbf{q}_1$  and  $\mathbf{r}_2, \mathbf{h}_2 = \mathbf{p} \oplus \mathbf{q}_2$ . If  $lp_i + 1 = x \in L_{\mathbf{r}_1}$  or  $L_{\mathbf{h}_1}$ , then  $x \notin L_{\mathbf{r}_2}$  or  $L_{\mathbf{h}_2}$ , and vice versa.

*Proof.* (Proof by contradiction) Suppose  $lp_i + 1 = x \in L_{\mathbf{r}_1}$  and  $x \in L_{\mathbf{r}_2}$ . Since  $x \in L_{\mathbf{r}_1}$ , we know that  $\langle x \rangle$  is an occurrence of  $\mathbf{r}_1$ . Similarly,  $\langle x \rangle$  is also an occurrence of  $\mathbf{r}_2$ . Obviously,  $\langle x \rangle$  cannot be two occurrences for two different patterns with the same length. Hence, this does not hold and the assumption is contradicted; that is,  $x \notin L_{\mathbf{r}_2}$ , and vice versa.  $\square$

**Theorem 3.** Suppose  $\mathbf{p}_1$  and  $\mathbf{p}_2$  can fuse with  $\mathbf{q}$ , i.e.,  $\mathbf{r}_1, \mathbf{h}_1 = \mathbf{p}_1 \oplus \mathbf{q}$  and  $\mathbf{r}_2, \mathbf{h}_2 = \mathbf{p}_2 \oplus \mathbf{q}$ . If  $lq_j = x \in L_{\mathbf{r}_1}$  or  $L_{\mathbf{h}_1}$ , then  $x \notin L_{\mathbf{r}_2}$  or  $L_{\mathbf{h}_2}$ , and vice versa.

*Proof.* The proof method is the same as for Theorem 2.  $\square$

**Theorem 4.** The screening strategy is correct.

*Proof.* According to Theorem 2,  $\langle lp_i \rangle$  belongs to only one pattern. Hence, if  $\langle lp_i \rangle$  is used to generate an occurrence of its super-pattern, then  $\langle lp_i \rangle$  can be pruned. Similarly, according to Theorem 3,  $\langle lq_j \rangle$  can also be pruned. We have therefore proved the correctness of the screening strategy.  $\square$

Example 6 is used to demonstrate that SPF-Pro outperforms SPF.

**Example 6.** We adopt the same data as in Example 5. We know that  $\mathbf{q}=(2,1)$ , and  $\mathbf{q}$  can fuse with  $\mathbf{q}$ , i.e.,  $\mathbf{e} = \mathbf{q} \oplus \mathbf{q} = (3,2,1)$ .  $\mathcal{P}_{\mathbf{q}} = \mathcal{S}_{\mathbf{q}} = L_{\mathbf{q}} = \{\langle 3 \rangle, \langle 5 \rangle, \langle 6 \rangle, \langle 7 \rangle, \langle 9 \rangle, \langle 11 \rangle, \langle 13 \rangle, \langle 15 \rangle\}$ . According to SPF, we know that  $L_{\mathbf{e}} = \{\langle 6 \rangle, \langle 7 \rangle\}$  and  $\text{sup}(\mathbf{e}) = 2$ .

We now show that SPF-Pro yields better performance than SPF. In Example 5, we know that the super-patterns  $\mathbf{r}$  and  $\mathbf{h}$  are generated. According to the screening strategy,  $\langle 3 \rangle$  is an occurrence of  $\mathbf{r} = (1,3,2)$ . Hence,  $3 \notin \mathcal{S}_{\mathbf{q}}$ , and  $3$  is pruned from  $\mathcal{S}_{\mathbf{q}}$ . Similarly, according to SPF-Pro, we know that  $\mathcal{S}_{\mathbf{q}} = \{\langle 6 \rangle, \langle 7 \rangle\}$ . SPF-Pro then uses  $\mathcal{P}_{\mathbf{q}} = \{\langle 3 \rangle, \langle 5 \rangle, \langle 6 \rangle, \langle 7 \rangle, \langle 9 \rangle, \langle 11 \rangle, \langle 13 \rangle, \langle 15 \rangle\}$  and  $\mathcal{S}_{\mathbf{q}} = \{\langle 6 \rangle, \langle 7 \rangle\}$  to calculate the support of  $\mathbf{e}$ . Moreover,  $L_{\mathbf{e}} = \{\langle 6 \rangle, \langle 7 \rangle\}$  and  $\text{sup}(\mathbf{e}) = 2$ , which are the same as for SPF. Now, we can see that in SPF,  $\mathcal{S}_{\mathbf{q}} = L_{\mathbf{q}} = \{\langle 3 \rangle, \langle 5 \rangle, \langle 6 \rangle, \langle 7 \rangle, \langle 9 \rangle, \langle 11 \rangle, \langle 13 \rangle, \langle 15 \rangle\}$ , while in SPF-Pro,  $\mathcal{S}_{\mathbf{q}} = \{\langle 6 \rangle, \langle 7 \rangle\}$ , with a size that is significantly smaller than in SPF. Hence, SPF-Pro outperforms SPF.

Pseudocode for SPF-Pro is given in Algorithm 1, which calculates the supports of the super-patterns using the pattern fusion strategy.

#### 4.4. Pruning candidate patterns

In this section, we propose a pruning strategy to further prune candidate patterns based on SPF-Pro.

**Pruning strategy.** If  $|\mathcal{P}_{\mathbf{p}}| < \text{minsup}$ , then  $\mathbf{p}$  as a prefix pattern will no longer generate frequent patterns. If  $|\mathcal{S}_{\mathbf{q}}| < \text{minsup}$ , then  $\mathbf{q}$  as a suffix pattern will no longer generate frequent patterns.

**Theorem 5.** The pruning strategy is correct.

*Proof.* Suppose pattern  $\mathbf{p}$  can fuse with pattern  $\mathbf{q}$ , i.e.,  $\mathbf{r}, \mathbf{h} = \mathbf{p} \oplus \mathbf{q}$ . Obviously, the sizes of  $L_{\mathbf{r}}$  and  $L_{\mathbf{h}}$  are not greater than the size of  $\mathcal{P}_{\mathbf{p}}$  or  $\mathcal{S}_{\mathbf{q}}$ , since according to SPF-Pro, if and only if  $lq_j = lp_i + 1$  ( $lp_i \in \mathcal{P}_{\mathbf{p}}$ ,  $lq_j \in \mathcal{S}_{\mathbf{q}}$ ),  $lq_j \in L_{\mathbf{r}}$ . Thus,  $|L_{\mathbf{r}}| \leq |\mathcal{P}_{\mathbf{p}}|$ . Therefore,  $|L_{\mathbf{r}}| < \text{minsup}$ , since  $|\mathcal{P}_{\mathbf{p}}| < \text{minsup}$ . Hence,  $\mathbf{p}$  as a prefix pattern will no longer generate frequent patterns. Similarly, we can prove that  $\mathbf{q}$  as a suffix pattern will no longer generate frequent patterns.  $\square$

---

**Algorithm 1** SPF-Pro

---

**Input:** Pattern  $\mathbf{p}$  and its matching result  $\mathcal{P}_{\mathbf{p}}$ , pattern  $\mathbf{q}$  and its matching result  $\mathcal{S}_{\mathbf{q}}$  **Output:** Super-patterns and their matching results, and  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$

```
1:  $L_{\mathbf{r}} = \{\}; L_{\mathbf{h}} = \{\};$ 
2:  $\mathbf{e} \leftarrow R(\text{prefix}(\mathbf{p}));$ 
3:  $\mathbf{k} \leftarrow R(\text{suffix}(\mathbf{q}));$ 
4: if  $\mathbf{k} == \mathbf{e}$  then
5:   if  $\mathbf{p}[0] == \mathbf{q}[m-1]$  then
6:      $\mathbf{r} \cup \mathbf{h} \leftarrow \mathbf{p} \oplus \mathbf{q};$ 
7:     Calculate  $L_{\mathbf{r}}$  and  $L_{\mathbf{h}}$ , and update  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$  according to Rule 4;
8:   else
9:      $\mathbf{r} \leftarrow \mathbf{p} \oplus \mathbf{q};$ 
10:    Calculate  $L_{\mathbf{r}}$ , and update  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$  according to Rule 3;
11:   end if
12: end if
13: Return  $\mathbf{r}$ ,  $\mathbf{h}$ ,  $\mathcal{P}_{\mathbf{p}}$  and  $\mathcal{S}_{\mathbf{q}}$ ;
```

---

Example 7 illustrates the effectiveness of pruning strategy.

**Example 7.** We use the same data as in Example 6. We know that  $\mathbf{p} = (1,2)$  and  $\mathbf{q} = (2,1)$ . According to Rule 4, after two patterns  $\mathbf{r} = (1,3,2)$  and  $\mathbf{h} = (2,3,1)$  are generated by  $\mathbf{p} \oplus \mathbf{q}$ , we know that  $\mathcal{S}_{\mathbf{q}} = \{\langle 6 \rangle, \langle 7 \rangle\}$ . Suppose  $\text{minsup} = 3$ . If we do not apply the pruning strategy, according to Example 6, we have to use  $\mathcal{P}_{\mathbf{q}} = \{\langle 3 \rangle, \langle 5 \rangle, \langle 6 \rangle, \langle 7 \rangle, \langle 9 \rangle, \langle 11 \rangle, \langle 13 \rangle, \langle 15 \rangle\}$  and  $\mathcal{S}_{\mathbf{q}} = \{\langle 6 \rangle, \langle 7 \rangle\}$  to calculate the support of  $\mathbf{e} = \mathbf{q} \oplus \mathbf{q} = (3,2,1)$ . We know that  $L_{\mathbf{e}} = \{\langle 6 \rangle, \langle 7 \rangle\}$  and  $\text{sup}(\mathbf{e}) = 2$ , and pattern  $\mathbf{e}$  is not a frequent pattern. However, according to the pruning strategy, we do not need to use  $\mathcal{P}_{\mathbf{q}}$  and  $\mathcal{S}_{\mathbf{q}} = \{\langle 6 \rangle, \langle 7 \rangle\}$  to calculate the support of  $\mathbf{e}$ , since  $|\mathcal{S}_{\mathbf{q}}| = 2 < \text{minsup}$ . Hence, we can avoid calculating  $\mathbf{q} \oplus \mathbf{q}$  using this approach.

#### 4.5. Mining OPPs

In this section, we introduce the EFO-Miner algorithm to discover frequent OPPs.

The steps of EFO-Miner are as follows.

Step 1: Scan the time series  $\mathbf{t}$  to calculate the matching results and the supports of patterns  $(1,2)$  and  $(2,1)$ . If the pattern is frequent, then it is stored into the frequent pattern set  $F_2$ ;

Step 2: Select any two patterns  $\mathbf{p}$  and  $\mathbf{q}$  in  $F_m$ . If pattern  $\mathbf{p}$  can fuse with pattern  $\mathbf{q}$ , then  $\mathbf{p} \oplus \mathbf{q}$  can generate candidate super-patterns  $\mathbf{r}$  and  $\mathbf{h}$ . If  $|\mathcal{P}_{\mathbf{p}}| \geq \text{minsup}$  and  $|\mathcal{S}_{\mathbf{q}}| \geq \text{minsup}$ , then use SPF-Pro to calculate the matching results and the supports of super-patterns  $\mathbf{r}$  and  $\mathbf{h}$ . If  $\mathbf{r}$  or  $\mathbf{h}$  is frequent, store it in the set  $F_{m+1}$ ;

Step 3: Iterate Step 2 until no  $(m+1)$ -length super-pattern is generated;

Step 4: Iterate Steps 2 and 3 until  $F_{m+1}$  is empty.

Finally, all frequent patterns  $F = F_2 \cup F_3 \cup \dots \cup F_m$ .

Example 8 illustrates the principle of EFO-Miner.

**Example 8.** We use the same data as in Example 5. Suppose  $\text{minsup} = 3$ . We can discover all frequent patterns as follows.

First, the matching sets of length-2 patterns  $\mathbf{p} = (1,2)$  and  $\mathbf{q} = (2,1)$  are  $L_{\mathbf{p}} = \{\langle 2 \rangle, \langle 4 \rangle, \langle 8 \rangle, \langle 10 \rangle, \langle 12 \rangle, \langle 14 \rangle, \langle 16 \rangle\}$  and  $L_{\mathbf{q}} = \{\langle 3 \rangle, \langle 5 \rangle, \langle 6 \rangle, \langle 7 \rangle, \langle 9 \rangle, \langle 11 \rangle, \langle 13 \rangle, \langle 15 \rangle\}$ , respectively. Therefore,  $\mathcal{P}_{\mathbf{p}} = \mathcal{S}_{\mathbf{p}} = L_{\mathbf{p}} = \{\langle 2 \rangle, \langle 4 \rangle, \langle 8 \rangle, \langle 10 \rangle, \langle 12 \rangle, \langle 14 \rangle, \langle 16 \rangle\}$  and  $\mathcal{P}_{\mathbf{q}} = \mathcal{S}_{\mathbf{q}} = L_{\mathbf{q}} = \{\langle 3 \rangle, \langle 5 \rangle, \langle 6 \rangle, \langle 7 \rangle, \langle 9 \rangle, \langle 11 \rangle, \langle 13 \rangle, \langle 15 \rangle\}$ . Since  $\text{sup}(\mathbf{p}) = 7$  and  $\text{sup}(\mathbf{q}) = 8$ , we know that  $F_2 = \{(1,2), (2,1)\}$ .

EFO-Miner now finds frequent patterns with length three.  $\mathbf{p} \oplus \mathbf{p} = (1,2) \oplus (1,2) = (1,2,3)$ . According to SPF-Pro,  $\text{sup}(1,2,3) = 0$ , and  $\mathcal{P}_{\mathbf{q}}$  and  $\mathcal{S}_{\mathbf{q}}$  are not changed. Similarly,  $\mathbf{p} \oplus \mathbf{q}$  generates two candidate patterns,  $(1,3,2)$  and  $(2,3,1)$ . SPF-Pro calculates  $L_{(1,3,2)} = \{\langle 3 \rangle, \langle 5 \rangle, \langle 9 \rangle, \langle 11 \rangle, \langle 13 \rangle, \langle 15 \rangle\}$ ,  $\text{sup}(1,3,2) = 6$  and  $\text{sup}(2,3,1) = 0$ . Meanwhile,  $\mathcal{P}_{\mathbf{p}} = \{16\}$  and  $\mathcal{S}_{\mathbf{q}} = \{\langle 6 \rangle, \langle 7 \rangle\}$ . Thus,  $(1,3,2)$  is a frequent pattern. When the pruning strategy is used,  $\mathbf{p}$  as a prefix pattern and  $\mathbf{q}$  as a suffix pattern will no longer generate frequent patterns. In a similar way,  $(2,1,3)$  can be found.

Hence, the length-3 frequent pattern set  $F_3 = \{(1,3,2), (2,1,3)\}$  is obtained. Moreover, length-4 frequent patterns can be calculated based on  $F_3$ . Finally, we get the frequent pattern set  $F = \{(1,2), (2,1), (1,3,2), (2,1,3), (1,3,2,4), (3,1,4,2)\}$ .

Pseudocode for EFO-Miner is given in Algorithm 2.

---

**Algorithm 2** EFO-Miner

---

**Input:** Time series  $\mathbf{t}$  and the minimum support threshold  $minsup$  **Output:** Frequent pattern set  $F$

```

1: Scan sequence  $\mathbf{t}$ , and use  $L_{(1,2)}$  and  $L_{(2,1)}$  to store the matching sets of (1,2) and (2,1), respectively. If the size of
    $L_{(1,2)}$  is no less than  $minsup$ , then add pattern (1,2) to  $F_2$ . Similarly, add pattern (2,1) to  $F_2$ ;
2:  $m \leftarrow 2$ ;
3: while  $F_m \neq \text{NULL}$  do
4:   for each  $\mathbf{p}$  in  $F_m$  do
5:      $\mathcal{P}_p \leftarrow \mathcal{S}_p \leftarrow L_p$ ;
6:   end for
7:   for each  $\mathbf{p}$  in  $F_m$  do
8:     for each  $\mathbf{q}$  in  $F_m$  do
9:       if  $\mathcal{P}_p.size \geq minsup \ \&\& \ \mathcal{S}_q.size \geq minsup$  then
10:        if  $\mathbf{p}$  can fuse with  $\mathbf{q}$  then
11:          Calculate the matching results of super-patterns  $\mathbf{r}$  and  $\mathbf{h}$  and update  $\mathcal{P}_p$  and  $\mathcal{S}_q$  using SPF-Pro;
12:          if  $L_r.size \geq minsup$  then
13:            Add pattern  $\mathbf{r}$  to  $F_{m+1}$ ;
14:          end if
15:          if  $L_h.size \geq minsup$  then
16:            Add pattern  $\mathbf{h}$  to  $F_{m+1}$ ;
17:          end if
18:        end if
19:      end if
20:    end for
21:  end for
22:   $m \leftarrow m+1$ 
23: end while
24: Return  $F$ ;

```

---

**Theorem 6.** *EFO-Miner is correct and complete.*

*Proof.* We know that EFO-Miner employs the pattern fusion strategy to generate candidate patterns, the screening strategy to calculate the supports of candidate patterns, and the pruning strategy to further prune candidate patterns. Theorems 1, 4, and 5 show the correctness and completeness of these strategies. Therefore, EFO-Miner is correct and complete.  $\square$

**Theorem 7.** *The space and time complexity of EFO-Miner are  $O(f \times n)$ , where  $f$  and  $n$  are the number of frequent patterns and the sequence length, respectively.*

*Proof.* The space complexity of EFO-Miner involves two parts: the frequent patterns and the matching results. Since the number of frequent patterns is  $f$ , the space complexity of frequent patterns is  $O(f \times m)$ , where  $m$  is the length of the longest pattern. For each pattern  $\mathbf{p}$ , the space complexity of the matching results is  $O(n)$ . Similarly, the space complexities of  $\mathcal{P}_p$  and  $\mathcal{S}_p$  are also  $O(n)$ . Since there are  $f$  frequent patterns, the space complexity of the matching results is  $O(f \times n)$ . Since  $m$  is far less than  $n$ , the space complexity of EFO-Miner is  $O(f \times (m + n)) = O(f \times n)$ . The time complexity of calculating the matching results for each pattern is  $O(n)$ . There are  $f$  patterns. Therefore, the time complexity of EFO-Miner is  $O(f \times n)$ .  $\square$

#### 4.6. Mining strong OPRs

In this section, we explore the use of OPR-Miner to mine strong OPRs from all frequent patterns using EFO-Miner.

A simple method is that we enumerate all OPRs according to Definition 6 and calculate their confidences. If the confidence is no less than the threshold, then the rule is a strong OPR. Obviously, this method is not efficient.

According to Algorithm 2, we know that pattern  $\mathbf{p}$  is the prefix OPP of patterns  $\mathbf{r}$  and  $\mathbf{h}$ . Therefore, we can discover the strong OPRs in the process of mining frequent OPPs. It means that if the support of pattern  $\mathbf{r}$  is no less than  $minsup/minconf$ , then  $\mathbf{p} \rightarrow \mathbf{r}$  is a strong OPR. Similarly,  $\mathbf{p} \rightarrow \mathbf{h}$  is a strong OPR. More importantly, this method has the same time and space complexities as those of EFO-Miner. Pseudocode for OPR-Miner is shown in Algorithm 3.

---

#### Algorithm 3 OPR-Miner

---

**Input:** Time series  $\mathbf{t}$ , frequent pattern set  $F$ , support of each frequent pattern  $sup$ , and the minimum confidence threshold  $minconf$  **Output:** Strong OPR set  $R$

- 1: **if**  $L_{\mathbf{r}}.size/L_{\mathbf{p}}.size \geq minconf$  **then** // Add these codes after Line 17 in Algorithm 2.
  - 2:     Add rule  $\mathbf{p} \rightarrow \mathbf{r}$  to  $R$ ;
  - 3: **end if**
  - 4: **if**  $L_{\mathbf{h}}.size/L_{\mathbf{p}}.size \geq minconf$  **then**
  - 5:     Add rule  $\mathbf{p} \rightarrow \mathbf{r}$  to  $R$ ;
  - 6: **end if**
- 

According to Algorithm 3, we know that OPR-Miner does not employ any strategy, only uses Definitions 7 and 8 to discover strong OPRs based on EFO-Miner. Theorem 6 shows that EFO-Miner is correct and complete. Therefore, OPR-Miner is also correct and complete.

Moreover, Example 9 illustrates the difference between all OPRs and strong OPRs.

**Example 9.** This example uses the frequent OPPs in Example 8. We know that  $(1,2)$  and  $(1,3,2)$  are two frequent patterns, where  $(1,2)$  is the prefix pattern of  $(1,3,2)$ . According to Definition 6,  $(1,2) \rightarrow (1,3,2)$  is an OPR. Similarly, we find all OPRs:  $(1,2) \rightarrow (1,3,2)$ ,  $(2,1) \rightarrow (2,1,3)$ ,  $(1,3,2) \rightarrow (1,3,2,4)$ , and  $(2,1,3) \rightarrow (3,1,4,2)$ .

However, according to Definition 7, the confidence of rule  $(2,1) \rightarrow (2,1,3)$  is  $conf((2,1) \rightarrow (2,1,3)) = sup(2,1,3)/sup(2,1) = 4/8 = 0.5$ . Since rules with low confidence have no practical meaning in most applications, we only discover the strong OPRs, that is, those for which the confidence level is higher than a certain threshold. For example, suppose the minimum confidence threshold  $minconf$  is 0.7. Thus, rule  $(2,1) \rightarrow (2,1,3)$  is not a strong OPR, since its confidence is 0.5. We know that  $sup((1,2)) = 7$  and  $sup(1,3,2) = 6$ . Hence, the confidence of rule  $(1,2) \rightarrow (1,3,2)$  is  $conf((1,2) \rightarrow (1,3,2)) = 6/7$ , which is greater than  $minconf$ , and rule  $(1,2) \rightarrow (1,3,2)$  is therefore a strong OPR. Similarly, we get the strong OPR set  $R = \{(1,2) \rightarrow (1,3,2), (2,1,3) \rightarrow (3,1,4,2)\}$ .

This example shows that the number of strong OPRs is less than that of all OPRs, since the confidence of a strong OPR is no less than  $minconf$ , while the general OPRs do not have such constraints.

## 5. Experimental results and analysis

Section 5.1 introduces the benchmark datasets and the baseline methods. Section 5.2 validates the running performance of EFO-Miner. Section 5.3 shows the scalability of EFO-Miner. Section 5.4 reports the influence of different  $minsup$ . Section 5.5 verifies the performance of OPR-Miner. Section 5.6 shows the the influence of different  $minconf$ . Section 5.7 demonstrates the advantages of OPR patterns.

### 5.1. Benchmark datasets and baseline methods

We use real stock, weather, and oil datasets as test sequences. The stock and oil datasets can be downloaded from <https://www.yahoo.com/>, the weather datasets can be downloaded from <https://archive.ics.uci.edu/ml/datasets.php/>, the daily new cases datasets can be downloaded from <https://coronavirus.jhu.edu/>, the sensor and spectro datasets can be downloaded from <http://www.timeseriesclassification.com/index.php/>, and the diagnosis fault datasets can

Table 3: Description of datasets

Name	Dataset	Type	Total length	Number of sequences	Number of labels
SDB1	Italian-temperature	Weather	256	1	/
SDB2	Italian-temperature	Weather	1,233	1	/
SDB3	1WTI-2	Oil	2,496	1	/
SDB4	Crude Oil	Oil	4,954	1	/
SDB5	Russell 2000	Stock	8,141	1	/
SDB6	Nasdaq	Stock	12,279	1	/
SDB7	S&P 500	Stock	23,046	1	/
SDB8	PRSA_Data_Nongzhanguan	Weather	34,436	1	/
SDB9	CSSE COVID19 Dataset	Daily new cases	2,715	15	15
SDB10	Car	Sensor	8,655	15	4
SDB11	Meat	Spectro	6,345	15	3
SDB12	Beef	Spectro	7,050	15	5
SDB13	Bearing fault-NR	Diagnosis fault	46,024	44	2
SDB14	Bearing fault-NI	Diagnosis fault	46,024	44	2
SDB15	Bearing fault-NO	Diagnosis fault	46,024	44	2
SDB16	New York Stock Exchange	Stock	60,000	1	/

Note: SDB13-SDB15 are part of the sequences selected from the bearing fault dataset, which records the bearing fault vibration signals. There are four bearing State labels representing different States. Normal, Inner, Outer, and Roller. Among them, 22 Normal and 22 Roller sequences are extracted from SDB13, 22 Normal and 22 Inner sequences are extracted from SDB14, and 22 Normal and 22 outer sequences are extracted from SDB15.

be downloaded from <http://jzw.ie.tsinghua.edu.cn/Show/index/cid/45/id/1568.html/>. A specific description of each dataset is given in Table 3.

All experiments were run on a computer with Intel(R) Core(TM) i5-3230U, 1.60 GHz CPU, 8.0 GB RAM, and a Win10 64-bit operating system, and the compilation environment was Dev C++ 5.4.0.

This paper proposes EFO-Miner and OPR-Miner to mine frequent OPPs and strong OPRs, respectively. OPR-Miner adds two branch statements on the basis of EFO-Miner, which hardly takes time. Therefore, we only validate the running performance of EFO-Miner, since the running performance of OPR-Miner is almost the same as EFO-Miner. Moreover, we verify the usefulness of strong OPRs mined by OPR-Miner.

#### For EFO-Miner:

1) Mat-Based: To verify the efficiency of EFO-Miner, we developed Mat-Based which employs the pattern fusion strategy to generate candidate patterns and adopts an OPP matching algorithm proposed in [18] to calculate the support for each candidate pattern.

2) OPP-Miner [21]: To validate the efficiency of EFO-Miner, we selected OPP-Miner as a competitive algorithm. OPP-Miner adopts a pattern matching strategy to calculate the support and needs to scan the sequence numerous times.

3) EFO-enum: To test the performance of the pattern fusion strategy in terms of generating super-patterns, we developed EFO-enum, which employs an enumeration strategy to generate super-patterns and SFP to calculate the support.

4) EFO-scrn: To verify the effect of the screening strategy on the calculation of supports, we developed EFO-scrn, which does not apply the screening strategy. Since the pruning strategy is based on the screening strategy, EFO-scrn employs neither pruning strategy nor screening strategy, and instead adopts pattern fusion to generate candidate patterns and SFP to calculate the support.

5) EFO-prun: To validate the performance of the pruning strategy, we proposed EFO-prun, which does not apply the pruning strategy, and instead adopts pattern fusion to generate candidate patterns and SFP-Pro to calculate the support.

**For OPR-Miner:** 6) OPR-Rule: To report the confidences of the strong rules mined by OPR-Miner, we explored

OPR-Rule to generate all OPRs based on all frequent OPPs.

### 5.2. Performance of EFO-Miner

To validate the performance of EFO-Miner, we used five competitive algorithms: Mat-Based, OPP-Miner, EFO-enum, EFO-scrn, and EFO-prun. We performed experiments on the SDB1–SDB8 datasets, and set the minimum support threshold  $minsup = 12$ . Since all six algorithms are complete, the mining results are the same, i.e., there are 17, 72, 160, 297, 497, 741, 1162, and 1023 frequent patterns for SDB1–SDB8, respectively. Comparisons of the running time and numbers of candidate patterns are shown in Figs. 4 and 5, respectively. We also show a comparison of the numbers of elements in the prefix and suffix arrays in Fig. 6 (this figure does not include both Mat-Based and OPP-Miner, since the two algorithms do not use prefix and suffix arrays to calculate the support).

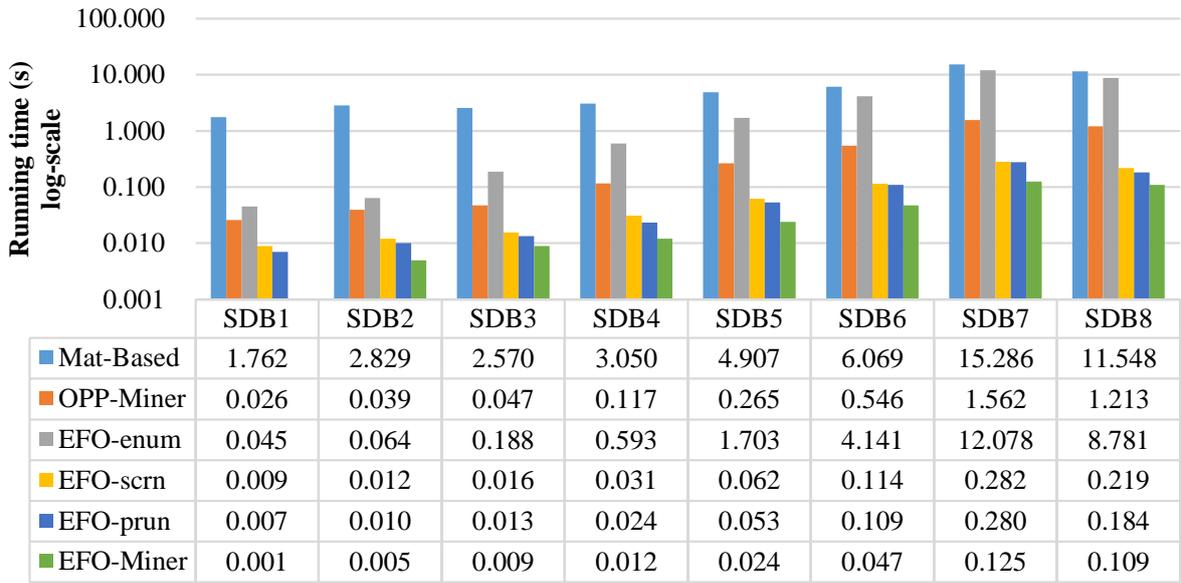


Figure 4: Comparison of running time on SDB1–SDB8

The results give rise to the following observations.

1) EFO-Miner gives better performance than both Mat-Based and OPP-Miner, since EFO-Miner not only runs faster than the two algorithms, but also checks fewer candidate patterns. For example, on SDB7, Fig. 4 shows that EFO-Miner takes 625 ms, while Mat-Based and OPP-Miner take 10,421 and 1,266 ms, respectively; Fig. 5 shows that EFO-Miner checks 2,838 candidate patterns, while Mat-Based and OPP-Miner check both 4,030. The same effect can be found on all the other datasets. The reasons for this are as follows. Mat-Based and OPP-Miner employ different pattern matching strategies that cannot use the results for the sub-patterns and has to scan the database repeatedly, which is inefficient. In contrast, EFO-Miner uses the results for the sub-patterns to calculate the occurrences of super-patterns, which can avoid redundant calculations and improve the efficiency. Moreover, although EFO-Miner, Mat-Based, and OPP-Miner adopt a pattern fusion strategy to generate candidate patterns, EFO-Miner employs a pruning strategy that can further reduce the number of candidate patterns. Hence, EFO-Miner checks fewer candidate patterns than both Mat-Based and OPP-Miner, and therefore outperforms them.

2) EFO-Miner outperforms EFO-enum, thus demonstrating that the pattern fusion strategy can efficiently prune candidate patterns. Fig. 4 shows that EFO-Miner runs faster than EFO-enum. For example, on SDB4, EFO-Miner takes 65.6 ms, while EFO-enum takes 2,594 ms. The same effect can be found on all the other datasets. The reason for this is that the pattern fusion strategy can effectively reduce the number of candidate patterns. For example, Fig. 5 shows that on SDB4, EFO-Miner generates 707 candidate patterns, while EFO-enum generates 1,992. From Fig. 6, we can see that on SDB4, EFO-Miner carries out 68,828 comparisons between elements, while for EFO-enum it

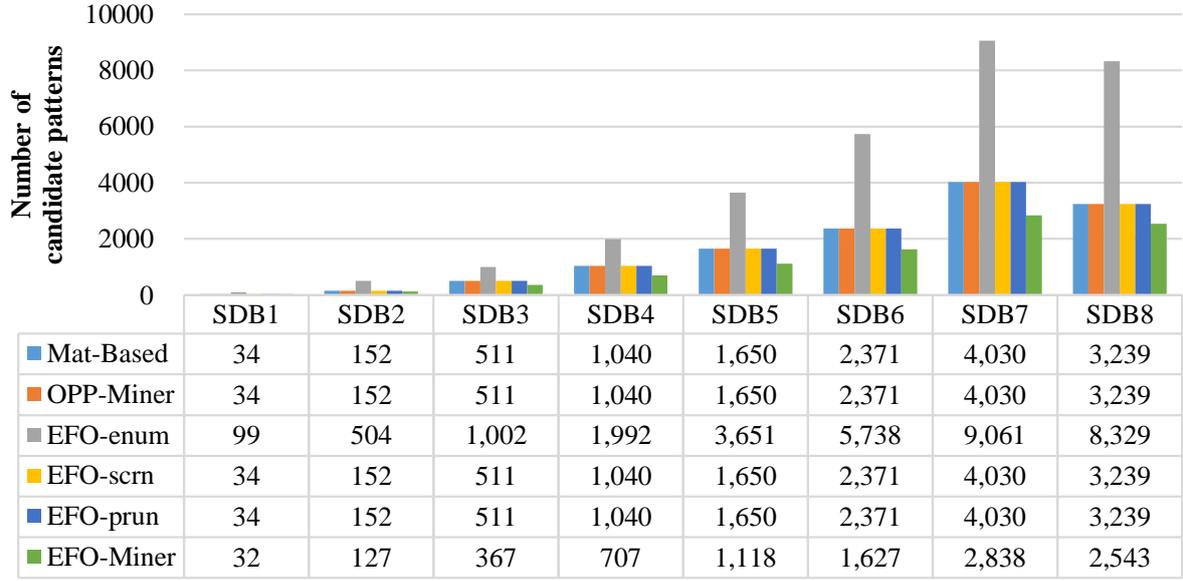


Figure 5: Comparison of numbers of candidate patterns for SDB1-SDB8

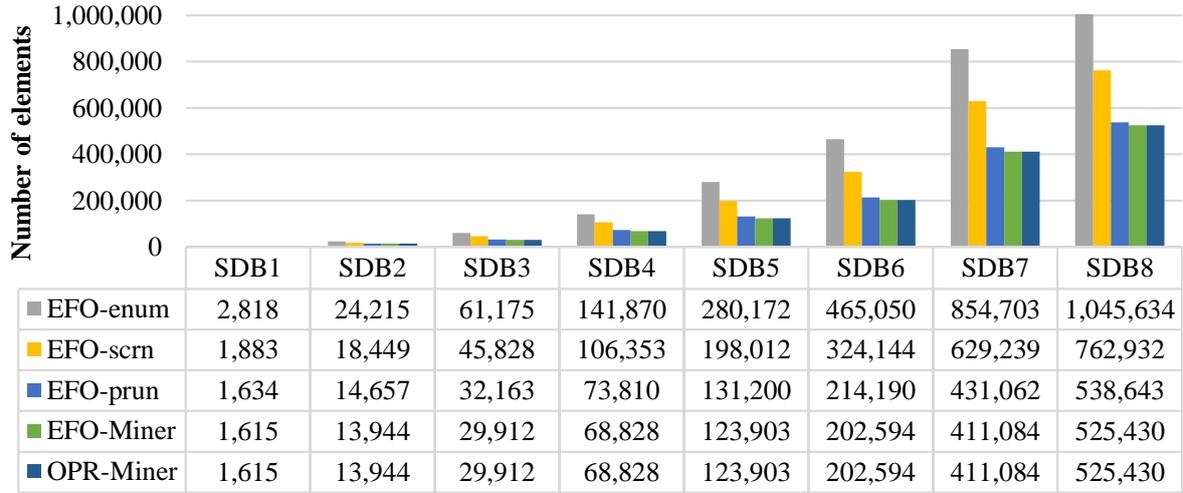


Figure 6: Comparison of numbers of elements in prefix and suffix arrays for SDB1-SDB8

is 141,870. The experimental results are therefore consistent with those in Example 4. We know that the lower the number of candidate patterns, the faster the algorithm runs. Hence, EFO-Miner runs faster than EFO-enum.

3) EFO-Miner outperforms EFO-prun, which indicates that the pruning strategy can efficiently reduce the number of candidate patterns. Fig. 4 shows that EFO-Miner runs faster than EFO-prun. For example, on SDB5, EFO-Miner takes 125 ms, while EFO-prun takes 234 ms, and the same effect can be found on the other datasets. The reason for this is that the pruning strategy can effectively reduce the number of candidate patterns. For example, from Fig. 5, we can see that EFO-Miner checks 1,118 candidate patterns for SDB5, while EFO-prun checks 1,650. From Figure 6, we see that on SDB5, EFO-Miner carries out 123,903 comparisons between elements, while for EFO-scrn it is 131,200. With a reduction in the number of candidate patterns, the number of comparisons is also reduced. The experimental results are therefore consistent with those in Example 7. We know that the lower the number of candidate patterns,

the faster the algorithm runs. EFO-Miner therefore runs faster than EFO-prun.

4) EFO-Miner outperforms EFO-scrn. More importantly, EFO-prun outperforms EFO-scrn, which indicates that the screening strategy can efficiently improve the mining performance. Fig. 4 shows that EFO-prun runs faster than EFO-scrn. For example, on SDB3, EFO-prun takes 37.3 ms, while EFO-scrn takes 44.8 ms, and the same effect can be found on all the other datasets. The reason for this is that the screening strategy can dynamically reduce the size of the prefix and suffix arrays. For example, Fig. 6 shows that on SDB3, EFO-prun carries out 32,163 comparisons between elements, while EFO-scrn carries out 45,828. The experimental results are therefore consistent with those in Example 6. The lower the sizes of the prefix and suffix arrays, the faster the algorithm runs, meaning that EFO-prun runs faster than EFO-scrn. We know that EFO-Miner runs faster than EFO-prun. Hence, EFO-Miner runs faster than EFO-scrn.

### 5.3. Scalability

In this section, to evaluate the scalability of EFO-Miner, we employed Mat-Based, EFO-enum, EFO-scrn, and EFO-prun as competitive algorithms. Moreover, we selected SDB8 as the experimental dataset, and created SDB8\_1, SDB8\_2, SDB8\_3, SDB8\_4, SDB8\_5, and SDB8\_6, which are one, two, three, four, five, and six times the size of SDB8, respectively. Obviously, if *minsup* is a constant, the longer the sequence, the more frequent patterns. The running time is positive related with number of frequent patterns and the sequence length according to Theorem 7. To avoid the impact of the different number of patterns on the running time, we set *minsup*=10, 20, 30, 40, 50, and 60 on SDB8\_1-SDB8\_6. All these algorithms mine 1243 patterns, and the comparison of running time is shown in Fig. 7.

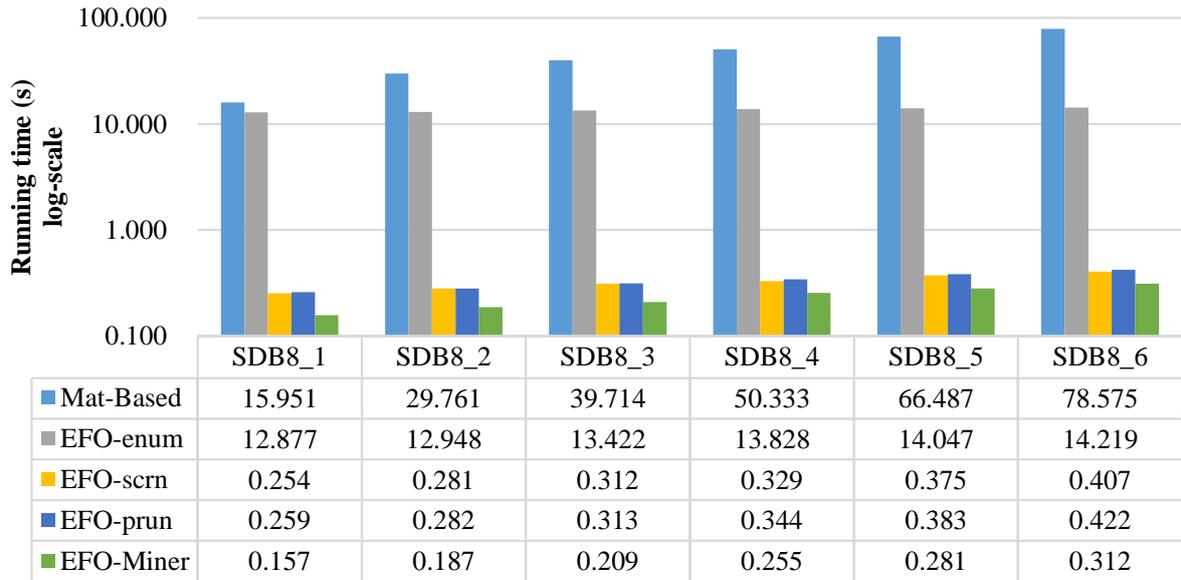


Figure 7: Comparison of running time with different dataset sizes

The results give rise to the following observations. From Fig. 7, we know that the running time of EFO-Miner grows slower than the dataset size. For example, the size of SDB8\_6 is six times of SDB8\_1, while EFO-Miner takes 0.331s on SDB8\_6, which is  $0.331/0.147=2.252$  times of SDB8\_1. This phenomenon can be found in all other datasets. The results indicate that the running time is positively correlated with the dataset size, which is consistent with the analysis of time complexity of EFO-Miner. More importantly, EFO-Miner runs significantly faster than other competitive algorithms, such as Mat-Based, EFO-enum, EFO-scrn, and EFO-prun. The reason is the same as the analysis in Section 5.2. Hence, EFO-Miner has strong scalability, since the mining performance does not degrade as the dataset size increases.

#### 5.4. Influence of different *minsup*

In this section, to report the influence of different *minsup* on number of patterns and running time of EFO-Miner, we selected Mat-Based, EFO-enum, EFO-scrn, and EFO-prun as competitive algorithms, and selected dataset SDB16 and expanded it by 10 times to obtain a larger dataset as the experimental dataset. We set *minsup*=1600, 1700, 1800, 1900, 2000, and 2100, respectively. The comparison of number of patterns and running time on SDB16 are shown in Figs. 8 and 9, respectively.

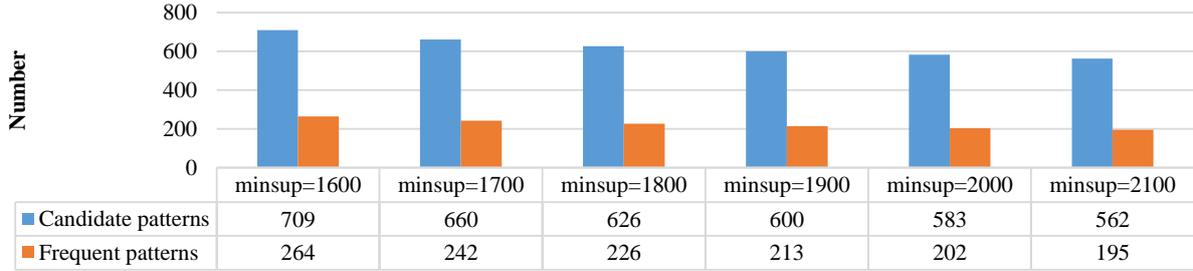


Figure 8: Comparison of number of patterns with different *minsup* on SDB16

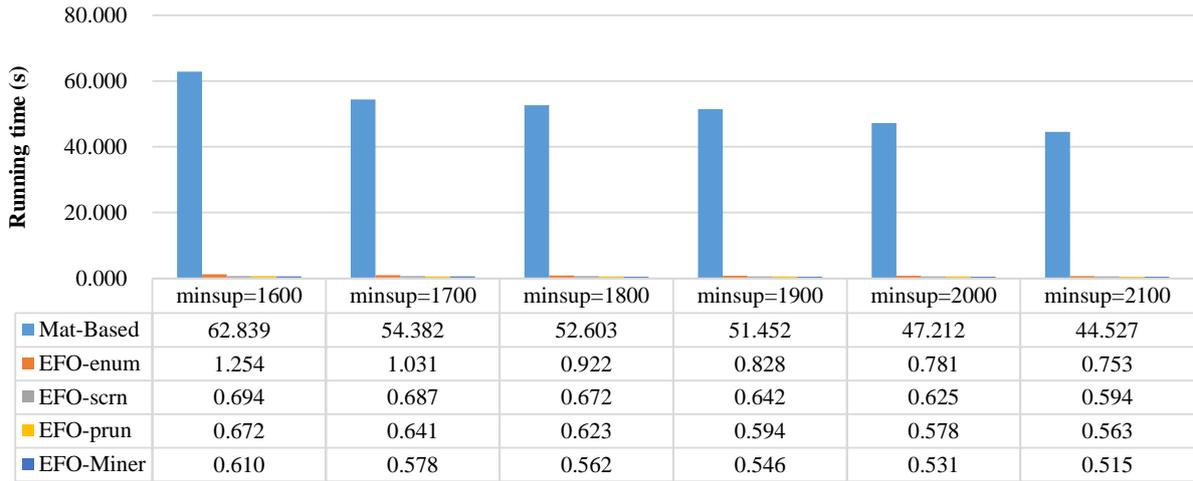


Figure 9: Comparison of running time with different *minsup* on SDB16

The results give rise to the following observations. With the increase of *minsup*, the number of patterns and running time decreases. For example, from Figs. 8 and 9, when *minsup*=1600, EFO-Miner discovers 264 OPPs and takes 0.610s, whereas when *minsup*=2100, EFO-Miner discovers 195 OPPs and takes 0.515s. This phenomenon can also be found in other competitive algorithms. The reason for this is as follows. With the increase of *minsup* value, the number of frequent patterns decreases. As a result, the running time also decreases. Moreover, EFO-Miner outperforms other competitive algorithms, which is consistent with the results of Section 5.2.

#### 5.5. Performance of OPR-Miner

In this case, OPR-Rule was selected as a comparison algorithm to generate all the OPRs, and experiments were carried out on SDB1–SDB8. We set the minimum support threshold *minsup*=12 and the minimum confidence threshold *minconf*=0.45. The number of generated rules is shown in Fig. 10. Moreover, Fig. 11 shows the comparison of the confidences of OPRs and strong OPRs for SDB3.

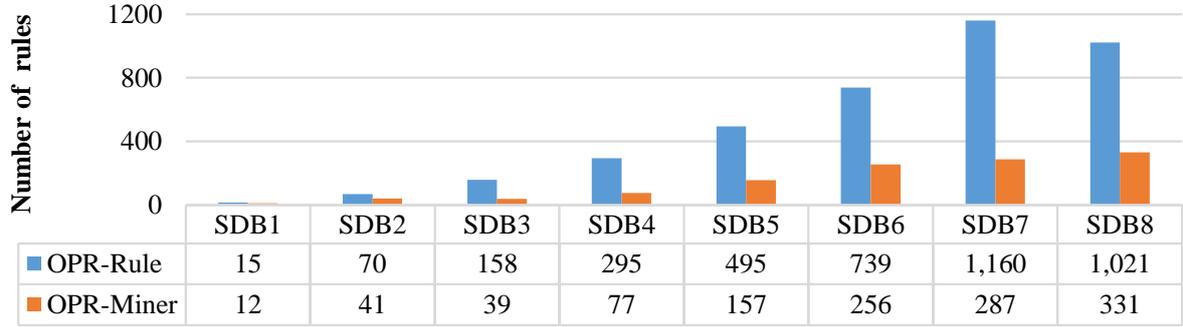


Figure 10: Comparison of number of rules on SDB1–SDB8

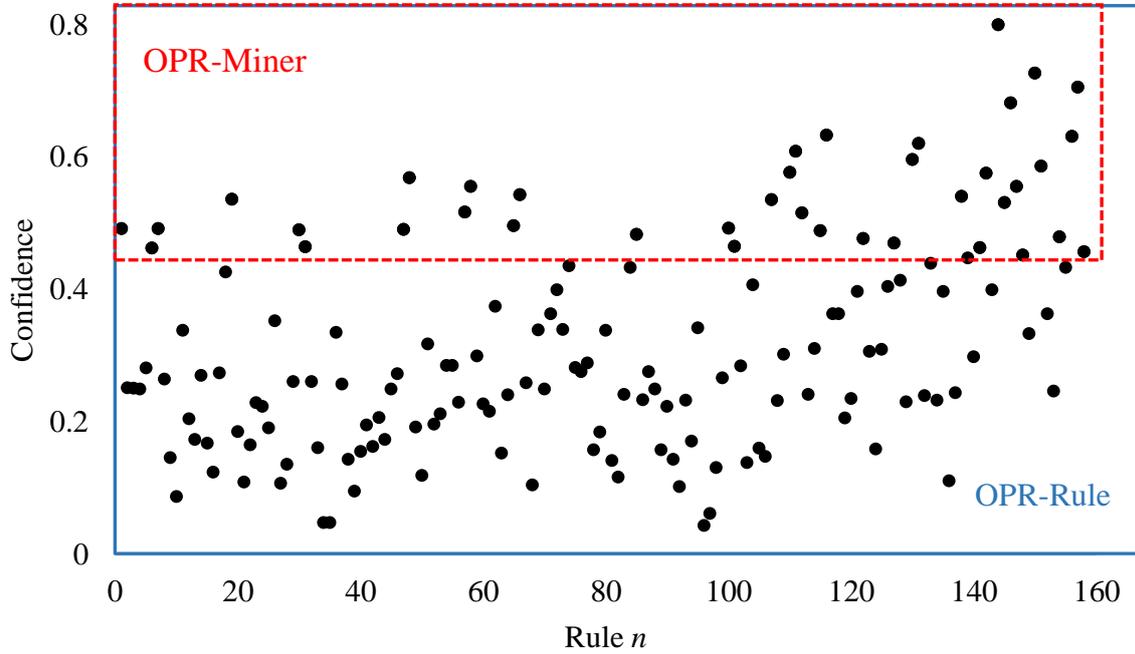


Figure 11: Comparison of confidences of OPRs and strong OPRs for SDB3. A point represents a rule, the X-axis represents the  $n$ -th rule, and the Y-axis represents the confidence level of the rule.

The results indicate that OPR-Miner outperforms OPR-Rule, thus validating that OPR-Miner can efficiently prune rules. For example, from Fig. 10, we know that on SDB3, OPR-Miner generates 39 candidate patterns, while OPR-Rule generates 158. Our experimental results are therefore consistent with those in Example 9. Moreover, Fig. 11 shows that the mined rules of OPR-Miner are a part of OPR-Rule. More importantly, OPR-Miner can mine rules with high confidences. Since we set the minimum confidence threshold  $minconf = 0.45$ , the confidences of the OPRs mined by OPR-Miner are no less than 0.45, while some of the confidences of OPRs mined by OPR-Rule are less than 0.45. Hence, OPR-Miner can find more useful rules than OPR-Rule.

### 5.6. Influence of different $minconf$

To report the influence of different  $minconf$  on the number of patterns and running time of OPR-Miner, we also selected dataset SDB16 and expanded it by 10 times to obtain a larger dataset as the experimental dataset. We selected

OPR-Rule as the competitive algorithm. We set  $minsup=1800$  and  $minconf=0.40, 0.45, 0.50, 0.55, 0.60,$  and  $0.65,$  respectively. The running time of OPR-Miner and OPR-Rule on all  $minconf$  is all about 0.563s, and the comparison of number of strong OPRs with different  $minconf$  is shown in Fig. 12.

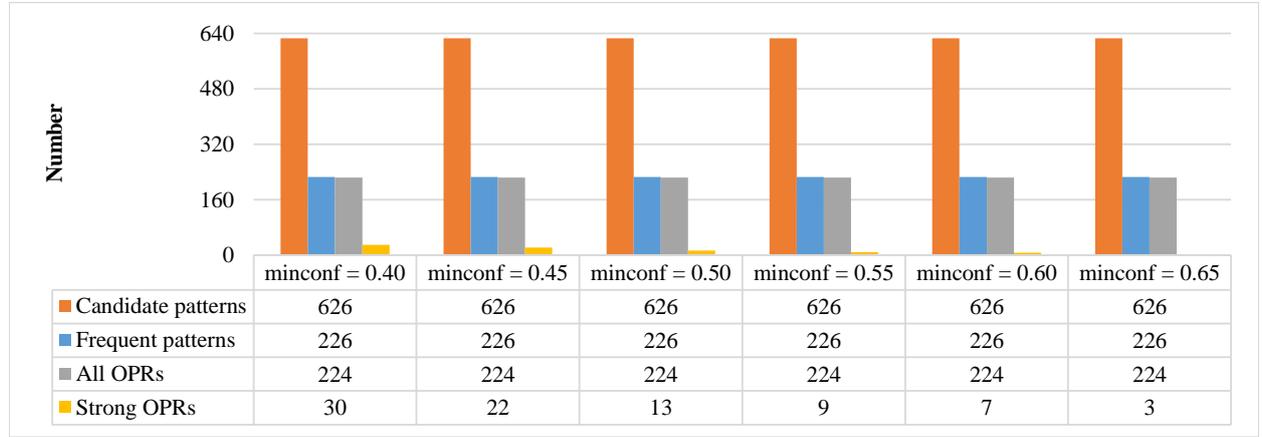


Figure 12: Comparison of number of strong OPRs with different  $minconf$

The results give rise to the following observations. The running time of OPR-Miner and OPR-Rule are almost the same, since OPR-Miner discovers a subset of OPR-Rule, and the process requires almost no time. Moreover, with the increase of  $minconf$ , the number of candidate patterns, frequent patterns, and all OPRs are constant, while the number of strong OPRs decreases. For example, no matter what  $minconf$  is, the number of candidate patterns, frequent patterns, and all OPRs are 626, 226, and 224, respectively. However, when  $minconf=0.40$ , the number of strong OPRs is 30, while when  $minconf=0.65$ , the number of strong OPRs is 3. The reason is as follows. We know that the number of candidate patterns, frequent patterns, and all OPRs are not related with the value of  $minconf$ . Thus, with the increase of  $minconf$ , the number of candidate patterns, frequent patterns, and all OPRs are constant. However, the number of strong OPRs is related with the value of  $minconf$ . With the increase of  $minconf$ , the number of strong OPRs decreases.

### 5.7. Case studies

In this section, we report the applications of rule mining from two aspects: clustering and classification. To evaluate the performance of OPR-Miner, we selected OPP-Miner as the competitive model. We know that each OPR can be represented by  $x \rightarrow y$ , where  $x$  and  $y$  are two frequent OPPs. Thus, each OPR is composed by two OPPs. If we mine  $t$  OPRs, then there will be  $k$  different OPPs, where  $k \leq 2 \times t$ , since some OPPs may be the same. For example,  $(1,2,3) \rightarrow (1,2,3,4)$  and  $(1,2,3,4) \rightarrow (1,2,3,4,5)$  are two different OPRs. However, there are only three different OPPs:  $(1,2,3)$ ,  $(1,2,3,4)$ , and  $(1,2,3,4,5)$ . We use  $k$  supports of corresponding OPPs to form a new dataset as OPRs. For fairness, we also selected top- $k$  supports of OPPs to form a new dataset as OPPs. The original dataset is called Raw.

#### 5.7.1. Clustering performance

To validate the clustering performance, a clustering experiment is conducted in this section. Since SDB1-SDB8 are single sequence datasets, clustering experiment cannot be conducted. Thus, we selected SDB9-SDB12 to conduct the experiment according to the following steps.

1. We employ OPR-Miner to mine OPRs and the parameters are  $minsup = 25$  and  $minconf = 0.65$ . We discover 6 OPRs corresponding to 8 OPPs on SDB9, 8 OPRs corresponding to 12 OPPs on SDB10, 8 OPRs corresponding to 12 OPPs on SDB11, and 8 OPRs corresponding to 11 OPPs on SDB12. Therefore, we discover top-8, top-12, top-12, and top-11 OPPs on SDB9, SDB10, SDB11, and SDB12, respectively. We show the comparison of the mined OPRs, their corresponding OPPs, top- $k$  OPPs, and shared OPPs in Table 4.

2. We adopt K-Means to cluster the Raw, OPPs, and OPRs data with parameter  $K = 7$ .

3. To evaluate the clustering performance, we select two criteria: Normalized Mutual Information ( $NMI$ ) [44] and Homogeneity ( $h$ ) [45], which can be calculated according to Equations 1 and 2, respectively.

Table 4: Comparison of mined patterns

Dataset	Type	Number	Mined OPRs or OPPs
SDB9	Strong OPRs	6	(1,2)→(1,2,3), (2,1)→(3,2,1), (1,2,3)→(1,2,3,4), (3,2,1)→(4,3,2,1), (1,2,3,4)→(1,2,3,4,5), (1,2,3,4,5)→(1,2,3,4,5,6)
	Corresponding OPPs	8	(1,2), (2,1), (1,2,3), (3,2,1), (1,2,3,4), (4,3,2,1), (1,2,3,4,5), (1,2,3,4,5,6)
	Top- <i>k</i> OPPs	8	(1,2), (1,2,3), (2,1), (1,2,3,4), (1,2,3,4,5), (3,2,1), (1,2,3,4,5,6), (1,2,3,4,5,6,7)
	Shared OPPs	7	(1,2), (2,1), (1,2,3), (3,2,1), (1,2,3,4), (1,2,3,4,5), (1,2,3,4,5,6)
SDB10	Strong OPRs	8	(1,2,3)→(1,2,3,4), (3,2,1)→(4,3,2,1), (1,2,3,4)→(1,2,3,4,5), (4,3,2,1)→(5,4,3,2,1), (1,2,3,4,5)→(1,2,3,4,5,6), (5,4,3,2,1)→(6,5,4,3,2,1), (1,2,3,4,5,6)→(1,2,3,4,5,6,7), (6,5,4,3,2,1)→(7,6,5,4,3,2,1)
	Corresponding OPPs	12	(1,2), (2,1), (1,2,3), (3,2,1), (1,2,3,4), (4,3,2,1), (1,2,3,4,5), (5,4,3,2,1), (1,2,3,4,5,6), (6,5,4,3,2,1), (1,2,3,4,5,6,7), (7,6,5,4,3,2,1)
	Top- <i>k</i> OPPs	12	(2,1), (3,2,1), (1,2), (4,3,2,1), (1,2,3), (5,4,3,2,1), (6,5,4,3,2,1), (1,2,3,4), (7,6,5,4,3,2,1), (1,2,3,4,5), (8,7,6,5,4,3,2,1), (9,8,7,6,5,4,3,2,1)
	Shared OPPs	10	(2,1), (3,2,1), (1,2), (4,3,2,1), (1,2,3), (5,4,3,2,1), (6,5,4,3,2,1), (1,2,3,4), (7,6,5,4,3,2,1), (1,2,3,4,5)
SDB11	Strong OPRs	8	(1,2,3)→(1,2,3,4), (3,2,1)→(4,3,2,1), (1,2,3,4)→(1,2,3,4,5), (4,3,2,1)→(5,4,3,2,1), (1,2,3,4,5)→(1,2,3,4,5,6), (5,4,3,2,1)→(6,5,4,3,2,1), (1,2,3,4,5,6)→(1,2,3,4,5,6,7), (6,5,4,3,2,1)→(7,6,5,4,3,2,1)
	Corresponding OPPs	12	(1,2,3), (1,3,2), (2,1,3), (3,2,1), (1,2,3,4), (4,3,2,1), (1,2,3,4,5), (5,4,3,2,1), (1,2,3,4,5,6), (6,5,4,3,2,1), (1,2,3,4,5,6,7), (7,6,5,4,3,2,1)
	Top- <i>k</i> OPPs	12	(2,1), (3,2,1), (4,3,2,1), (5,4,3,2,1), (6,5,4,3,2,1), (7,6,5,4,3,2,1), (8,7,6,5,4,3,2,1), (9,8,7,6,5,4,3,2,1), (10,9,8,7,6,5,4,3,2,1), (11,10,9,8,7,6,5,4,3,2,1), (12,11,10,9,8,7,6,5,4,3,2,1), (1,2)
	Shared OPPs	5	(3,2,1), (4,3,2,1), (5,4,3,2,1), (6,5,4,3,2,1), (7,6,5,4,3,2,1)
SDB12	Strong OPRs	8	(2,1)→(3,2,1), (1,2,3)→(1,2,3,4), (3,2,1)→(4,3,2,1), (1,2,3,4)→(1,2,3,4,5), (4,3,2,1)→(5,4,3,2,1), (1,2,3,4,5)→(1,2,3,4,5,6), (5,4,3,2,1)→(6,5,4,3,2,1), (1,2,3,4,5,6)→(1,2,3,4,5,6,7)
	Corresponding OPPs	11	(1,2), (2,1), (1,2,3), (3,2,1), (1,2,3,4), (4,3,2,1), (1,2,3,4,5), (5,4,3,2,1), (1,2,3,4,5,6), (6,5,4,3,2,1), (1,2,3,4,5,6,7)
	Top- <i>k</i> OPPs	11	(2,1), (1,2), (3,2,1), (1,2,3), (4,3,2,1), (5,4,3,2,1), (1,2,3,4), (6,5,4,3,2,1), (1,2,3,4,5), (7,6,5,4,3,2,1), (1,2,3,4,5,6)
	Shared OPPs	10	(1,2), (2,1), (1,2,3), (3,2,1), (1,2,3,4), (4,3,2,1), (1,2,3,4,5), (5,4,3,2,1), (1,2,3,4,5,6), (6,5,4,3,2,1)

$$NMI(X, Y) = \frac{\sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} P(i, j) \log \left( \frac{P(i, j)}{P(i)P(j)} \right)}{\sqrt{\sum_{i=1}^{|X|} P(i) \log P(i) \times \sum_{j=1}^{|Y|} P(j) \log P(j)}} \quad (1)$$

$$h(X, Y) = 1 - \frac{-\sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} P(i, j) \log P(i | j)}{-\sum_{i=1}^{|X|} P(i) \log P(i)} \quad (2)$$

Both  $NMI$  and  $h$  reflect the similarity between the clustering results and the actual values. The greater the  $NMI$  and  $h$ , the greater the similarity, i.e., the better the clustering performance. The comparison of clustering performances is shown in Fig. 13.

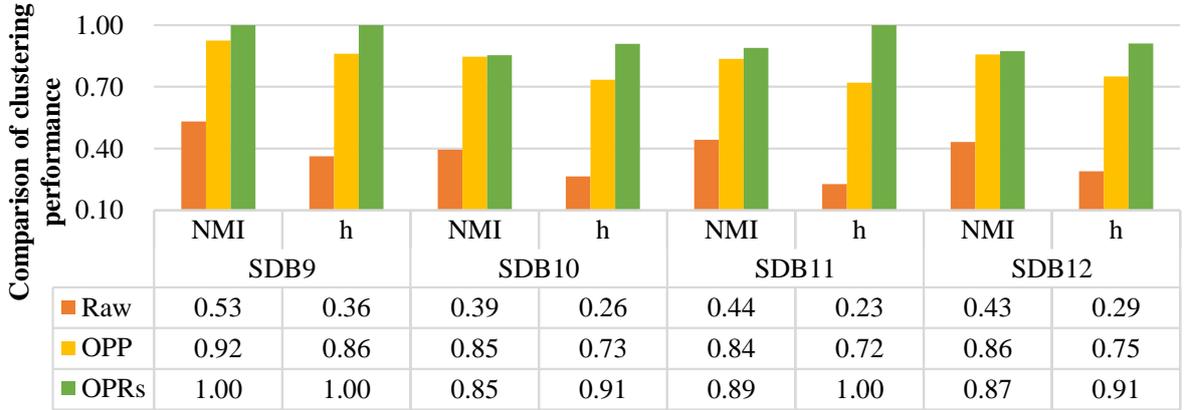


Figure 13: Comparison of clustering performances on SDB9–SDB12

The results give rise to the following observations.

1. Both OPP-Miner and OPR-Miner can effectively extract the critical information from the original time series. For example, the values of  $NMI$  of Raw, OPPs, and OPRs on SDB9 are 0.57, 0.85, and 0.89, respectively. The performances of OPP-Miner and OPR-Miner are better than Raw. The same effect can be found on all other datasets. The reason is that the original data may contain much redundant information, which can affect the clustering performance, while both OPP-Miner and OPR-Miner use the frequent trends to represent the original time series, which are more critical information with high support and high confidence. The results indicate that OPP-Miner and OPR-Miner can be used for feature selection for clustering task.

2. OPR-Miner has better performance than OPP-Miner. For example, the values of  $NMI$  of OPPs and OPRs on SDB10 are 0.86 and 0.88, respectively. The same effect can be found on all other datasets except SDB9 for  $h$ . The reason is that although top- $k$  OPPs are very critical information with high supports, some OPPs have lower confidence. However, OPR-Miner can extract the critical information with high support and high confidence, which can improve the clustering performance.

3. It is a very interesting phenomenon that some datasets share many common OPPs, while others share few. For example, on SDB9, OPR-Miner discovers six strong OPRs which are composed of eight patterns, and among them, seven patterns are Top-8 OPPs. But on SDB11, OPR-Miner discovers eight strong OPRs which are composed of 12 patterns, and among them, only five patterns are Top-12 OPPs. This result indicates that there is no clear relationship between top- $k$  OPPs and strong OPRs. For a specific time series clustering problem, how to extract effective features to achieve high-quality clustering performance is worthy of further study.

### 5.7.2. Classification performance

To validate the classification performance, a classification experiment is conducted in this section. We conducted the experiment on SDB13-SDB15. We chose five classical classification algorithms: SVM with Polynomial kernel function, C4.5, CART, AdaBoost, and KNN, which are Top 10 algorithms in data mining [46].

To mine OPRs, the parameters are  $minsup = 15$  and  $minconf = 0.25$ . We discover 7 OPRs corresponding to 12 OPPs on SDB13, 5 OPRs corresponding to 10 OPPs on SDB14, and 7 OPRs corresponding to 10 OPPs on SDB15. Since the three datasets are binary classification datasets, we adopt the prediction accuracy as the criterion. Moreover, we employ three-fold cross-validation to verify the classification performance. The comparisons of accuracy on SDB13, SDB14, and SDB15 are shown in Figs. 14, 15, and 16, respectively.

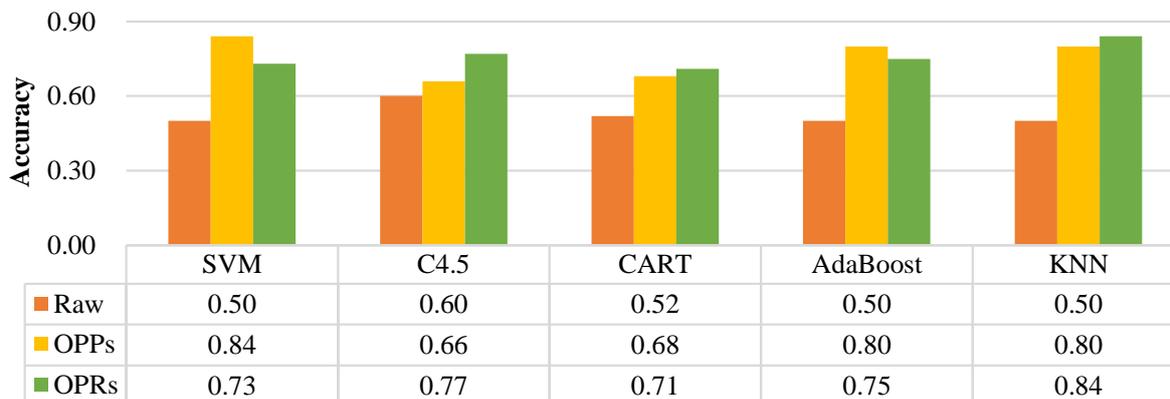


Figure 14: Comparison of accuracy on SDB13

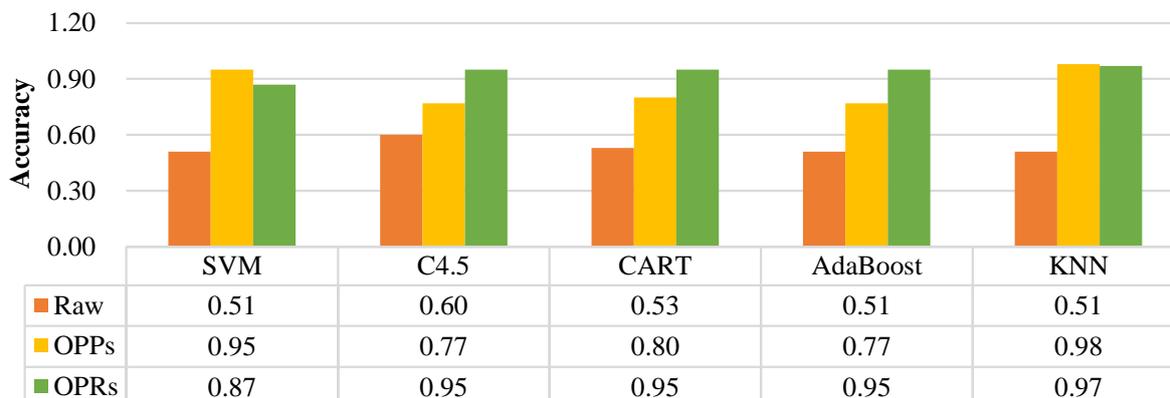


Figure 15: Comparison of accuracy on SDB14

From Figs. 14-16, we observe that both OPP-Miner and OPR-Miner can effectively improve the classification performance. For example, in Fig. 14, if we use C4.5 as the classifier, the accuracy of the original data on SDB13 is 0.60, while those of OPPs and OPRs are 0.66 and 0.77, respectively. The classification performance is significantly improved. This effect can be found on all the other datasets. Moreover, OPR-Miner has better classification performance than OPP-Miner. The reason is the same as that in clustering experiments.

## 6. Conclusion

To improve the efficiency of OPP mining and mine the implicit relations between OPPs, we have addressed the issue of OPR mining and proposed an effective mining algorithm called OPR-Miner. In this approach, the key step is finding frequent OPPs. To mine these frequent OPPs, we proposed an algorithm called EFO-Miner consisting of four parts. To reduce the number of candidate patterns, EFO-Miner adopts a pattern fusion strategy to generate

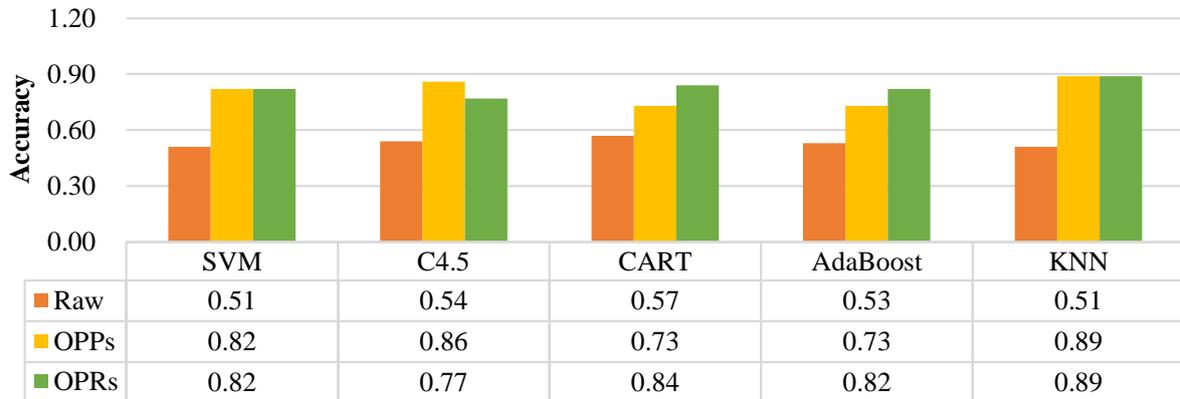


Figure 16: Comparison of accuracy on SDB15

candidate patterns. Moreover, to calculate the supports of super-patterns, EFO-Miner uses the matching results of sub-patterns based on the pattern fusion strategy. To improve the efficiency of support calculations, EFO-Miner employs a screening strategy to dynamically reduce the size of the matching results for sub-patterns. To avoid useless support calculations, EFO-Miner applies a pruning strategy to dynamically prune the sub-patterns for which the size of the matching results is less than the minimum support threshold. Experimental results from weather, oil, and stock datasets verify that OPR-Miner gives better performance than other competitive algorithms. More importantly, clustering and classification experiments validate that OPR-Miner can be used to realize feature extraction and achieve good performance.

## Acknowledgement

This work was partly supported by National Natural Science Foundation of China (61976240, 52077056, 62120106008), National Key Research and Development Program of China (2016YFB1000901), and Natural Science Foundation of Hebei Province, China (Nos. F2020202013, E2020202033).

## References

- [1] C. Dai, J. Wu, D. Pi, S. I. Becker, L. Cui, Q. Zhang, and B. Johnson, "Brain EEG time-series clustering using maximum-weight clique," *IEEE Trans. Cybern.*, vol. 52, no. 1, pp. 357-371, 2022.
- [2] Q. Li, J. Tan, J. Wang, and H. Chen, "A multimodal event-driven LSTM model for stock prediction using online news," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 10, pp. 3323-3337, 2021.
- [3] Z. Karevan and J. A.K. Suykens, "Transductive LSTM for time-series prediction: An application to weather forecasting," *Neural Networks*, vol. 125, pp. 1-9, 2020.
- [4] R. Wu and E. Keogh, "Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress," *IEEE Trans. Knowl. Data Eng.*, DOI: 10.1109/TKDE.2021.3112126, 2021.
- [5] R. Rezvani, P. M. Barnaghi, and S. Enshaeifar, "A new pattern representation method for time-series data," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 7, pp. 2818-2832, 2021.
- [6] Y. Wu, X. Wang, Y. Li, L. Guo, Z. Li, J. Zhang, and X. Wu, "OWSP-Miner: Self-adaptive one-off weak-gap strong pattern mining," *ACM Trans. Manag. Inf. Syst.*, vol. 13, no. 3, pp. 25, 2022.
- [7] H. Mannila and H. Toivonen, "Discovering generalized episodes using minimal occurrences," *KDD*, vol. 96, pp. 146-151, 1996.
- [8] P. Fournier-Viger, P. Yang, J. C. W. Lin, and U. Yun, "HUE-Span: Fast high utility episode mining," *International Conference on Advanced Data Mining and Applications*, pp. 169-184, 2019.
- [9] X. Ao, P. Luo, J. Wang, F. Zhuang, and Q. He, "Mining precise-positioning episode rules from event sequences," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 3, pp. 530-543, 2017.
- [10] Y. Chen, P. Fournier-Viger, F. Nouioua, and Y. Wu, "Sequence prediction using partially-ordered episode rules," *2021 International Conference on Data Mining Workshops (ICDMW)*, pp. 574-580, 2021.
- [11] E. Keogh, K. Chakrabarti, and M. Pazzani, "Locally adaptive dimensionality reduction for indexing large time series databases," *ACM SIGMOD Conference on Management of Data*, pp. 151-162, 2001.

- [12] J. Lin, E. Keogh, L. Wei, and S. Lonardi. "Experiencing SAX: a novel symbolic representation of time series," *Data Min. Knowl. Discov.*, vol. 15, no.2, pp. 107-144, 2007.
- [13] Y. Li, L. Yu, J. Liu, L. Guo, Y. Wu, and Xindong Wu, "NetDPO: (delta, gamma)-approximate pattern matching with gap constraints under one-off condition," *Appl. Intell.*, vol. 52, no. 11, pp. 12155–12174, 2022.
- [14] Y. Wu, J. Fan, Y. Li, L. Guo, and X. Wu, "NetDAP: (delta, gamma)-Approximate pattern matching with length constraints," *Appl. Intell.*, vol. 50, no. 11, pp. 4094-4116, 2020.
- [15] Y. Wu, Z. Yuan, Y. Li, L. Guo, P. Fournier-Viger, and Xindong Wu, "NWP-Miner: Nonoverlapping weak-gap sequential pattern mining," *Inf. Sci.*, vol. 588, pp. 124-141, 2022.
- [16] F. Min, Z. Zhang, W. Zhai, and R. Shen, "Frequent pattern discovery with tri-partition alphabets," *Inf. Sci.*, vol. 507, pp. 715-732, 2020.
- [17] Y. Wu, L. Luo, Y. Li, L. Guo, P. Fournier-Viger, X. Zhu, and X. Wu, "NTP-Miner: Nonoverlapping three-way sequential pattern mining," *ACM Trans. Knowl. Discov. Data*, vol. 16, no. 3, pp. 51, 2022.
- [18] J. Kim, P. Eades, R. Fleischer, S. Hong, C. S. Iliopoulos, K. Park, S. J. Puglisi, and T. Tokuyama, "Order-preserving matching," *Theor. Comput. Sci.*, vol. 525, pp. 68-79, 2014.
- [19] S. Cho, J. C. Na, K. Park, and J. S. Sim, "A fast algorithm for order-preserving pattern matching," *Inf. Process. Lett.*, vol. 115, no. 2, pp. 397-402, 2015.
- [20] M. Kubica, T. Kulczynski, J. Radoszewski, W. Rytter, and T. Walen, "A linear time algorithm for consecutive permutation pattern matching," *Inf. Process. Lett.*, vol. 113, no. 12, pp. 430-433, 2013.
- [21] Y. Wu, Q. Hu, Y. Li, L. Guo, X. Zhu, and X. Wu, "OPP-Miner: Order-preserving sequential pattern mining for time series," *IEEE Trans. Cybern.*, DOI: 10.1109/TCYB.2022.3169327, 2022.
- [22] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C. Wu, and V. S. Tseng, "SPMF: A java open-source pattern mining library," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3389-3393, 2014.
- [23] J. S. Okolica, G. L. Peterson, R. F. Mills, and M. R. Grimaila, "Sequence pattern mining with variables," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 1, pp. 177-187, 2020.
- [24] Y. Wu, L. Wang, J. Ren, W. Ding, and X. Wu, "Mining sequential patterns with periodic wildcard gaps," *Appl. Intell.*, vol. 41, no. 1, pp. 99-116, 2014.
- [25] X. Dong, P. Qiu, J. Lu, L. Cao, and T. Xu, "Mining Top-k useful negative sequential patterns via learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 9, pp. 2764-2778, 2019.
- [26] Y. Wu, M. Chen, Y. Li, J. Liu, Z. Li, J. Li, and X. Wu, "ONP-Miner: One-off negative sequential pattern mining," *ACM Trans. Knowl. Discov. Data*, doi: 10.1145/3549940, 2022.
- [27] W. Gan, J. C. W. Lin, P. Fournier-Viger, H.-C. Chao, V. S. Tseng, P. S. Yu, "A survey of utility-oriented pattern mining," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1306-1327, 2021.
- [28] W. Song, L. Liu, and C. Huang, "Generalized maximal utility for mining high average-utility itemsets," *Knowl. Inf. Syst.* vol. 63, pp. 2947-2967, 2021.
- [29] T. Truong, H. V. Duong, B. Le, and P. Fournier-Viger, "Efficient vertical mining of high average-utility itemsets based on novel upper-bounds," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 301-314, 2019.
- [30] H. Mannila, H. Toivonen, and A. Inkeri Verkamo, "Discovery of frequent episodes in event sequences," *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 259-289, 1997.
- [31] X. Ao, P. Luo, C. Li, F. Zhuang, and Q. He, "Online frequent episode mining," *IEEE 31st International Conference on Data Engineering*, pp. 891-902, 2015.
- [32] S. Ghosh, J. Li, L. Cao, and K. Ramamohanarao. Hsu, "Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns," *J. Biomed. Informatics*, vol. 66, pp. 19-31, 2017.
- [33] Y. Li, S. Zhang, L. Guo, J. Liu, Y. Wu, and X. Wu, "NetNMSF: Nonoverlapping maximal sequential pattern mining," *Appl. Intell.*, vol. 52, no. 9, pp. 9861-9884, 2022.
- [34] N. Nishimura, N. Sukegawa, Y. Takano, and J. Iwanaga, "A latent-class model for estimating product-choice probabilities from clickstream data," *Inf. Sci.*, vol. 429, pp. 406-420, 2018.
- [35] T. Wang, L. Duan, G. Dong, and Z. Bao, "Efficient mining of outlying sequence patterns for analyzing outlierness of sequence data," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 5, pp. 62, 2020.
- [36] Y. Wu, Y. Wang, Y. Li, X. Zhu, and X. Wu, "Top-k self-adaptive contrast sequential pattern mining," *IEEE Trans. Cybern.*, DOI: 10.1109/TCYB.2021.3082114, 2021.
- [37] J. D. Smedt, G. Deeva, and J. D. Weerd, "Mining behavioral sequence constraints for classification," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1130-1142, 2020.
- [38] Y. Wu, Y. Tong, X. Zhu, and X. Wu, "NOSEP: Nonoverlapping sequence pattern mining with gap constraints," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2809-2822, 2018.
- [39] P. Zhang, and A. Mikhail, "On approximate pattern matching with thresholds," *Inf. Process. Lett.*, vol. 123, no. 7, pp. 21-26, 2017.
- [40] P. Fournier-Viger, U. Faghihi, R. Nkambou, and E. M. Nguifo, "CMRules: Mining sequential rules common to several sequences," *Knowl. Based Syst.*, vol. 25, no. 1, pp. 63-76, 2012.
- [41] T. Pham, J. Luo, T. Hong, and B. Vo, "An efficient method for mining non-redundant sequential rules using attributed prefix-trees," *Eng. Appl. Artif. Intell.*, vol. 32, pp. 88-99, 2014.
- [42] P. Fournier-Viger, C.-W. Wu, V.S. Tseng, L. Cao, R. Nkambou, "Mining partially-ordered sequential rules common to multiple sequences," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2203-2216, 2015.
- [43] T. Chhabra, and J. Tarhio. "A filtration method for order-preserving matching," *Inf. Process. Lett.*, vol. 116 no. 2, pp. 71-74, 2016.
- [44] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, pp. 09008, 2005.
- [45] A. Rosenberg, and J. Hirschberg, "V-Measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of Empirical*

- Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 410-420.
- [46] X. Wu, V. Kumar, J. R. Quinlan, J. H. Motoda, G. J. McLachlan, A. F. M. Ng, B. Liu, P. S. Yu, Z. Zhou, M. S. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1-37, 2008.