IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 21, 2021, accepted August 11, 2021, date of publication August 18, 2021, date of current version September 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3105966

A Novel Operation Sequence Similarity-Based Approach for Typical Process Route Knowledge Discovery

BINZI XU[1,2,3], **YAN WANG**[2], **AND ZHICHENG JI**[2]

[1]Key Laboratory of Electric Drive and Control, Anhui Higher Education Institutes, Anhui Polytechnic University, Wuhu 241000, China
[2]School of IoT and Engineering, Jiangnan University, Wuxi 214122, China
[3]School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand

Corresponding author: Yan Wang (wangyan88@jiangnan.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61973138, in part by the 111 Project under Grant B12018, in part by the Open Research Fund of Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Anhui Polytechnic University, under Grant DTESD2020B01, in part by the Scientific Research Foundation for Introduced Talent Scholars, Anhui Polytechnic University, under Grant 2020YQQ041, in part by the Natural Science Foundation of the Anhui Higher Education Institutions of China under Grant KJ2016A795 and Grant KJ2020A0345, and in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2020-069.

ABSTRACT A typical process route essentially represents the commonly used process planning-related knowledge and can be modified to generate new process routes easily. Hence, its quality directly affects the performance of newly generated process routes and thereby the goodness of products. To effectively discover typical process route knowledge, a reasonable similarity measure and a clustering method specifically for process routes are required. However, existing operation sequence similarity coefficients often assign coarse-grained similarities, which leads to inaccurate clustering results. For the clustering problem, most researchers have not considered the practical constraints during typical process route discovery. In this paper, an operation sequence similarity-based discovery method is presented. First, the characteristics and information requirements of the operation sequence similarity problem are analysed, and a novel comprehensive similarity coefficient combined with a modified pseudo-longest-common-subsequence (pseudo-LCS) and Jaccard similarity coefficient is proposed based on this analysis and principal component analysis (PCA). This coefficient considers the precedence relationship, the number of common operations, and the operation similarity simultaneously to handle all the potential similarity situations. Second, two soft constraints, namely, quantity constraint and size constraint, are introduced in the traditional process route clustering problem to ensure the quality and validity of the discovered typical process routes. To solve this more practical problem and achieve a balance between these two conflicting constraints, the K-medoids method is improved with an adjustment mechanism to generate valid results under these two soft constraints. Finally, numerical illustrations are presented to verify the effectiveness of the proposed methods. The results show that compared with existing similarity coefficients, the proposed comprehensive similarity coefficient is more sensitive and much better at distinguishing the tiny difference between the process routes. In addition, the modified K-medoids method can perform much better than existing methods on process route discovery data sets under two conflicting soft constraints.

INDEX TERMS Operation sequence similarity, knowledge discovery, typical process route, soft constraint, manufacturing.

I. INTRODUCTION

As customer demand becomes more personalized, modern manufacturing prefers the production mode with multiple varieties and small batches. Hence, industries require a more efficient process planning method to rapidly respond to changing market demands [1], [2].

Related studies have indicated that process routes in the same part family have 70% $\sim$ 80% similarity with each other [3], [4]. There are two main reasons for this situation:

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Tsun Cheng.

(1) the process routes in the same part family have global similarity; and (2) often, some local machining operation sequences are highly reused. Therefore, typical process route-based process route planning is seen as a promising method because it can effectively reuse existing process planning knowledge. To plan a suitable process route for a new part, the technician only needs to first find the corresponding typical process route in the same part family and then adjust this typical process route by adding/removing/editing operations. This kind of process planning method has been widely used in modern computer-aided process planning (CAPP) [5], [6]. Another potential application of the typical process route is that it can facilitate the construction of a specific reconfigurable manufacturing system (RMS) configuration [7] or cellular manufacturing system (CMS) [8], [9] in such a way that they have the capacity to handle the process requirements of all parts in the corresponding part family by minor adjustment.

A typical process route is the most general operation sequence in a part family, which means that it is essentially a process template that contains highly reused common process information and knowledge among all the historical process routes in the same part family. From the clustering perspective, the typical process route is the centre (or medoid) of a part family, which is close to any other process route in the same cluster. Hence, a typical process route can be seen as process planning-related knowledge summarized from existing process routes. A new process route can be easily obtained for the new part by modifying the typical process route properly. This typical process route-based process planning method can not only greatly improve the efficiency and quality of the process planning for new products but also facilitate the standardization of process routes. Since the process route is a kind of explicit expression of process technicians' experiences and knowledge, how to elicit typical process routes based on their similarity is an urgent problem for knowledge acquisition, accumulation, and sharing [10].

However, it is difficult for entrepreneurs to extract process route knowledge accurately and effectively because most of the knowledge is tacit. Specifically, such knowledge is deeply embedded in individuals' experience, skill, and even preferences [11]. Recently, data mining techniques have shown their effectiveness in manufacturing systems [12], such as process engineering design [13], quality improvement [14], cloud manufacturing [15], and semiconductor manufacturing [16]. With regard to typical process route knowledge discovery, clustering analysis is frequently applied.

A similarity definition is the foundation of clustering analysis [17]. Generally, the operation sequence similarity measure is one of the sequence similarity problems in sequence clustering. Unlike the applications to DNA/protein sequences [18], sentences [19] and process mining [20], the operation sequence similarity problem has some unique characteristics:

(1) There are various types of operations in manufacturing, and the similarities between different operation pairs are different.
(2) The lengths of different operation sequences are different.
(3) The precedence constraint relationship is the key information of operation sequences.
(4) The number of common operations between operation sequences should also be considered.

Most of the existing operation sequence similarity coefficients focus only on the precedence constraints and ignore the similarities between the operations themselves. This coarse-grained similarity could bring inaccurate clustering results because operation similarity directly affects the operation sequence similarity and thereby impacts the clustering result. It is difficult for clustering algorithms to make correct clustering decisions based solely on global similarity. In addition, the number of common operations becomes more significant when the precedence constraint relationship is similar. It is necessary to consider all the useful information to enable the similarity coefficient to measure the process route similarities more accurately and then guide the clustering algorithm to find more rational typical process routes.

On the other hand, the typical process route discovery problem is an exemplar-based clustering problem. Instead of clustering all the process routes [21], this approach tends to find the exemplars to represent other process routes properly. In other words, it pays more attention to the similarities between process routes and their exemplars rather than the intracluster and intercluster similarity. In addition, the typical process route discovery problem, as a practical matter, suffers from some practical constraints.

According to the literature [22], a valid typical process route should contain sufficient and representative common process-related information. Hence, two conditions should be considered in the typical process route discovery problem:

(1) The number of process routes in a cluster cannot be smaller than n, which is the minimum number of process routes in a valid process route cluster. This constraint is also called the quantity constraint.
(2) The radius of a cluster, i.e., the maximal distance between the process route and its exemplar in one cluster, should be smaller than ε, which is the maximum radius of a valid process route cluster. This constraint is also called the size constraint.

The quantity constraint indicates that a valid typical process route should be summarized from a sufficient number of existing process routes. In this way, its embedded common process-related information and knowledge are sufficiently supported, which means that it is frequently reused by many other process routes. If the quantity constraint is not satisfied, the corresponding typical process route (or process-related knowledge) is seen as being too infrequent to offer information with a high and wide reference value. The size constraint guarantees the representativeness of typical process routes. More specifically, if the radius of a cluster is too large, then similarities between the typical process route (i.e., the medoid) and the process routes in the marginal area are too small such that it is inappropriate for the typical process route to represent those marginal process routes. In other words, the common information between them is too small, which does not conform with the requirement of being a typical process route.

Note that the cluster (or typical process route) is invalid, which does not mean that the process routes in this cluster are irrelevant. These two constraints only serve the requirement of this specific practical problem (i.e., the validity of typical process routes).

However, most current studies simply treated the typical process route discovery problem as a traditional clustering problem and ignored the practical requirement of this specific problem, which could generate some unqualified exemplars (i.e., the typical process routes) [23], [24]. Some work [22] has considered these two conditions to be hard constraints, which makes the quality of the clustering results seriously dependent on the manually designed constraint parameters. Large constraint parameters will make the constraints less important, while small constraint parameters will make it impossible to find valid clustering results. Since the shape of the process route data set is unknown, it is difficult to set rational constraint parameters manually. Hence, the two constraints mentioned above are treated as soft constraints in this paper. In this way, the goal of the clustering process is to find the exemplar set that optimizes not only the total distance between the process routes and their exemplars but also the penalties of the two soft constraints.

This study represents the first time that these two constraints are considered as soft constraints simultaneously in the manufacturing field. In doing so, we can loose the dependence between the constraint parameters and the clustering results because the soft constraint is used to *guide* the clustering process, not *force* it, as the hard constraint does. Specifically, compared with hard constraints, soft constraints have higher fault tolerance and less sensitivity to constraint parameters. In addition, the penalty coefficients in this novel problem can be seen as the indexes that measure the importance of the corresponding soft constraints and that reflect the will of the process technicians. Therefore, it is necessary to consider these two conditions as soft constraints.

Essentially, the size constraint and quantity constraint conflict with each other. The size of a cluster increases when that cluster contains more process routes. Assume that the expected cluster number is 5. If $K < 5$, the number of data instances in the clusters will increase, and the sizes of these clusters can correspondingly increase, in which case the clustering procedure mainly suffers from the size constraint. If $K > 5$, the number of data instances in the clusters decreases, and the sizes of the clusters could also be smaller, in which case the clustering procedure mainly suffers from the quantity constraint. Therefore, how to achieve a balance between these two soft constraints is one of the issues that this paper attempts to solve.

To overcome the above drawbacks, the main contributions of this paper include the following:

(1) The characteristics of the operation sequence similarity problem are deeply analysed to understand the specific information requirement of this issue. Based on this analysis, all the potential similarity situations that could occur in the operation sequence similarity problem are given with corresponding examples.

(2) Based on the analysis above, a novel similarity coefficient is presented to consider not only the precedence constraint of the process routes but also the similarity between the operations and the number of common operations simultaneously. A modified pseudo-longest common-subsequence (pseudo-LCS) coefficient is proposed to quantify the information of the precedence constraint and the operation similarity between process routes. The Jaccard similarity coefficient is introduced in this paper to measure the number of common operations. Then, principal component analysis (PCA) is used to generate a comprehensive similarity coefficient based on these two similarity coefficients. This novel similarity coefficient can effectively assign suitable similarity degrees for different similarity situations.

(3) The quantity constraint and size constraint are treated as two soft constraints in this paper. To handle these two conflicting soft constraints and obtain suitable clustering results, a modified exemplar-based clustering algorithm with an adjustment mechanism is proposed based on the K-medoids method. The numerical illustration based on the generated process route sets shows that the proposed algorithm can effectively handle the typical process route discovery problem considering two conflict soft constraints and obtain appropriate clustering results with better performance.

In accomplishing the goal of overcoming the drawbacks mentioned above, the improvements that our methods bring for the current related technologies can be summarized as follows:

(1) A novel comprehensive operation sequence similarity coefficient is proposed to measure the similarity between two operation sequences more precisely. The numerical illustrations indicated that this novel similarity coefficient can handle both artificial data sets and real data sets better than the existing 10 coefficients. More specifically, our coefficient can effectively find the tiny similarity difference of operation sequence pairs and assign rational similarity values while others cannot, which means that it is a fine-grained similarity coefficient, as we expected.

(2) The traditional K-medoids method has been improved in this paper under the consideration of both quantity constraint and size constraint, which are seen as soft constraints. In this manner, the manually designed

constraint-related parameters have less influence on the clustering result. Hence, the required precision of the parameter settings can be lower, making it easier and more practical for workers to determine the constraint parameters. The numerical illustrations show that the modified K-medoids method can effectively find a trade-off between these two conflicting soft constraints.

The remainder of this paper is organized as follows. Section II describes the related work of the typical process route discovery problem, including the operation sequence similarity design and an exemplar-based clustering method. Then, the analysis of the operation sequence similarity problem is given in Section III, and the proposed comprehensive similarity coefficient is introduced in Section IV. Section V presents the proposed exemplar-based clustering method. Section VI gives a numerical illustration of the novel similarity coefficient and the modified clustering method. Section VII presents the conclusions and future work of this paper.

II. RELATED WORK

From a technical point of view, there are two main sub-problems in the typical process route knowledge discovery problem: selecting a suitable operation sequence-based similarity coefficient to describe the similarity between process routes quantitatively and selecting an appropriate clustering algorithm to discover typical process route knowledge hidden behind a large quantity of process route data.

This section also discusses some limitations of the existing work and highlights the contributions of this paper.

A. OPERATION SEQUENCE-BASED SIMILARITY COEFFICIENT

As a sequence similarity problem, operation sequence similarity is different from traditional sequence similarity problems such as DNA/protein sequence similarity [18], text similarity [19] and process mining [20]. DNA/protein sequences have only 4 types of nucleotides or 20 amino acids in the biological sequences. For this reason, letter sequence representation (LSR) is suitable for this situation [25]. Text similarity mainly focuses on the semantic similarity and word frequencies between two sentences. Process mining pays more attention to the event log, which considers the succession relationship rather than the precedence relationship of activities.

Considering the unique characteristics of the operation sequence similarity problem, many operation sequence-based similarity coefficients have been developed to measure the similarity between process routes rationally [26].

Choobineh [27] first considered the precedence relationship of operation sequences and proposed a novel similarity measure based on the Jaccard similarity coefficient. Tam [28] used the Levenshtein distance (also called the Edit distance) to describe the dissimilarity between sequences. Ho *et al.* [29] presented a compliant index generated from the forward compliant index and backward index, which can be seen as an improved LCS method. Askin and Zhou [30] also proposed a similarity coefficient based on the LCS, which is found by the common sequence tree. Irani and Huang [31] took the gaps between matching operations into consideration and developed a novel similarity index called the merger coefficient. Huang [32] modified this similarity index, which took the length information into consideration. Goyal *et al.* [33] analysed the advantages and disadvantages of the aforementioned similarity coefficients and proposed a novel bypassing moves and idle machines (BMIM) similarity coefficient based on LCS and the shortest common supersequence (SCS). Zhou and Dai [11], inspired by bioinformatics technology, used the Needleman-Wunsch (NW) algorithm to find the best alignment of two operation sequences. Wang *et al.* [34] also considered bypassing moves and idle machines and presented a novel similarity coefficient. Zhou and Dai [35] focused on part features to establish a fuzzy similarity matrix of process routes and obtained typical process sequences through granular computing. Navaei and Elmaraghy [36] first considered the similarity of product variants whose operation sequences are reticulated and applied the average linkage clustering (ALC) algorithm to cluster part/product variants. Wang *et al.* [22] presented an attributed directed graph to describe process routes and measure the similarities between them. Wu *et al.* [9] proposed an improved similarity coefficient for the cell formation problem, in which machine choice and usage are also considered.

Although much effort has been made to solve the operation sequence similarity problem, there is no existing work to deeply analyse the information requirements and characteristics of this problem. The lack of this kind of analysis makes it difficult for existing work to handle all possible similarity situations in the operation sequence similarity problem.

Most of the similarity coefficients mentioned above did not consider the similarity between the operations themselves. They simply treat two similar operations as two totally different operations, in which case the similarity values will be the same for the process route pairs with and without similar operations. Liu *et al.* [37] were the first to introduce the operation code scheme into the typical process route discovery problem and use this code scheme to calculate the operation similarities. This improvement has made typical process route discovery more realistic and effective. However, the authors did not consider the operation similarity information and the precedence relationship simultaneously, which could lead to inaccurate similarity measures because these two are closely related. Currently, no sequence similarity measure can rationally address the relationship between the precedence constraint and the operation similarity. In addition, these two pieces of information cover only partial information of common operations. Not having this information would make it difficult for coefficients to judge whether the process route pairs have only one common operation or have the same operation but the precedence is totally different.

B. CLUSTERING ALGORITHMS

The second phase of typical process route knowledge discovery is clustering process routes based on the proposed operation sequence similarity coefficient. The ALC algorithm [34], [36] has been widely used in previous studies. However, the purpose of this method is to ensure that the data points in the same cluster are as close as possible and that the distances between different clusters are as far as possible [38]. This approach is different from the aim of typical process route knowledge discovery, which is to find exemplars to represent other data.

Another commonly used clustering algorithm is the granular computing (GrC)-based method [11], [35]. This method first clusters objects into several classes based on different granule thresholds and then determines an optimal granular layer based on the information entropy, information gain, and other related information.

In addition to ALC and GrC, the K-means [39] and affinity propagation (AP) algorithm [22] have been used in the typical process route knowledge discovery problem. However, they also did not consider the validity of the clustering result. In addition, these clustering algorithms only output the clustering result and need an extra step to find the exemplar of each class.

As one of the most well-known exemplar-based clustering methods, the K-medoids method is suitable for typical process route knowledge discovery. There are several kinds of K-medoids method in the literature, such as partitioning around medoids (PAM) and clustering large applications (CLARA) [40]. The main idea of the K-medoids method is to find K data points that are defined as the centre of each cluster (i.e., exemplars). However, PAM requires considerable time for large data sets, and CLARA is not as powerful as PAM. Therefore, Park and Jun [41] proposed an improved K-medoids method, which effectively reduces the computing time of PAM.

To the best of our knowledge, no exemplar-based clustering algorithm takes the quantity soft constraint and size soft constraint into consideration simultaneously. Existing work often treated these two constraints as hard constraints and considered them only after clustering. Therefore, the related parameter setting is so important that it has a strong impact on the number of valid clusters. In fact, the size constraint often has a very large radius setting in most previous work [22] to ensure that enough valid clusters can be obtained. In other words, most of the existing work has difficulty handling the conflict between these two constraints when they are seen as hard constraints. Considering this practical requirement of the typical process route discovery problem, the K-medoids method is chosen as the clustering algorithm in this paper and improved with an adjustment mechanism in such a way that the K-medoids method can consider these two soft constraints simultaneously during clustering and achieve a balance between them.

III. ANALYSIS OF THE OPERATION SEQUENCE SIMILARITY PROBLEM

The operation sequence similarity problem can be mathematically described as follows: given an operation sequence pair set $\mathcal{P} = \left\{\left(x_i, x_j\right) \mid i, j = 1, \cdots, N\right\}$, where x_i and x_j are operation sequences, the goal is to design a proper similarity coefficient that is sensitive to the tiny difference between operation sequence pairs in the set $\mathcal{P}$. Mathematically, the following index should be maximized:

$$R\left(\mathcal{P}, s\right) = \frac{\left|\left\{\left(x_i, x_j\right) \mid r_r\left(x_i, x_j\right) = r_s\left(x_i, x_j\right)\right\}\right|}{\left|\left\{\left(x_i, x_j\right) \mid i, j = 1, \cdots, N\right\}\right|} \qquad (1)$$

where s is a similarity coefficient and $r_r\left(x_i, x_j\right)$ and $r_s\left(x_i, x_j\right)$ are the real and measured similarity relationships between the operation sequences x_i and x_j, respectively.

Eq. (1) indicates that this index is essentially the ratio of the number of operation sequence pairs with rational similarity relationships (i.e., $r_r\left(x_i, x_j\right) = r_s\left(x_i, x_j\right)$) that is assigned by similarity coefficient s to the total number of operation sequence pairs in the set $\mathcal{P}$. Specifically, $r_r\left(x_i, x_j\right) = r_s\left(x_i, x_j\right)$ is met when two conditions are satisfied: (1) the operation sequence pair can be distinguished from other pairs based on their difference, and (2) the operation sequence pair with a larger amount of common information should be assigned a higher similarity value and vice versa. The index R measures the sensitivity of the similarity coefficient s.

In this section, the specific information requirement of the operation sequence similarity problem is analysed, which captures the manifestations and characteristics of process routes. Then, different potential similarity cases of operation sequences are defined with corresponding examples to describe different information requirements in the practical application of typical process route knowledge discovery.

A. INFORMATION REQUIREMENT OF THE OPERATION SEQUENCE SIMILARITY PROBLEM

Different sequence similarity problems have their own characteristics and emphasis points, which are reflected by the information requirements of these specific problems. Therefore, information requirement analysis is a key step for designing the operation sequence similarity measure.

Based on previous work and the demand for real issues, three kinds of information are needed in the typical process route discovery problem: the precedence relationship, the number of common operations and the operation similarity.

1) PRECEDENCE RELATIONSHIP

The main idea of the precedence relationship has been commonly considered in previous work because it essentially reflects the process constraints in manufacturing. In the real situation, these process constraints determine which operation should be performed first and which operation should be performed later. However, many researchers treated it as a succession relationship and described it with the longest common substring.

Although succession relationships and precedence relationships both represent the order of operations, they pay more attention to different aspects, as shown in Figure 1 (a). The succession relationship requires the operations to be successive, which means that the first operation goes first and the next operation follows closely. A precedence relationship has no such requirement. In other words, the first operation goes first, and the next operation can be placed after several operations. Therefore, the longest common substring is often used for succession relationships, while the longest common subsequence is suitable for precedence relationships.

2) NUMBER OF COMMON OPERATIONS

Although LCS can effectively measure the precedence relationship between operation sequences, it requires two potential conditions: (1) the corresponding operations are identical; and (2) their precedence relationship is the same. Hence, it cannot measure the situation in which the operations are identical but the precedence relationship is different, similar to case 4 shown in Figure 1 (b). In this case, these common operations provide little similarity for two operation sequences from the point view of LCS. In other words, common operations with different precedence relationships are not considered to be similarity information for LCS.

The number of common operations is considered by the Jaccard similarity coefficient in this paper to measure the operation set similarity of operation sequences. In this way, this kind of information is separated from the precedence relationship and can also be considered when the precedence relationship is totally different.

3) OPERATION SIMILARITY

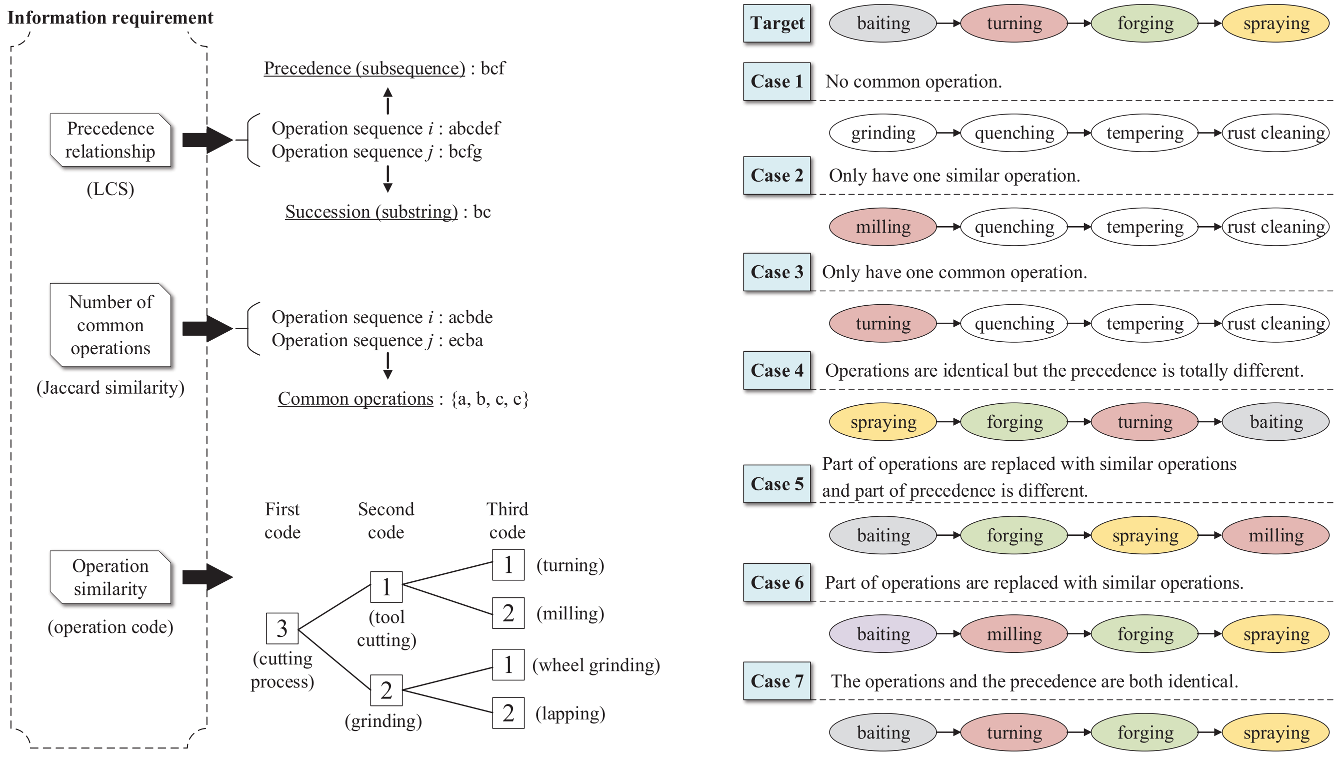
The information requirement of operation similarity differentiates the operation sequence similarity problem from the traditional sequence similarity problems. In fact, the operation itself contains multiple process-related information, including processing methods, specific requirements, tool usage, technological requirements, and other related characteristics.

Most previous work ignored such information, which leads to the scarcity of information for the similarity measure and further causes a coarse-grained clustering result. By considering the operation similarity, the proposed similarity measure can offer a fine-grained and multilevel similarity value and be more sensitive to the subtle difference of operation sequences.

As shown in Figure 1 (a), the operation similarity is described by the operation code in this paper, which uses a three-digit code to represent the process-related information from three levels. In Figure 1, operation "turning" is represented as "311", and "milling" is "312".

B. DIFFERENT SIMILARITY CASES OF OPERATION SEQUENCES

The combination of the three kinds of information discussed above leads to different operation sequence similarity cases,


Information requirement
Precedence relationship
(LCS)
Precedence (subsequence) : bcf
Operation sequence *i* : abcdef
Operation sequence *j* : bcfg
Succession (substring) : bc
Number of common operations
(Jaccard similarity)
Operation sequence *i* : acbde
Operation sequence *j* : ecba
Common operations : {a, b, c, e}
Operation similarity
(operation code)
First code
Second code
Third code
3
(cutting process)
1
(tool cutting)
2
(grinding)
1
(turning)
2
(milling)
1
(wheel grinding)
2
(lapping)
Target
baiting
turning
forging
spraying
Case 1
No common operation.
grinding
quenching
tempering
rust cleaning
Case 2
Only have one similar operation.
milling
quenching
tempering
rust cleaning
Case 3
Only have one common operation.
turning
quenching
tempering
rust cleaning
Case 4
Operations are identical but the precedence is totally different.
spraying
forging
turning
baiting
Case 5
Part of operations are replaced with similar operations and part of precedence is different.
baiting
forging
spraying
milling
Case 6
Part of operations are replaced with similar operations.
baiting
milling
forging
spraying
Case 7
The operations and the precedence are both identical.
baiting
turning
forging
spraying

(a) Information requirement of operation sequence similarity problem

(b) Different similarity cases of operation sequences

FIGURE 1. Analysis of the operation sequence similarity problem.

in which the degree of common information contained in the operation sequence pair determines their similarity. Based on this aspect, Figure 1 (b) shows different similarity situations and their corresponding examples for operation sequence pairs that potentially occur in the operation sequence clustering problem. For a clear explanation, a target operation sequence is given first, and other operation sequences are designed based on different situations. Note that the numerical example cases shown in Figure 1 (b) are not the typical process routes that we mentioned above. In fact, they are used as examples to numerically illustrate the similarity situations described in Figure 1 (b).

Case 1 shown in Figure 1 (b) is the most dissimilar situation, and its sequence similarity with the target sequence is the lowest among 7 cases because there is no common information between case 1 and the target sequence. More specifically, there is no common or similar operation between them, let alone the precedence constraint relationship. In case 2, operation ''milling'' replaces ''grinding'' in case 1. Considered from the technological angle, ''turning'' and ''milling'' are both cutting processes that use cutting tools to cut and remove workpiece material (as shown in Figure 1 (a)), while ''grinding'' is the method to refine the surface of workpieces. Therefore, the operation similarity between ''turning'' and ''milling'' is higher than that between ''turning'' and ''grinding'', and case 2 is better than case 1. Compared with case 2, the operation sequence in case 3 shares one common operation with the target sequence. For case 4, all of its operations are the same as the target sequence, but the precedence relationship is totally different. In case 5, part of the sequence's precedence relationship is consistent with the target sequence, and part of the operations in the target sequence are replaced with similar operations. As shown in Figure 1 (b), operation ''turning'' in the target sequence is replaced with ''milling'' in case 5. In addition, the precedence relationship of ''baiting'', ''forging'' and ''spraying'' in case 5 is consistent with the target sequence, and only the position of ''turning''/''milling'' is different. Case 6 is better than case 5 because the precedence relationship of this situation has no difference from the target sequence, and ''milling'' is highly similar to ''turning'', as discussed before. Case 7 is the opposite of case 1 because all three kinds of information are the same for case 7 and the target sequence, which means that the similarity between the operation sequences in case 7 and the target sequence should be 1.

Based on the analysis above, the relationships of these 7 cases should be

$$s_1 < s_2 < s_3 < s_4 < s_5 < s_6 < s_7,$$

where s_i is the similarity between the operation sequence in case i and the target sequence.

For a better understanding of the similarity cases shown in Figure 1 (b), Figure 2 gives the common information degree of three information types between the target operation sequence and different cases, in which the operation similarity is also considered in the precedence relationship to clarify the subtle difference between case 5 and case 6.

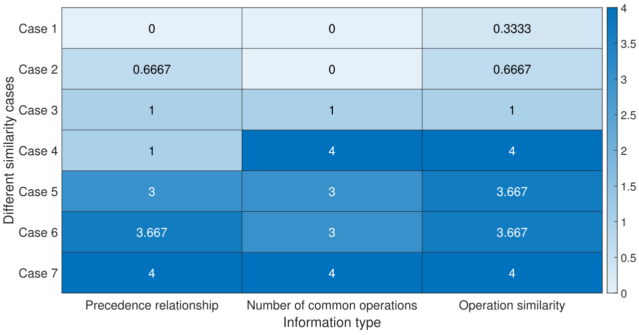


FIGURE 2. **Common information degree of three information types between the target operation sequence and different cases.**

It is noteworthy that the operation sequence similarity measure in the typical process route knowledge discovery problem is different from the same problem in the RMS [30], [33], [34] and CMS [5], [23]. The RMS and CMS are both used to manufacture a part family, and their operation sequences are slightly different [33]. Therefore, the number of idle machines and bypassing moves should be considered, which requires a strictly successive operation sequence with the same precedence constraint relationship. Typical process route knowledge discovery aims to find a universal process route for workpieces in such a way that it can reduce the time and labour cost for process planning and improve the efficiency and quality of process planning. Hence, succession relations are not key information in this problem.

IV. A NOVEL OPERATION SEQUENCE SIMILARITY COEFFICIENT

From the analysis above, an effective operation sequence similarity measure needs to cover all three kinds of information to comprehensively define the operation sequence similarity. This section proposes a modified pseudo-LCS similarity coefficient to measure the precedence constraint relationship and operation similarity simultaneously. Then, a corresponding backtracking algorithm is proposed to find the matching operation subsequence correctly for pseudo-LCS. The Jaccard similarity coefficient is also used in this paper to describe the information of common operations. Based on the pseudo-LCS and Jaccard similarity coefficient, a comprehensive similarity coefficient is then proposed by PCA. Figure 3 gives the hierarchical relationship between the required information and similarity measures.

A. SIMILARITIES BETWEEN OPERATIONS

There are many operations used in the manufacturing system, and their similarities are different from the point view of technology. Inspired by the work of Liu *et al.* [37], an operation code is introduced in this paper to represent the different operations. As shown in Figure 1 (a), each operation is represented by a unique three-digit code, and each digit of this code is denoted by numbers from 0 to 9. This operation

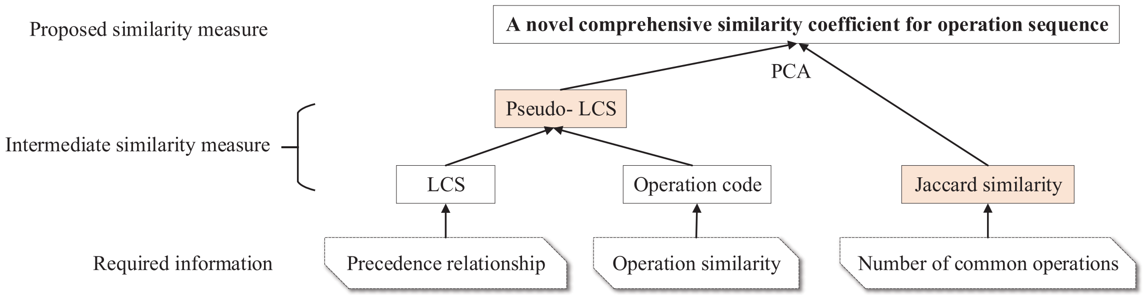

Proposed similarity measure
A novel comprehensive similarity coefficient for operation sequence
PCA
Pseudo- LCS
Intermediate similarity measure
LCS
Operation code
Jaccard similarity
Required information
Precedence relationship
Operation similarity
Number of common operations

FIGURE 3. Hierarchical relationship between the required information and similarity measures.

code essentially clusters operations based on their processing method, specific requirements, tool usage, and other machining-related information and uses a three-digit code representation to show common clustering information of operations from multiple granularities. The first digit is the roughest class, which mainly considers the processing method. The second digit shows a finer class than the first digit. For the second digit, information including the tool usage, specific requirements, and processing purposes are considered to define and distinguish the operations. The third digit is the finest class, and it clusters operations based on the processing surface, temperature requirement, deformation characteristics of workpieces, and so on.

In conclusion, the three-digit code representation of operations enables externalization of the machining-related information behind the text and makes it easier to measure the operation similarities. Table 1 shows the three-digit codes of process routes in Figure 1 (b).

TABLE 1. Three-digit codes of process routes in Figure 1 (b).

	Three-digit codes of process routes
Target	711-311-111-672
Case 1	321-513-515-359
Case 2	312-513-515-359
Case 3	311-513-515-359
Case 4	672-111-311-711
Case 5	711-111-672-312
Case 6	711-312-111-672
Case 7	711-311-111-672

Then, the similarities between different operations are defined based on the exclusive or (XOR) operation:

$$s_o(x, y) = 1 - \left(\sum_{m=1}^{3} |x_m \wedge y_m|\right)/3, \tag{2}$$

where $(\sum_{m=1}^{3} |x_m \wedge y_m|)/3$ is the distance between operations x and y, and x_m and y_m are the first m digit codes of operations x and y, respectively. When x_m and y_m are equivalent, $|x_m \wedge y_m| = 0$; when x_m and y_m are different, $|x_m \wedge y_m| = 1$.

According to this definition, the similarities between those operations in Figure 1 (b) are calculated and are shown in Table 2.

TABLE 2. Similarities between operations.

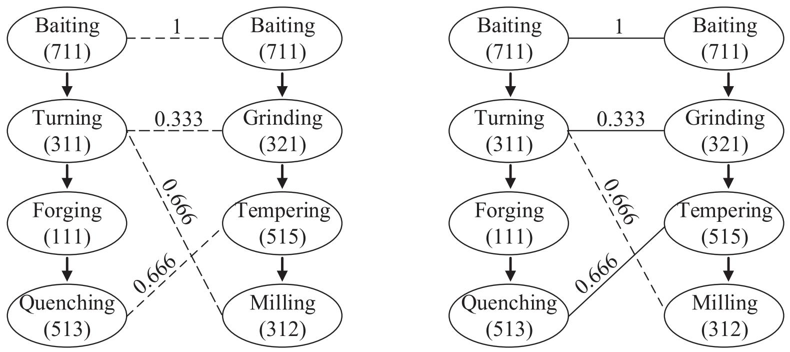
Operation	711	311	111	672	321	513	515	359	312
711	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
311	0.000	1.000	0.000	0.000	0.333	0.000	0.000	0.333	0.666
111	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
672	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
321	0.000	0.333	0.000	0.000	1.000	0.000	0.000	0.333	0.333
513	0.000	0.000	0.000	0.000	0.000	1.000	0.666	0.000	0.000
515	0.000	0.000	0.000	0.000	0.000	0.666	1.000	0.000	0.000
359	0.000	0.333	0.000	0.000	0.333	0.000	0.000	1.000	0.333
312	0.000	0.666	0.000	0.000	0.333	0.000	0.000	0.333	1.000

B. THE MODIFIED PSEUDO-LCS SIMILARITY COEFFICIENT

The LCS problem aims to find the longest common subsequence of two arbitrary strings. For example, both "CEF" and "CDF" are the longest common subsequences of strings "CDEF" and "CEDF". Wagner and Fischer [42] first proposed the classic dynamic programming method to solve the LCS problem. However, the traditional LCS problem considers only two cases, namely, identical characters and totally different characters, which suggests that the similarity between two characters has only two situations: 1 or 0.

As we mentioned before, the operation similarity is also a piece of key information that should be considered. In this paper, similarity between operations is seen as the weight between characters, and then, the traditional LCS problem becomes a weighted LCS problem, which attempts to find the subsequence pair with the maximal weighted sum. In the work of Lu [43], this maximal weighted subsequence pair is called pseudo-LCS because it is essentially an improved version of LCS. Figure 4 gives an example of finding the subsequence pair with the maximal weighted sum between two process routes, in which there is no connection (shown as the dashed line) between operations if the similarity of this operation pair is 0. As shown in Figure 4, there are two possible subsequence pairs: (baiting-turning-quenching and baiting-grinding-tempering) and (baiting-turning and baiting-milling). Their weighted sums are 2 and 1.666, respectively, which means that the maximal weighted subsequence pair of these two operation sequences is the former (see Figure 4 (b) with solid lines). This maximal weighted subsequence pair can be seen as the extension or variant of traditional LCS, and the weighted sum, therefore, is the pseudo-LCS similarity of these two operation sequences.

Lu [43] first considered the similarity between the characters in the LCS problem and improved the classical dynamic



(a) Similarity/Weight between operations (b) Subsequence pair with maximal weighted sum

FIGURE 4. **Weighted bipartite graph matching for two process routes.**

programming method to calculate the sentence similarity with Eq. (3), in which t is the threshold value.

$$c[i,j]=\begin{cases}0, & \\ & i=1\ or\ j=0\\ \max\{c[i-1,j-1]+sim(x,y), & \\ c[i-1,j],c[i,j-1]\}, & \\ & sim(x,y)>t\\ \max\{c[i-1,j],c[i,j-1]\}, & \\ & sim(x,y)\le t\end{cases}\tag{3}$$

In Eq. (3), the threshold value t is used to determine whether these two words are similar or not. If $sim(x,y) \le t$, then the common information embedded in $sim(x,y)$ will be ignored. When it comes to the operation sequence similarity problem, the smallest common information is critical for distinguishing different similarity cases. Hence, we define that all the operation pairs are similar and the difference relies on the degree of similarity in this paper. We set $t = 0$ so that no common information (i.e., similarity) will be ignored. Based on the previous work, the modified pseudo-LCS of two operation sequences can be obtained by Eq. (4).

$$c[i,j]=\begin{cases}0, & \\ & i=0\ or\ j=0\\ \max\{c[i-1,j-1]+s_o(x,y), & \\ c[i-1,j],c[i,j-1]\}, & \\ & s_o(x,y)\ge 0\end{cases}\tag{4}$$

Eq. (4) is essentially a special case of Eq. (3), in which the threshold value $t = 0$. Eq. (4) combines the last two situations into one, where $c[i,j] = \max\{c[i-1,j-1], c[i-1,j], c[i,j-1]\}$ when $s_o(x,y)$=0. It is based on the fact that $c[i-1,j-1] \le \max\{c[i-1,j], c[i,j-1]\}$, which can be easily proven by reduction ad absurdum.

However, when we attempt to find the pseudo-LCS of operation sequences, the backtracking method that Lu [43] proposed is not quite correct. They simply thought that if $c[i,j] > c[i-1,j-1]$, then x_i and y_j are matching operations. Taking $x = [711\ 311\ 111\ 672]$ and $y = [711\ 310\ 672\ 670]$ as an example, Figure 5 shows the result obtained according to their method.

TABLE 3. **Pseudo-LCS similarity coefficient between operation sequences in Figure 1 (b).**

	Target	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Target	1.0000	0.0833	0.1667	0.2500	0.2500	0.7500	0.9167	1.0000
Case 1	0.0833	1.0000	0.8333	0.8333	0.0833	0.0833	0.0833	0.0833
Case 2	0.1667	0.8333	1.0000	0.9167	0.1667	0.2500	0.2500	0.1667
Case 3	0.2500	0.8333	0.9167	1.0000	0.2500	0.1667	0.1667	0.2500
Case 4	0.2500	0.0833	0.1667	0.2500	1.0000	0.4167	0.2500	0.2500
Case 5	0.7500	0.0833	0.2500	0.1667	0.4167	1.0000	0.7500	0.7500
Case 6	0.9167	0.0833	0.2500	0.1667	0.2500	0.7500	1.0000	0.9167
Case 7	1.0000	0.0833	0.1667	0.2500	0.2500	0.7500	0.9167	1.0000

Figure 5 indicates that the pseudo-LCS of x and y is 711-311-672 and 310-672-670, respectively, but the real result is 711-311-672 and 711-310-672. To address this problem, Algorithm 1 is proposed to find the right pseudo-LCS.

Figure 6 shows the right matrix and trace based on Algorithm 1. The pseudo-LCS of x and y in Figure 6 is 711-311-672 and 711-310-672, respectively, which is consistent with the practical situation. Numerical illustration shows that this algorithm is also workable for operation sequence pairs with multiple pseudo-LCSs.

According to the analysis above, the pseudo-LCS similarity coefficient between two operation sequences is calculated by Eq. (5). This definition contains information about the precedence relationship and operation similarity simultaneously. Specifically, pseudo-LCS, namely, the maximal weighted subsequence pair, reflects the precedence relationship, and pseudo-LCS similarity is calculated based on the operation similarity.

$$s_s(x,y) = sim_{p-LCS} / \max(|x|\,, |y|),\tag{5}$$

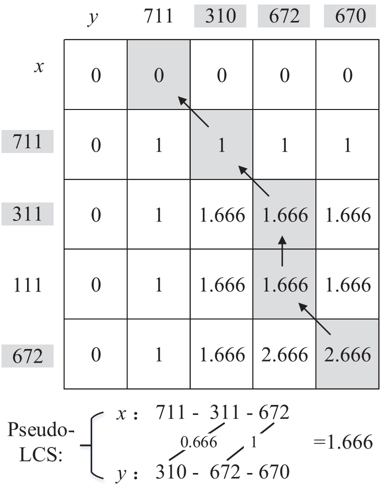
where sim_{p-LCS} is the weighted sum of pseudo-LCS, namely, the maximal value in the matrix of Figure 6. Here, $|x|$ and $|y|$ are the lengths of the operation sequences x and y, respectively. Then, the pseudo-LCS similarity matrix between the operation sequences in Figure 1 (b) is shown in Table 3.

C. A NOVEL COMPREHENSIVE SIMILARITY COEFFICIENT

From Table 3, it can be seen that the pseudo-LCS similarity coefficient cannot distinguish case 3 and case 4 because their

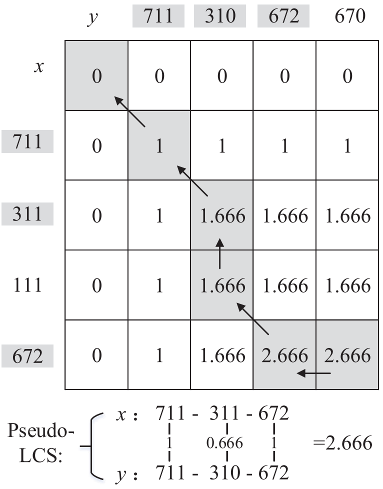
Algorithm 1 A Novel Backtracking Algorithm to Find the Pseudo-LCS

```
Let $i$ and $j$ start from $m$ and $n$, respectively, where $m$ and $n$ represent the length of two process routes;
while $i > 0$ or $j > 0$ do
    if $s_o(x_i, y_j) = 1$ then
        $x_i$ and $y_j$ are matching operations, $i = i - 1, j = j - 1$;
    else if $1 > s_o(x_i, y_j) > 0$ then
        find the maximum value between $c[i-1, j-1] + s_o(x, y)$, $c[i-1, j]$ and $c[i-1, j-1]$;
        if $c[i-1, j-1] + s_o(x, y)$ is maximal, then $i = i - 1, j = j - 1$ and $x_i, y_j$ are matching operations;
        if $c[i-1, j]$ is maximal, then $i = i - 1, j = j$;
        if $c[i-1, j-1]$ is maximal, then $i = i, j = j - 1$;
    else if $s_o(x_i, y_j) = 0$ then
        find the maximum value between $c[i-1, j]$ and $c[i-1, j-1]$;
        if $c[i-1, j]$ is maximal, then $i = i - 1, j = j$;
        if $c[i-1, j-1]$ is maximal, then $i = i, j = j - 1$;
```



	y	711	310	672	670
x	0	0	0	0	0
711	0	1	1	1	1
311	0	1	1.666	1.666	1.666
111	0	1	1.666	1.666	1.666
672	0	1	1.666	2.666	2.666

FIGURE 5. Matrix and trace of *x*, *y* by Lu's method.



	y	711	310	672	670
x	0	0	0	0	0
711	0	1	1	1	1
311	0	1	1.666	1.666	1.666
111	0	1	1.666	1.666	1.666
672	0	1	1.666	2.666	2.666

FIGURE 6. Matrix and trace by the novel backtracking algorithm.

pseudo-LCS similarity coefficients are both 0.25. The reason is that the pseudo-LCS similarity coefficient strictly requires the precedence constraint relationship of sequences. For the operation sequence in case 4, the precedence constraint relationship is totally different from the target sequence, although their operations are identical. In this case, there are four pseudo-LCSs (i.e., spraying, forging, turning, and baiting) between case 4 and the target, and their pseudo-LCS similarities are all 1, which are exactly the same pseudo-LCS similarities as in case 3 and the target. The main difference between case 4 and case 3 is the number of pseudo-LCSs because case 3 only shares one pseudo-LCS (i.e., turning) with the target.

To address this drawback, the Jaccard similarity coefficient originally proposed by Paul Jaccard [44] is also introduced in this paper because it focuses on only the number of common operations.

However, the pseudo-LCS similarity coefficient itself already contains partial information about common operations because pseudo-LCS reflects the identical precedence of similar operations, which includes identical operations. In other words, these two similarity coefficients are correlative. Taking the operation sequences in Table 1 as an example, the correlation coefficient of these two indexes is 0.6706. The relationship between these two similarity coefficients is shown in Figure 7.

Because it is difficult to quantify the overlapping portion in Figure 7, PCA is used to obtain a comprehensive similarity coefficient based on these two similarity coefficients, considering them equally. The greatest strength of PCA is that it can generate several independent indexes (also called principal components) based on some relevant coefficients. These principal components contain most of the information of original coefficients, and the duplicate information can only be counted once. Therefore, the PCA-based analysis of the pseudo-LCS similarity coefficient s_s and Jaccard similarity coefficient s_j is shown in Table 4.

Since there are only two principal components in this operation sequence similarity problem and the first principal component calculated by PCA contains 89.48% information

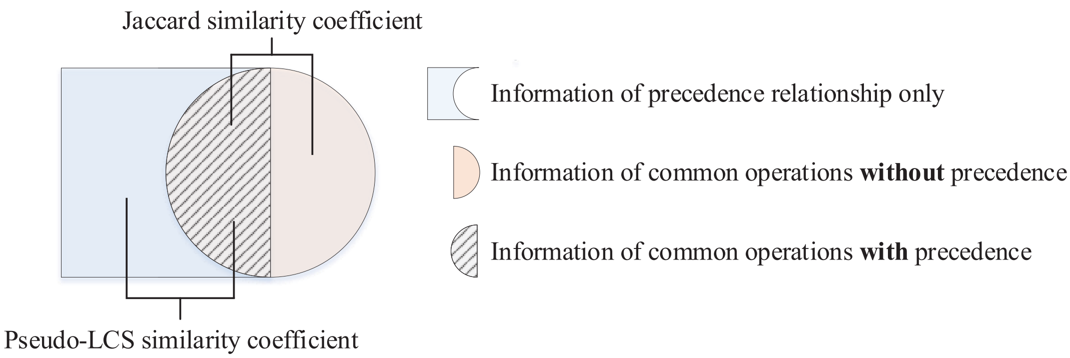

Jaccard similarity coefficient
Information of precedence relationship only
Information of common operations **without** precedence
Information of common operations **with** precedence
Pseudo-LCS similarity coefficient

FIGURE 7. Relationship of two similarity coefficients.

TABLE 4. Calculation coefficient and cumulative contribution proportions of principle components.

Principal component	s_s	s_j	Eigenvalue	Proportion (%)
y_1	0.7071	0.7071	1.7896	89.48%
y_2	-0.7071	0.7071	0.2104	10.52%

of s_s and s_j, it is chosen as the comprehensive similarity coefficient to measure the similarity between two operation sequences. Hence, the normalized comprehensive similarity matrix between operation sequences in Figure 1 (b) is given in Table 5.

It is noteworthy that a comprehensive similarity coefficient of 0 does not mean the worst similarity case that $s_j = s_s = 0$ in Table 5. Analogously, a comprehensive similarity coefficient of 1 does not mean the best similarity case that $s_j = s_s = 1$. It only represents the worst and best situations in the operation sequence data set. As shown in Figure 1 (b), case 1 is not the worst case in the actual situation because $s_o(turing, grinding) = 0.333$. Hence, $s_s(Target, Case\ 1) = 0.0833$. Since case 1 is the worst case in Figure 1 (b), the normalized comprehensive similarity coefficient is 0. Table 5 shows that the novel comprehensive similarity coefficient can distinguish all the different cases in Figure 1 (b), and the calculated result is also consistent with reality.

V. MODIFIED K-MEDOIDS METHOD FOR TYPICAL PROCESS ROUTE KNOWLEDGE DISCOVERY

The quantity soft constraint and size soft constraint are necessary for typical process route knowledge discovery because they essentially reflect the granularity requirement of discovered typical process routes. Technically, the clustering granularity directly affects the quality and contains information about the discovered typical process routes. These two constraints are seen as soft constraints in this paper because it is difficult for technicians to set parameters perfectly and because soft constraints are less sensitive and more tolerated than hard constraints. However, introducing these two soft constraints into the ordinary typical process route discovery problem makes it more complicated to solve.

In this section, the K-medoids method, a widely used exemplar-based clustering algorithm, is modified to meet the quantity soft constraint and size soft constraint simultaneously. For a better explanation, we first give the mathematical description of the typical process route knowledge discovery problem with two soft constraints. Then, the modified algorithm and its pseudocodes are presented.

TABLE 5. Normalized comprehensive similarity coefficients.

	Target	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Target	1.0000	0.0000	0.0460	0.1626	0.5860	0.6644	0.7564	1.0000
Case 1	0.0000	1.0000	0.7104	0.7104	0.0000	0.0000	0.0000	0.0000
Case 2	0.0460	0.7104	1.0000	0.7564	0.0460	0.1626	0.1626	0.0460
Case 3	0.1626	0.7104	0.7564	1.0000	0.1626	0.0460	0.0460	0.1626
Case 4	0.5860	0.0000	0.0460	0.1626	1.0000	0.4804	0.3884	0.5860
Case 5	0.6644	0.0000	0.1626	0.0460	0.4804	1.0000	0.8620	0.6644
Case 6	0.7564	0.0000	0.1626	0.0460	0.3884	0.8620	1.0000	0.7564
Case 7	1.0000	0.0000	0.0460	0.1626	0.5860	0.6644	0.7564	1.0000

A. MATHEMATICAL DESCRIPTION OF THE TYPICAL PROCESS ROUTE KNOWLEDGE DISCOVERY PROBLEM

Given a process route set $\mathcal{X} = \{x_1, \cdots, x_N\}$, let $\mathcal{S} = \left\{s\left(x_i, x_j\right)\right\}$ be the similarity set of all process route pairs, and $s\left(x_i, x_j\right)$ is calculated based on the novel comprehensive similarity coefficient proposed in Section IV-C. Let $q_e \geq 0$ be the penalty for the quantity constraint and let $q_r \geq 0$ be the penalty for the size constraint. Then, the goal of the clustering algorithm is to find the exemplar set $\mathcal{O} = \{o_1, \cdots, o_K\}$ that maximizes the following objective function:

$$\underset{\mathcal{O}}{arg\,max} \sum_{i=\{1,2,\cdots,N\}} s\left(x_i, o_{x_i}\right) - q_e\left|\mathcal{X}_e\right| - \sum_{x_i\in\mathcal{X}_r}\left(\frac{\varepsilon - s\left(x_i, o_{x_i}\right)}{\varepsilon}q_r\right), \tag{6}$$

where o_{x_i} is the exemplar of the ith process route x_i, $\mathcal{X}_e$ and $\mathcal{X}_r$ are the sets of process routes that violate the quantity constraint and size constraint, respectively, and ε is the maximum radius of a valid process route cluster.

B. MODIFIED K-MEDOIDS METHOD CONSIDERING TWO SOFT CONSTRAINTS

The K-medoids method is a classic approach to finding the K exemplars in the clustering problem. To ensure that the clustering result satisfies the soft constraints mentioned above, a modified K-medoids method with an adjustment mechanism is proposed. The modified K-medoids method for typical process routes is shown in Algorithm 2.

Algorithm 2 A Modified K-Medoids Method for Typical Process Routes

```
Randomly select $K$ process routes as the initial medoids and assign other process routes to the nearest medoid with the
formula shown below, which considers the similarity and size constraint simultaneously;

    $s_{ij}^* = s\left(x_i, o_j\right) + \min\left[-\frac{\varepsilon - s\left(x_i, o_j\right)}{|\varepsilon|} q_r, 0\right]$

while the objective value (see Eq. (6)) of the new clustering result is better than that of the old result do
    Find $\mathcal{O}^L$ and $\mathcal{O}^F$, which are the set of unqualified clusters that violate the quantity constraint and the set of qualified
    clusters in which the number of process routes is larger than $n$;
    // Adjust the clustering result
    while $\mathcal{O}^L$ and $\mathcal{O}^F$ are not empty do
        for $i = 1: \left|\mathcal{O}^L\right|$ do
            choose an unqualified cluster $O_i^L$ from $\mathcal{O}^L$ with the corresponding medoid $o_i^L$;
            select a process route $x$ that is closest to $o_i^L$ according to $s_{xo_i^L}^*$ from each cluster in $\mathcal{O}^F$, and then, obtain a
            process route queue with $\left|\mathcal{O}^F\right|$ process routes;
            sort this queue in descending order based on $s_{xo_i^L}^*$ ;
            while queue is not empty do
                choose the first process route in this queue;
                if $s_{xo_j^F}^* - s_{xo_i^L}^* < q_e$, where $o_j^F$ is the original medoid of $x$ then
                    move this process route to the unqualified cluster $O_i^L$;
                    break;
                else if $s_{xo_j^F}^* - s_{xo_i^L}^* \geq q_e$ then
                    delete this process route from this queue;
        foreach $O_i^L \in \mathcal{O}^L$ do
            delete $O_i^L$ from $\mathcal{O}^L$ if $\left|O_i^L\right| = n$ or $O_i^L$ obtains nothing from the queue;
        foreach $O_j^F \in O^F$ do
            delete $O_j^F$ from $\mathcal{O}^F$ if $\left|O_i^F\right| = n$ ;
    // End of adjustment
    Find a new medoid for each cluster according to $s_{ij}^*$ and form a new clustering result based on these new medoids;
```

In Algorithm 2, the adjustment mechanism starts after a new clustering result is generated in each generation because the new clustering result is obtained under the consideration of similarity and size constraint only. Therefore, the main idea and purpose of the adjustment mechanism are to reassign process routes based on the quantity constraint and its penalty q_e. In this way, the operation sequence similarity and two soft constraints are all considered during the clustering, which ensures the validation of the final clustering results and the discovery of typical process routes.

VI. NUMERICAL ILLUSTRATION

In this section, the performance of the proposed comprehensive similarity coefficient and the modified K-medoids method is evaluated. In addition, related parameters of the proposed K-medoids method are analysed.

A. PERFORMANCE OF THE PROPOSED COMPREHENSIVE SIMILARITY COEFFICIENT

The sequence similarity measure problem is a well-known problem in many fields. To evaluate the performance of the proposed comprehensive similarity coefficient, it has been compared with 10 other existing similarity coefficients of operation sequences in the literature. In this section, two process route sets are chosen as the test sets, i.e., the process routes shown in Table 1 and the process routes obtained from the literature [37]. Here, Eq. (1) is used as a performance measure to describe the goodness of the similarity coefficients. A similarity coefficient having a larger R value means that it has better performance.

1) COMPARISON BASED ON PROCESS ROUTES IN TABLE 1

The process routes shown in Table 1 have considered all the information required in the typical process route discovery problem; hence, they reflect all the potential similarity cases

TABLE 6. Comparison of different similarity coefficients based on process routes in Table 1.

Coefficients	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	R
Choobineh [27] (1988)	**0.0000**	**0.0000**	0.0625	0.2500	**0.2708**	**0.2708**	1.0000	0.4286
Tam [28] (1990)	**0.0000**	**0.0000**	0.0715	0.5000	0.5500	0.6750	1.0000	0.7143
Ho et al. [29] (1993)	**0.0000**	**0.0000**	**0.2500**	**0.2500**	**0.7500**	**0.7500**	1.0000	0.1429
Askin and Zhou [30] (1998)	**0.0000**	**0.0000**	**0.2500**	**0.2500**	**0.7500**	**0.7500**	1.0000	0.1429
Irani and Huang [31] (2000)	**0.2000**	**0.2000**	**0.4000**	**0.4000**	**0.8000**	**0.8000**	1.0000	0.1429
Huang [32] (2003)	**0.0000**	**0.0000**	**0.2500**	**0.2500**	0.7500	0.6875	1.0000	0.4286
Liu et al. [37] (2007)	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.1340	0.8333	1.0000	0.4286
Goyal et al. [33] (2013)	**0.0556**	**0.0556**	0.4286	0.5714	0.7167	0.6333	1.0000	0.7143
Wang et al. [34] (2016)	**0.0000**	**0.0000**	0.2000	0.2500	0.6666	0.6000	1.0000	0.7143
Wang et al. [22] (2016)	0.0417	0.0833	**0.1250**	**0.1250**	0.3750	0.4583	**0.5000**	0.5714
Proposed coefficient	0.0000	0.0460	0.1626	0.5860	0.6644	0.7564	1.0000	1.0000

in real manufacturing for typical process route discovery. Table 6 shows the calculated operation sequence similarities between 7 similarity cases and the target based on 11 similarity coefficients, and the bold results highlight the cases that this coefficient cannot distinguish.

In Choobineh's similarity coefficient [27], $L = 4$ because it is the maximum length of the longest common substring of all the operation sequence pairs in Table 1 (i.e., the pair of target operation sequence and case 7). In Tam's dissimilarity coefficient [28], $w_n = w_c = 0.5$, and $w_s = w_d = w_i = 1$. For a better comparison, Tam's dissimilarity coefficient is converted to a similarity coefficient by subtracting it from 1. In Wang's similarity coefficient [22], $\omega = 0.5$.

Based on Table 6, it can be seen that the proposed coefficient performs the best, with the highest R value, which means that it can effectively distinguish all the operation sequence pairs in Table 1 and assign rational similarity values as the analysis in Section III-B. In contrast, the other existing similarity coefficients are too coarse to find the subtle difference between 7 similarity cases.

As shown in Table 6, similarity coefficients in the literature [27], [28], [29], [30], [31], [32], [33] and [34] ignore the similarity between the operations themselves, which makes it difficult to identify the tiny difference between case 1 and case 2 and to measure the process-related information hidden behind the operations. Although Liu's and Wang's similarity coefficients considered this information, Liu *et al.* [37] used the Euclidean distance to measure the distance between the process routes, which essentially did not consider the precedence relationship, and Wang *et al.* [22] focused on only the precedence relationship between 2-operation substrings and did not consider these two pieces of information comprehensively. Therefore, the proposed similarity measurement is more reasonable and conforms to reality.

2) COMPARISON BASED ON PROCESS ROUTES FROM THE LITERATURE [37]

According to the literature [37], the process routes shown in Table 7 are obtained from a real manufacturing enterprise and are used to machine axle sleeve parts. There are two characteristics of these process routes: (1) their lengths are different; and (2) the process route can contain several identical operations. Four example sets are generated based on these 8 process routes to evaluate the performance of different similarity coefficients. Note that the operation similarity has less influence on the performance in this section than in Section VI-A1, although these process routes are also represented as three-digit codes.

TABLE 7. Process routes after coding from the literature [37].

No.	Part No.	Part name	Operation						
			1	2	3	4	5	6	7
x_1	J2.P5CK100-1	Screw	012	311	311	312	810	611	000
x_2	F121005-6	Guiding post	012	311	311	312	810	000	000
x_3	F122040-1	Plug screw	012	311	311	312	312	810	611
x_4	F122011-3	Screw-thread bush	012	311	311	312	810	611	000
x_5	YH451220-4	Orifice sleeve	012	311	810	000	000	000	000
x_6	F851342-4	Screw 2	012	311	311	312	810	653	000
x_7	GZ11100-40	Pressure block	012	311	312	311	810	000	000
x_8	P128421	Spread segment	012	311	311	311	311	810	000

Table 8 shows the comparison results of different similarity coefficients, in which the bold results are the limitations of the corresponding similarity coefficients. Here, $L = 6$ for Choobineh's similarity coefficient [27], and the other parameters are the same as those in the previous section.

As shown in Table 8, the similarity coefficient proposed in the literature [31], [32], [33], [34] and the proposed similarity coefficient all have the best performance in this process route data set. The performance index R of other similarity coefficients mainly suffers from the disadvantages of being less sensitive (e.g., most of the bold results) and illogical similarity values (e.g., example set 1 of reference [22]).

From Table 8, it can be seen that most of the existing coefficients cannot distinguish the tiny difference between process route pairs, especially example set 3, in which x_5 itself is the longest common subsequence. In this case, the information of another process route in the process route pair, namely, x_4, x_7 and x_8, plays a dominant role in the similarity measure. Another interesting observation from the table is that the coefficient proposed by Wang *et al.* [22] cannot ensure that the similarity value between the same process routes is the same, as shown in example set 1. This finding can be explained by the mechanism of their coefficient. Although Wang *et al.* [22] considered the similarity between the operations, the process route similarity is calculated based on the adjacency matrix. For two identical process routes, the number of -1 in the structural comparison matrix is determined by the adjacency relationship of the operations. Therefore, the similarity values of different process route pairs in example set 1 vary with the process route itself. In this section, the proposed similarity coefficient also shows good performance with the maximal

TABLE 8. Comparison of different similarity coefficients based on process routes from literature [37].

Coefficients	Example set 1				Example set 2			Example set 3			Example set 4		R
	$\{x_1, x_1\}$	$\{x_1, x_4\}$	$\{x_2, x_2\}$	$\{x_8, x_8\}$	$\{x_3, x_1\}$	$\{x_3, x_2\}$	$\{x_3, x_6\}$	$\{x_5, x_4\}$	$\{x_5, x_7\}$	$\{x_5, x_8\}$	$\{x_8, x_6\}$	$\{x_8, x_7\}$	
Choobineh [27]	1.00	1.00	1.00	1.00	0.49	0.49	0.38	0.25	**0.33**	**0.33**	0.26	0.18	0.83
Tam [28]	1.00	1.00	1.00	1.00	0.88	0.65	0.58	0.43	**0.63**	**0.63**	0.43	0.63	0.83
Ho et al. [29]	1.00	1.00	1.00	1.00	0.86	**0.71**	**0.71**	**0.50**	0.60	**0.50**	**0.67**	**0.67**	0.50
Askin and Zhou [30]	1.00	1.00	1.00	1.00	**1.00**	**1.00**	0.83	**1.00**	**1.00**	**1.00**	**0.67**	**0.67**	0.42
Irani and Huang [31]	1.00	1.00	1.00	1.00	0.98	0.98	0.84	0.92	0.85	0.88	0.83	0.81	1.00
Huang [32]	1.00	1.00	1.00	1.00	0.97	0.96	0.81	0.88	0.89	0.84	0.81	0.77	1.00
Liu et al. [37]	1.00	1.00	1.00	1.00	**0.35**	**0.35**	**0.35**	**0.24**	0.35	**0.24**	**0.45**	**0.45**	0.42
Goyal et al. [33]	1.00	1.00	1.00	1.00	0.92	0.79	0.70	0.58	0.63	0.54	0.58	0.60	1.00
Wang et al. [34]	1.00	1.00	1.00	1.00	0.86	0.77	0.67	0.55	0.60	0.50	0.53	0.57	1.00
Wang et al. [22]	**0.52**	**0.52**	**0.53**	**0.36**	0.49	0.42	0.37	0.36	0.45	0.40	0.30	0.35	0.67
Proposed coefficient	1.00	1.00	1.00	1.00	0.86	0.50	0.41	0.07	0.33	0.50	0.35	0.51	1.00

R for the real process route set, and it can effectively assign suitable similarity values for different process route pairs.

From Tables 6 and 8, it can be seen that although the similarity coefficients proposed in the literature [31], [32], [33], and [34] perform well in Table 8, they all have limitations in Table 6. Comprehensively, the proposed coefficient is more general than other existing coefficients because it can handle all the similarity situations in both data sets.

B. PERFORMANCE ANALYSIS OF THE MODIFIED K-MEDOIDS METHOD

The typical process route discovery problem is a practical problem in which the shape of a process route set is unknown. To comprehensively analyse the performance of the modified K-medoids method, which is represented as mKM in this section, two process route data sets are generated randomly to test the performance of the proposed algorithm.

To verify the effectiveness of the modified K-medoids method, especially the adjustment mechanism proposed in Section V-B, the original K-medoids method (represented as oKM) and ALC [34], [36] are also tested in this section. In the previous work, they did not consider the size constraint and quantity constraint as soft constraints during the clustering. These two constraints are often treated as hard constraints and are used to find valid clusters after clustering. In this way, the role of the adjustment mechanism proposed in this paper can be evaluated clearly.

1) PERFORMANCE MEASURE

Typically, the main task of clustering is to ensure that data points in the same cluster are as close to each other as possible, while the distance between different clusters should be as far as possible [45]. However, since process route clustering is a practical problem, it is impossible to know whether the data set is spherical or arbitrary, in which case most of the existing internal clustering validation might not be suitable for this problem.

Because the goal of this paper is to find the typical process routes, the sum of the similarities between process routes and their exemplars and the penalties of two soft constraints are chosen as the performance measure in this section, as shown in Eq. (6).

2) PERFORMANCE ANALYSIS BASED ON PROCESS ROUTE SETS

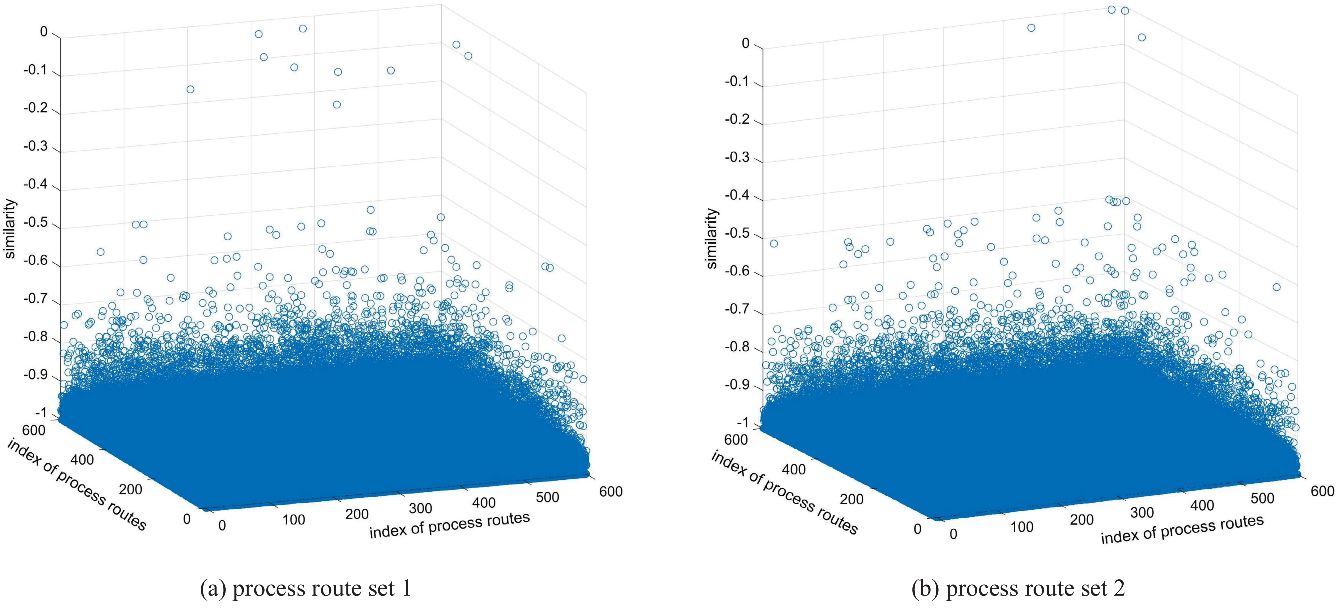
In this section, two process route sets with 600 process routes represented by three-digit codes are generated to evaluate the performance of clustering algorithms. Each digit code of the operation is designed from a discrete uniform distribution between 000 and 999. The maximum length of the process routes is 6, and their lengths also obey a discrete uniform distribution.

Figure 8 visualizes the similarities of two process route sets, in which the similarities are all minus 1, i.e., -1, which means the worst situation, while 0 is the best. In this way, 0 is the ideal clustering result, which can be seen as a visualized standard line in the figures. It is observed that most of the similarity values are less than -0.5. This finding arises because the proposed similarity measure takes a large amount of information into consideration (especially the precedence relationship), which makes it difficult for randomly generated process routes to obtain high similarity values. Process route pairs with maximal similarity, i.e., zero, all have only one operation, in which case the precedence relationships are identical. Because of this kind of distribution, size constraints can have less influence on process route sets, as most process route pairs have similar distances. Parameters in this section are set to be the following:

$$n = round(N/K), \quad \varepsilon = min\ s(x_i,\ x_j)/K,$$
$$q_e = 0.5, \quad q_r = 1.$$

where n is the minimum number of process routes in a valid process route cluster for the quantity constraint, ε is the maximum radius of a valid process route cluster for the size constraint, and q_e and q_r are the penalties for the quantity constraint and size constraint, respectively.

The parameter settings of n and ε ensure that each K is the optimal choice for the current constraint situation. Among these four parameters, q_e and q_r are designed based on the minimum similarity (i.e., -1) and the calculation methods of two soft constraint penalties. Specifically, according to Eq. (6), the penalty of the quantity constraint is determined directly by the number of process routes that violate the quantity constraint, and the penalty of the size constraint is determined by the ratio between the similarity of the process route to its exemplar and ε. Hence, the penalty of the quantity



(a) process route set 1

(b) process route set 2

FIGURE 8. Similarity distributions of two process route sets.

constraint is always larger than or equal to the penalty of the size constraint for the same process route when $q_e = q_r$. To prevent penalties from dominating the clustering performance, q_e and q_r should be less than the maximum distance of the process route pairs, and $q_e < q_r$ for the same reason.

For the numerical illustration design, two K-medoids methods, i.e., mKM and oKM, were rerun 3×10^4 times for each run with different K because of the undetermined shape of the two process route sets, which makes them more difficult to solve. Most data point pairs in process route data sets shown in Figure 8 share a similar pattern since most similarities are less than -0.5. In this case, the process route sets have C_{600}^{20} potential exemplar sets, while the traditional spherical data set only has approximately C_{400}^{20} or C_{300}^{20} potential exemplar sets when the sizes of the data sets are both 600 and $K = 20$. The reason is that the data points near the edge and centre area are hardly exemplars when $K = 20$. However, this problem does not exist for process route sets because they do not have a clear edge and centre area. As shown in Figure 8, every process route can be an exemplar because the similarities between each process route to all the other routes are too close. This feature is the most important feature for process route sets. In fact, we found that the average numbers of iterations for the two process route sets are 1.81 and 1.83 when $K = 20$. This finding indicates that the solution spaces of the process route sets have many local optima, in which case the K-medoids method often reaches the stop criteria easily and improves little from the initial solution. Hence, 3×10^4 reruns are necessary for mKM and oKM in this section.

Figure 9 shows the overall clustering results of the three clustering methods. This figure indicates that the modified K-medoids method performs best among the three methods and that their performance gap widens as K increases. The beam-like pattern in Figure 9 indicates that the clustering performance worsens as K increases. This finding is explainable because the penalty of the soft size constraint dominates the final performance for the process route sets. More specifically, the performance of the clustering result depends on three parts, as shown in Eq. (6): the similarity sum of each process route to its exemplar, the penalties of the quantity constraint and the size constraint. Because of the special shape of the process route sets, as we discussed above, the increase in cluster number K has little influence on the similarity sum and the penalty of the quantity constraint. Different choices of exemplars cannot bring significantly higher similarities for most process routes. For the penalty of the quantity constraint, the increasing K makes both n and the number of process routes in each cluster decrease. The penalty of the size constraint, however, increases as K increases. The reason is that ε decreases gradually, while the size of each cluster (which is determined by the farthest process route in the cluster) changes slightly since the similarity of any process route pair is close, as shown in Figure 8. Hence, the performance of the clustering results decreases linearly for the process route sets.

Tables 9 and 10 show the mean and standard deviation performance (calculated by Eq. (6) in Section V-A) of the clustering algorithms on two process route sets, respectively. The bold results in these two tables indicate that this result is significantly better than others based on Wilcoxon's rank sum test (5% significance level) [46].

The most obvious observation from Tables 9 and 10 is that the proposed K-medoids method (i.e., mKM) is significantly better than the other two clustering methods in most cases, although the improvement is not very obvious compared with that of oKM. It can also be seen that the similarity standard deviations of both mKM and oKM in Tables 9 and 10 are very

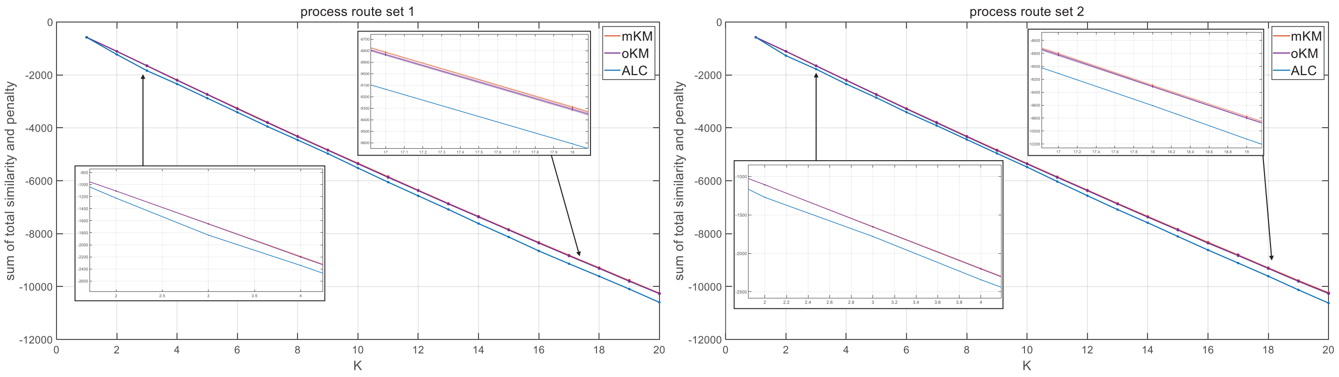


FIGURE 9. Overall performances and changing trends of clustering algorithms on process route data sets.

small. This finding further proves that the local optimums of the Kth solution space are close. This observation is also consistent with the previous conclusion that the improvement spaces of the clustering results for the process route sets are small because the average number of iterations is small. Considering this situation, a small improvement in mKM is sufficient to show the effectiveness of the adjustment mechanism proposed in Section V-B.

C. ANALYSIS OF RELATED PARAMETERS

This section analyses how the parameter setting of mKM affects the clustering performance, including the data size and related parameters of two soft constraints.

1) ANALYSIS OF THE INFLUENCE OF THE DATA SET SIZE

The performance analyses in Section VI-B have shown that the shape of the data set affects the performance of clustering algorithms. The reason is that the shape of the data set directly affects the shape of the solution space, which has a great impact on the difficulty of solving the clustering problem.

In this section, the relationship between the performance of the modified K-medoids method and the data set size is analysed based on the process route sets mentioned in Section VI-B2. Several process routes are randomly selected from the original data sets to form new data sets with different sizes. The parameter settings are the same as those in Section VI-B2, and the result is shown in Figure 10.

Figure 10 shows a similar pattern of performance changes as in Figure 9, in which the average performance values of different data sets also have similar behaviours. Concretely, these performance values exhibit an obvious linear decreasing trend when K increases. This finding is consistent with Figure 9, which shows similar decreasing trends for the three algorithms.

Regarding the influence of the data set size, Figure 10 suggests that a larger data set means a larger decreasing slope of the performance value. Under this situation, the algorithm

TABLE 9. The mean and standard deviation performances of clustering algorithms (*process route set* 1).

K	mKM	oKM	ALC	K	mKM	oKM	ALC
1	-578.1 (0.0)	-578.1 (0.0)	-578.1 (0.0)	11	**-5848.6 (9.2)**	-5869.1 (8.9)	-6052.6 (0.0)
2	**-1106.7 (0.0)**	-1107.0 (0.0)	-1226.0 (0.0)	12	**-6358.8 (4.2)**	-6372.0 (5.0)	-6569.5 (0.0)
3	**-1648.9 (0.0)**	-1654.1 (1.5)	-1839.1 (0.0)	13	**-6859.1 (2.5)**	-6874.7 (6.9)	-7084.0 (0.0)
4	**-2191.3 (1.3)**	-2201.8 (2.6)	-2341.8 (0.0)	14	-7349.8 (7.9)	-7363.8 (11.9)	-7617.3 (0.0)
5	**-2730.0 (1.3)**	-2738.7 (2.1)	-2878.4 (0.0)	15	**-7842.2 (5.0)**	-7858.7 (3.4)	-8125.9 (0.0)
6	**-3260.4 (4.0)**	-3276.9 (2.9)	-3411.1 (0.0)	16	-8332.7 (7.1)	-8351.7 (10.6)	-8651.9 (0.0)
7	**-3789.4 (2.4)**	-3805.3 (2.4)	-3950.1 (0.0)	17	**-8815.1 (4.5)**	-8835.5 (7.0)	-9135.9 (0.0)
8	**-4315.8 (3.1)**	-4330.6 (5.2)	-4458.8 (0.0)	18	**-9291.3 (5.4)**	-9310.4 (8.5)	-9609.4 (0.0)
9	**-4831.6 (5.2)**	-4845.0 (3.9)	-4972.2 (0.0)	19	-9774.6 (13.8)	-9794.8 (11.3)	-10093.0 (0.0)
10	**-5338.8 (3.9)**	-5357.8 (6.0)	-5518.7 (0.0)	20	-10252.4 (5.3)	-10265.8 (9.7)	-10596.0 (0.0)

TABLE 10. The mean and standard deviation performances of clustering algorithms (*process route set* 2).

K	mKM	oKM	ALC	K	mKM	oKM	ALC
1	-577.9 (0.0)	-577.9 (0.0)	-577.9 (0.0)	11	**-5854.0 (5.3)**	-5870.5 (5.2)	-6028.5 (0.0)
2	-1108.0 (0.0)	-1108.0 (0.0)	-1271.8 (0.0)	12	**-6353.7 (4.2)**	-6372.6 (6.6)	-6564.4 (0.0)
3	**-1651.1(0.2)**	-1656.2 (1.0)	-1780.2 (0.0)	13	**-6858.4 (0.9)**	-6875.9 (1.7)	-7089.0 (0.0)
4	**-2195.7 (0.9)**	-2202.8 (1.6)	-2343.0 (0.0)	14	**-7349.5 (6.4)**	-7371.8 (2.8)	-7586.4 (0.0)
5	**-2734.2 (1.9)**	-2744.7 (1.8)	-2859.7 (0.0)	15	**-7840.3 (7.1)**	-7859.0 (8.2)	-8104.8 (0.0)
6	**-3270.1 (2.4)**	-3283.6 (1.7)	-3409.8 (0.0)	16	**-8328.6 (7.1)**	-8355.0 (2.8)	-8617.6 (0.0)
7	-3793.2 (7.1)	-3808.6 (10.1)	-3914.8 (0.0)	17	-8805.5 (14.1)	-8824.2 (15.6)	-9110.6 (0.0)
8	**-4318.7 (4.3)**	-4334.8 (5.4)	-4441.8 (0.0)	18	**-9293.5 (5.9)**	-9313.0 (6.3)	-9606.5 (0.0)
9	**-4836.7 (1.5)**	-4854.7 (4.5)	-4965.7 (0.0)	19	**-9773.9 (3.7)**	-9798.5 (7.5)	-10124.0 (0.0)
10	**-5347.8 (2.1)**	-5363.9 (4.1)	-5482.8 (0.0)	20	**-10235.4 (10.9)**	-10261.0 (10.1)	-10623.0 (0.0)

has worse performance for the same K when the data set size increases. This outcome is expected since the data set size has an impact only on the number of process routes in a cluster. Hence, a larger data set size means a larger similarity sum because more process routes are considered when the number of exemplars is the same. Another observation from Figure 10 is that the standard deviation of all the data sets increases when K increases, which is consistent with the conclusions on the process route data sets. It is noticeable that a data set with a larger size does not mean a larger standard deviation in general. However, the data set with 50 process routes has the smallest standard deviation on both process route data sets.

2) ANALYSIS OF THE INFLUENCE OF THE QUANTITY CONSTRAINT

The purpose of this section is to further analyse the influence of the quantity constraint on the performance of the proposed K-medoids method, which includes the changes of n (i.e., the minimum number of process routes for a valid process route cluster) and q_e (i.e., the penalty of the quantity constraint).

a: ANALYSIS OF THE INFLUENCE OF PARAMETER SETTING n

To deeply understand how the change in n affects the performance of the modified K-medoids method, 5 scenarios with different parameter settings for the size constraint are designed in this section (i.e., $\varepsilon = -1/5, -1/10, -1/15$ and $-1/20$), and the corresponding K is also set to 5, 10, 15, and 20, respectively, to eliminate the influence of the size constraint. For each scenario, there are n changes based on the expected cluster number of the quantity constraint K_e (which varies from 1 to 20), and $n = round(N/K_e)$. In addition, $q_e = 0.5$, $q_r = 1$. The result is shown in Figure 11.

From Figure 11, it can be seen that the influence of parameter n on the performance shares similar patterns in the 5 scenarios. When $K_e = 1$, the corresponding performance is the worst. The reason is that the corresponding $n = 600$, while the average process route quantity of clusters is much smaller than 600. With the increase in K_e, n decreases sharply, but the average process route quantity of clusters changes little because K does not change in this section. Therefore, the performance is improving. When $K_e > K$, the quantity constraint can be easily satisfied, in which case the performance of the modified K-medoids method is only slightly further improved. Another interesting observation from Figure 11 is that the standard deviation of the scenario with $\varepsilon = -1/5$ is the best, while the scenario with $\varepsilon = -1/20$ is generally the worst. This finding is still consistent with our previous conclusion that the standard deviation is closely related to K, and the corresponding K for scenarios with $\varepsilon = -1/5$ and $\varepsilon = -1/20$ are 5 and 20, respectively.

b: ANALYSIS OF THE INFLUENCE OF PARAMETER SETTING q_e

To analyse the effect of q_e on the clustering performance, 5 scenarios with different n values calculated by K_e are designed to reflect the different quantity constraint requirements (i.e., the expected cluster number for the quantity constraint). In these 5 scenarios, $\varepsilon = -1/10$ with $q_r = 1$. Hence, there exists conflict between the two constraints. Figure 12 shows the optimal performance and the corresponding K for each q_e.

Figure 12 shows that the optimal performance value and the corresponding K tend to be stable when $q_e > 0.8$, which indicates that $q_e = 0.8$ has already been strong enough to overwhelm the size constraint. In this case, the clustering result changes little when $q_e > 0.8$. The same conclusion can also be found in figures with corresponding K, in which the optimal cluster number tends to be consistent with the corresponding K_e as q_e increases. When q_e is small (e.g., $q_e = 0$), the size constraint dominates the clustering process, and the optimal cluster number is 20 because the size constraint causes a preference for small clusters. As q_e increases, the modified K-medoids method needs to achieve a tradeoff between two conflict constraints. For example, the optimal cluster number of the scenario with $K_e = 5$ is 14 when $q_e = 0.2$ on process route set 1, which is neither the optimal cluster number of the size constraint (i.e., 20) nor the expected cluster number of the quantity constraint (i.e., 5). This finding proves that the proposed K-medoids method can effectively obtain a balance between the two conflicting soft constraints based on their penalty settings.

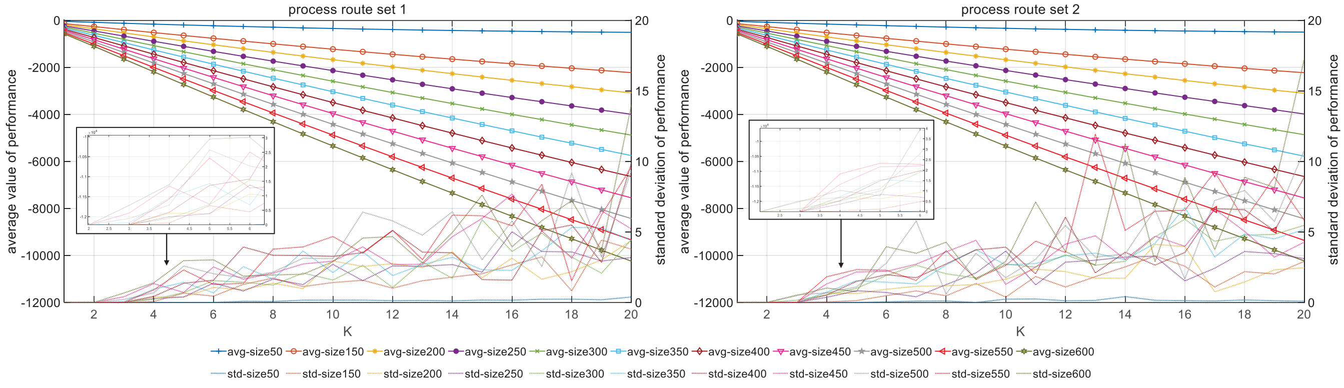


FIGURE 10. **Analysis of the influence of the process route set size.**

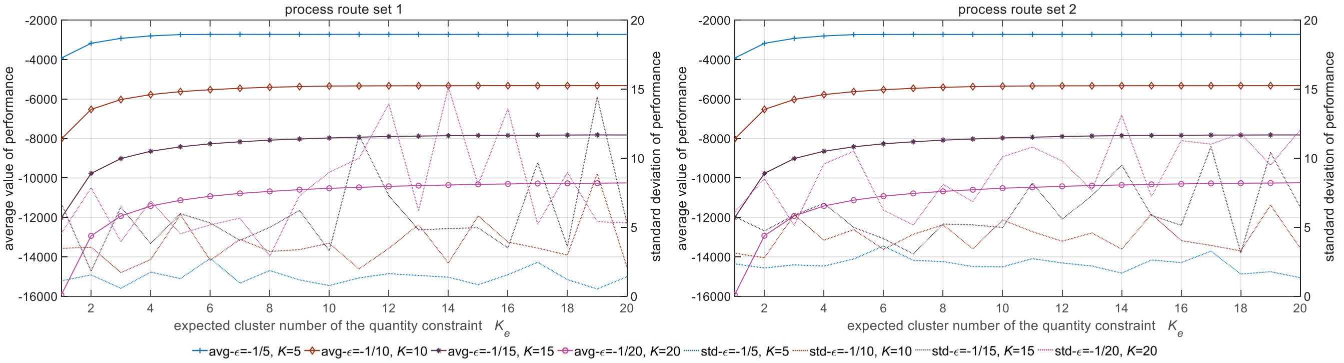


FIGURE 11. Analysis of the influence of parameter setting *n*.

3) ANALYSIS OF THE INFLUENCE OF THE SIZE CONSTRAINT

Similarly, this section aims to analyse the influence of the size constraint on the performance of the proposed K-medoids method, which includes the changes in ε (i.e., the maximum radius for a valid process route cluster) and q_r (i.e., the penalty of the size constraint).

a: ANALYSIS OF THE INFLUENCE OF PARAMETER SETTING ε

For a comprehensive analysis of ε, 5 scenarios with different parameter settings for the quantity constraint are designed, in which n is 120, 60, 40, and 30 (i.e., the expected cluster number of the quantity constraint $K_e = 5$, 10, 15, 20). The corresponding cluster number K in these scenarios is also 5, 10, 15, and 20 to eliminate the influence of the quantity constraint. The maximum radius of the clusters ε varies from -1 to $-1/20$ in each scenario, which is calculated by $\varepsilon = -1/K_r$ (K_r is the expected cluster number of the size constraint). The settings of q_e and q_r are consistent with Section VI-C2, and the result is shown in Figure 13.

At first glance, the performance of the modified K-medoids method in all 5 scenarios decreases rapidly when K_r increases (i.e., ε decreases), which is similar to that in Figure 9. The reason is that the increase in K_r essentially indicates a smaller ε, in which case the penalty of the size constraint increases almost linearly while the cluster radius changes little for a given K. When taking a closer look at Figure 13, it can be found that the performance gap is small compared with that in Figure 11. This finding can also be explained by the relationship between the cluster radius and the size constraint penalty. As we discussed in Section VI-B2, the variation in K has little effect on the cluster radius for process route sets. Hence, for a given K_r (i.e., a given ε), the size constraint penalty does not change substantially. In addition, it can also be found that the standard deviations in the performance values increase as K_r increases, which actually reflects the increasing difficulty of solving the clustering problem as it becomes more difficult to find an optimal solution.

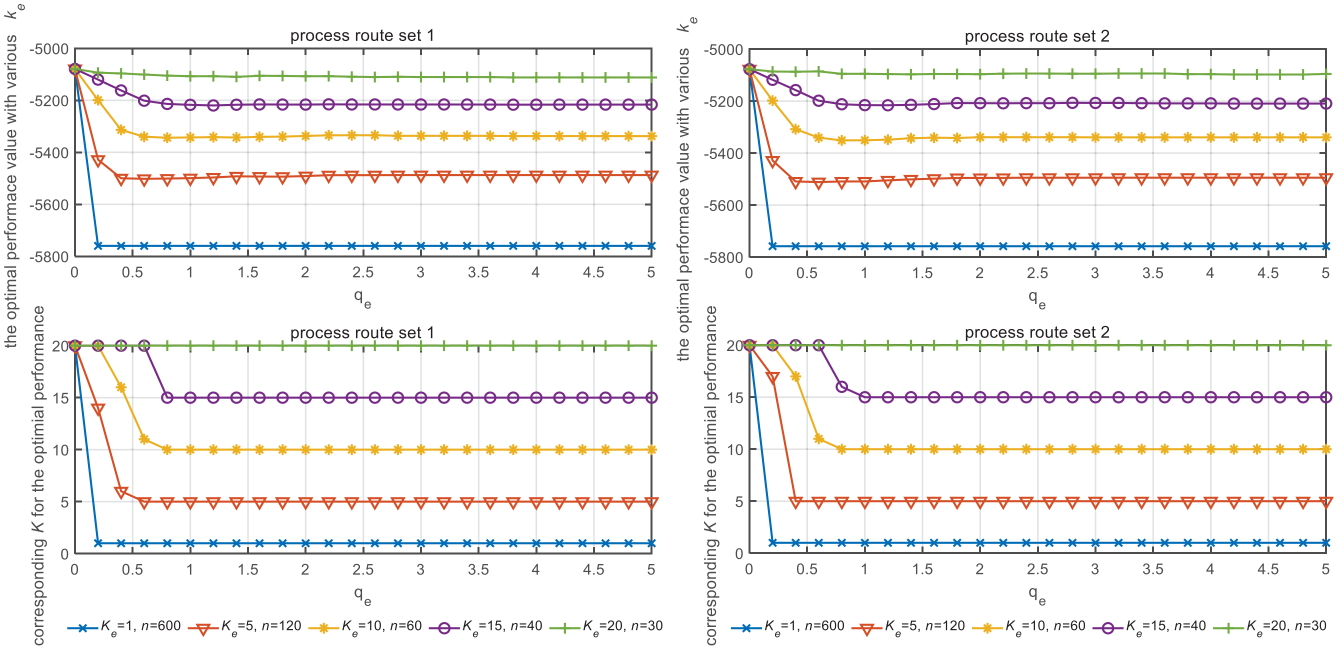


FIGURE 12. Analysis of the influence of parameter setting q_e.

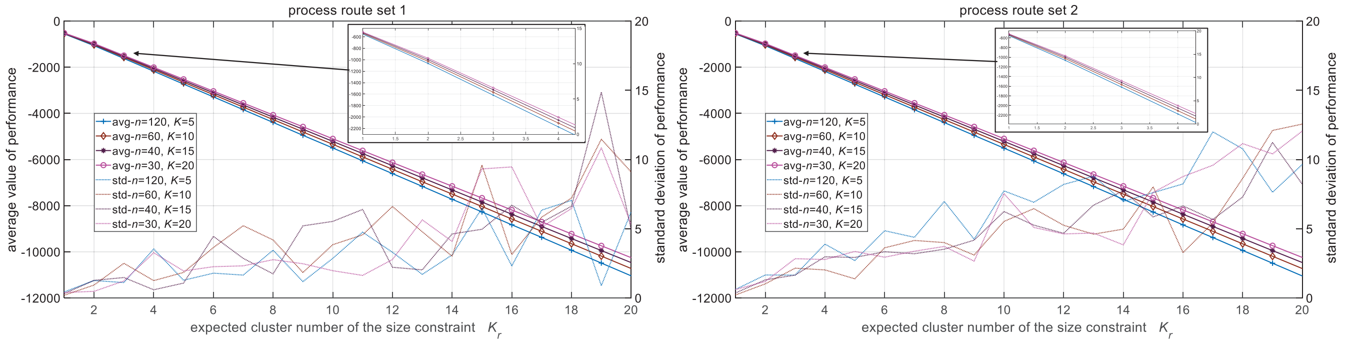


FIGURE 13. Analysis of the influence of parameter setting ε.

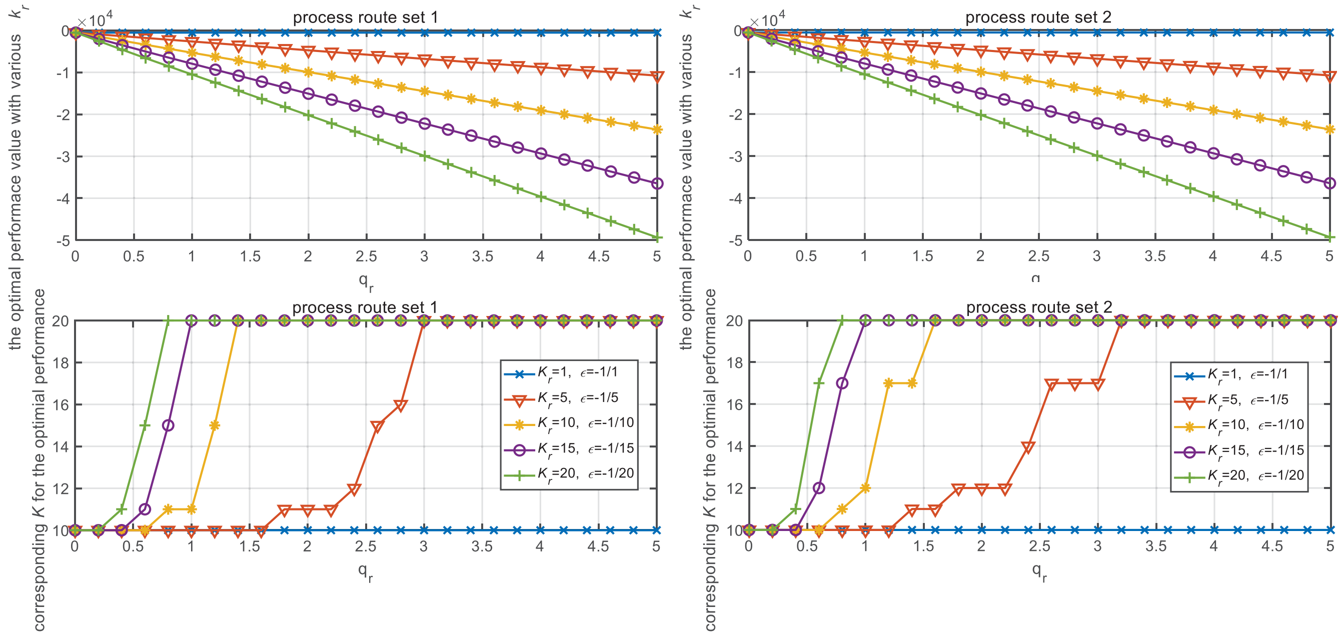


FIGURE 14. Analysis of the influence of parameter setting q_r.

b: ANALYSIS OF THE INFLUENCE OF PARAMETER SETTING q_r

In this section, 5 scenarios with different ε that were calculated by K_r are given to show different size constraint requirements. For each scenario, the parameter settings of the quantity constraint are $n = round(600/10) = 60$ and q_e=0.5. The penalty of the size constraint q_r varies from 0 to 5, and the overall result is shown in Figure 14.

An overall conclusion obtained from Figure 14 is that the optimal performance value decreases quickly when q_r increases. At the same time, the corresponding K of the optimal performance increases from 10 to 20. Figure 14 first indicates that the optimal performance value behaves similarly to Figure 9 and Figure 13. Among the 5 scenarios, the performance value of scenarios with $K_r = 5$, 10, 15, and 20 decreases almost linearly, while scenarios with $K_r = 1$ remain the same because its performance value is always -538.946. When $K_r = 1$, the corresponding ε is -1, in which case the size constraint has no effect on the performance because the minimum similarity of the process route sets is -1. When $K_r > 1$, the corresponding ε decreases rapidly, while the size of the generated cluster changes slightly as the cluster number K varies. Therefore, the related size constraint penalty accumulates when q_r increases, which leads to a decrease in the performance value. On the other hand, Figure 14 also shows that the optimal cluster number K increases as q_r increases. The reason is that the size constraint prefers small clusters (i.e., large cluster numbers) to enable this constraint to be satisfied as much as possible (i.e., reduce the number of process routes that can violate the size constraint).

VII. CONCLUSION AND FUTURE WORK

The process route plays an important role in manufacturing systems, which directly affects the quality of products, the performance of the system, and other aspects. It is difficult for technicians to design a good process route from scratch considering all related aspects. The goal of this paper is to find typical process route knowledge from historical data sets to reuse hidden knowledge. To achieve this goal, this paper proposed a discovery method for typical process routes that contains two parts: a novel comprehensive similarity coefficient for process routes and a modified exemplar-based clustering algorithm that considers two soft constraints during the clustering.

To effectively measure the similarity of the operation sequences, a deep analysis was performed to determine the information requirements and characteristics of the operation sequence similarity problem. Then, all the potential similarity cases in this problem are given. The analysis indicates that the similarity coefficient needs to consider three kinds of information: precedence relationships, the number of common operations, and operation similarities. Therefore, the modified pseudo-LCS is proposed to record the first two pieces of information, and a corresponding backtracking algorithm is also presented. The Jaccard similarity coefficient is used here to measure the last information. These two similarity coefficients are combined based on PCA to generate a novel comprehensive similarity coefficient. The numerical illustration result shows that it can distinguish all the different cases with rational similarity values.

Since typical process route discovery is a practical problem, two conflicting soft constraints are introduced into the traditional typical process route discovery problem, which makes it more complicated and practical. Because most of the previous related work did not consider these two soft constraints simultaneously or simply treated them as hard constraints, the modified K-medoids method is proposed to solve this novel problem. The illustration indicates that the proposed clustering method has better performance than existing algorithms, which ignore the constraints during clustering. If these soft constraints are treated as hard constraints, there is no valid cluster if the related parameter setting is strict. In other words, considering these two constraints as soft constraints can improve the fault tolerance and alleviate the sensitivity of manually designed constraint parameters. Further analysis also indicates that the performance of the modified K-medoids method is affected by the data set size and the parameter setting of the constraints, which suggests that the proposed method does consider this information during clustering.

There still exist some potential issues in our work, which can be addressed in the future. On the one hand, the situation of repeated operations in the process route also should be considered. In addition, the process route is often presented as a reticular structure in modern manufacturing, in which case the current similarity coefficient is no longer suitable. On the other hand, the clustering problem with two soft constraints can also be solved by optimization algorithms (e.g., genetic algorithms) and neural networks. Moreover, the outliers in the process route set should be handled to avoid clustering bias.

ACKNOWLEDGMENT

The authors would like to thank Dr. Yi Mei from Victoria University of Wellington for helping them to polish this article.

REFERENCES

[1] M. Gadalla and D. Xue, "Recent advances in research on reconfigurable machine tools: A literature review," *Int. J. Prod. Res.*, vol. 55, no. 5, pp. 1440–1454, Mar. 2017.

[2] D. Chen, Z. Jiang, S. Zhu, and H. Zhang, "A knowledge-based method for eco-efficiency upgrading of remanufacturing process planning," *Int. J. Adv. Manuf. Technol.*, vol. 108, no. 4, pp. 1153–1162, Feb. 2020.

[3] S. P. L. Kumar, "Knowledge-based expert system in manufacturing planning: State-of-the-art review," *Int. J. Prod. Res.*, vol. 57, nos. 15–16, pp. 4766–4790, Aug. 2019.

[4] C. Rösiö and J. Bruch, "Exploring the design process of reconfigurable industrial production systems: Activities, challenges, and tactics," *J. Manuf. Technol. Manage.*, vol. 29, no. 1, pp. 85–103, Jan. 2018.

[5] M. Chattopadhyay, S. Sengupta, T. Ghosh, P. K. Dan, and S. Mazumdar, "Neuro-genetic impact on cell formation methods of cellular manufacturing system design: A quantitative review and analysis," *Comput. Ind. Eng.*, vol. 64, no. 1, pp. 256–272, Jan. 2013.

[6] H. Ma, X. Zhou, W. Liu, Q. Niu, and C. Kong, "A customizable process planning approach for rotational parts based on multi-level machining features and ontology," *Int. J. Adv. Manuf. Technol.*, vol. 108, no. 3, pp. 647–669, May 2020.

[7] S. Huang and Y. Yan, "Part family grouping method for reconfigurable manufacturing system considering process time and capacity demand," *Flexible Services Manuf. J.*, vol. 31, no. 2, pp. 424–445, Jun. 2019.

[8] E. Ostrosi and A.-J. Fougères, "Intelligent virtual manufacturing cell formation in cloud-based design and manufacturing," *Eng. Appl. Artif. Intell.*, vol. 76, pp. 80–95, Nov. 2018.

[9] L. Wu, L. Li, L. Tan, B. Niu, R. Wang, and Y. Feng, "Improved similarity coefficient and clustering algorithm for cell formation in cellular manufacturing systems," *Eng. Optim.*, vol. 52, no. 11, pp. 1923–1939, Nov. 2020.

[10] G.-C. Vosniakos and T. Giannakakis, "A knowledge-based manufacturing advisor for pressworked sheet metal parts," *J. Intell. Manuf.*, vol. 24, no. 6, pp. 1253–1266, Dec. 2013.

[11] D. Zhou and X. Dai, "Integrating granular computing and bioinformatics technology for typical process routes elicitation: A process knowledge acquisition approach," *Eng. Appl. Artif. Intell.*, vol. 45, pp. 46–56, Oct. 2015.

[12] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114060.

[13] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017.

[14] G. Köksal, I. Batmaz, and M. C. Testik, "A review of data mining applications for quality improvement in manufacturing industry," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13448–13467, Sep. 2011.

[15] M. Yuan, K. Deng, W. A. Chaovalitwongse, and H. Yu, "Research on technologies and application of data mining for cloud manufacturing resource services," *Int. J. Adv. Manuf. Technol.*, vol. 99, nos. 5–8, pp. 1061–1075, Nov. 2018.

[16] P. Espadinha-Cruz, R. Godina, and E. M. G. Rodrigues, "A review of data mining applications in semiconductor manufacturing," *Processes*, vol. 9, no. 2, p. 305, Feb. 2021.

[17] K. Yasuda and Y. Yin, "A dissimilarity measure for solving the cell formation problem in cellular manufacturing," *Comput. Ind. Eng.*, vol. 39, nos. 1–2, pp. 1–17, Feb. 2001.

[18] H. Hu, Z. Li, H. Dong, and T. Zhou, "Graphical representation and similarity analysis of protein sequences based on fractal interpolation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 1, pp. 182–192, Jan. 2017.

[19] Y.-S. Lin, J.-Y. Jiang, and S.-J. Lee, "A similarity measure for text classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1575–1590, Jul. 2014.

[20] C. D. S. Garcia, A. Meincheim, E. R. F. Junior, M. R. Dallagassa, D. M. V. Sato, D. R. Carvalho, E. A. P. Santos, and E. E. Scalabrin, "Process mining techniques and applications—A systematic mapping study," *Expert Syst. Appl.*, vol. 133, pp. 260–295, Nov. 2019.

[21] F. Alhourani, "Clustering algorithm for solving group technology problem with multiple process routings," *Comput. Ind. Eng.*, vol. 66, no. 4, pp. 781–790, Dec. 2013.

[22] L. Wang, Y. Zhang, and S. Zhong, "Typical process discovery based on affinity propagation," *J. Adv. Mech. Des. Syst. Manuf.*, vol. 10, no. 1, pp. 1–13, Jan. 2016.

[23] M. Imran, C. Kang, Y. H. Lee, M. Jahanzaib, and H. Aziz, "Cell formation in a cellular manufacturing system using simulation integrated hybrid genetic algorithm," *Comput. Ind. Eng.*, vol. 105, pp. 123–135, Mar. 2017.

[24] A. M. Zohrevand, H. Rafiei, and A. H. Zohrevand, "Multi-objective dynamic cell formation problem: A stochastic programming approach," *Comput. Ind. Eng.*, vol. 98, pp. 323–332, Aug. 2016.

[25] T. Ma, Y. Liu, Q. Dai, Y. Yao, and P.-A. He, "A graphical representation of protein based on a novel iterated function system," *Phys. A, Stat. Mech. Appl.*, vol. 403, pp. 21–28, Jun. 2014.

[26] X. Li, S. Zhang, R. Huang, B. Huang, C. Xu, and Y. Zhang, "A survey of knowledge representation methods and applications in machining process planning," *Int. J. Adv. Manuf. Technol.*, vol. 98, nos. 9–12, pp. 3041–3059, Jul. 2018.

[27] F. Choobineh, "A framework for the design of cellular manufacturing systems," *Int. J. Prod. Res.*, vol. 26, no. 7, pp. 1161–1172, Jul. 1988.

[28] K. Y. Tam, "An operation sequence based similarity coefficient for part families formations," *J. Manuf. Syst.*, vol. 9, no. 1, pp. 55–68, Jan. 1990.

[29] Y.-C. Ho, C.-E.-C. Lee, and C. L. Moodie, "Two sequence-pattern, matching-based, flow analysis methods for multi-flowlines layout design," *Int. J. Prod. Res.*, vol. 31, no. 7, pp. 1557–1578, Jul. 1993.

[30] R. G. Askin and M. Zhou, "Formation of independent flow-line cells based on operation requirements and machine capabilities," *IIE Trans.*, vol. 30, no. 4, pp. 319–329, Apr. 1998.

[31] S. A. Irani and H. Huang, "Custom design of facility layouts for multiproduct facilities using layout modules," *IEEE Trans. Robot. Autom.*, vol. 16, no. 3, pp. 259–267, Jun. 2000.

[32] H. Huang, "Facility layout using layout modules," Ph.D. dissertation, Dept. Ind., Welding Syst. Eng., Ohio State Univ., Columbus, OH, USA, 2003.

[33] K. K. Goyal, P. K. Jain, and M. Jain, "A comprehensive approach to operation sequence similarity based part family formation in the reconfigurable manufacturing system," *Int. J. Prod. Res.*, vol. 51, no. 6, pp. 1762–1776, Mar. 2013.

[34] G.-X. Wang, S.-H. Huang, X.-W. Shang, Y. Yan, and J.-J. Du, "Formation of part family for reconfigurable manufacturing systems considering bypassing moves and idle machines," *J. Manuf. Syst.*, vol. 41, pp. 120–129, Oct. 2016.

[35] D. Zhou and X. Dai, "A method for discovering typical process sequence using granular computing and similarity algorithm based on part features," *Int. J. Adv. Manuf. Technol.*, vol. 78, nos. 9–12, pp. 1781–1793, Jan. 2015.

[36] J. Navaei and H. ElMaraghy, "Grouping part/product variants based on networked operations sequence," *J. Manuf. Syst.*, vol. 38, pp. 63–76, Jan. 2016.

[37] S. Liu, Z. Zhang, and X. Tian, "A typical process route discovery method based on clustering analysis," *Int. J. Adv. Manuf. Technol.*, vol. 35, nos. 1–2, pp. 186–194, Oct. 2007.

[38] S. M. Hasan, A. A. Baqai, S. U. Butt, and U. K. Q. Zaman, "Product family formation based on complexity for assembly systems," *Int. J. Adv. Manuf. Technol.*, vol. 95, nos. 1–4, pp. 569–585, Mar. 2018.

[39] A. Gupta, P. K. Jain, and D. Kumar, "A novel approach for part family formation using K-means algorithm," *Adv. Manuf.*, vol. 1, no. 3, pp. 241–250, Sep. 2013.

[40] R. Alqaisi, W. Ghanem, and A. Qaroush, "Extractive multi-document Arabic text summarization using evolutionary multi-objective optimization with K-medoid clustering," *IEEE Access*, vol. 8, pp. 228206–228224, Dec. 2020.

[41] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, Mar. 2009.

[42] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *J. ACM*, vol. 21, no. 1, pp. 168–173, Jan. 1974.

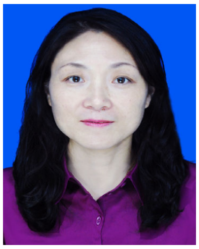
[43] Y. Lu, "Research on similar sentence retrieval technology for patents," M.S. thesis, Knowl. Eng. Centre, Shenyang Aerospace Univ., Shenyang, China, 2010.

[44] P. Jaccard, "The distribution of the flora in the Alpine zone," *New Phytol.*, vol. 11, no. 2, pp. 37–50, Feb. 1912.

[45] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, "Understanding and enhancement of internal clustering validation measures," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 982–994, Jun. 2013.

[46] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, Dec. 1945.

BINZI XU received the Ph.D. degree in control science and engineering from Jiangnan University, Wuxi, China, in 2019. He is currently a Lecturer with the School of Electrical Engineering, Anhui Polytechnic University, Wuhu, China. His main research interests include knowledge automation, intelligent manufacturing, job shop scheduling, and genetic programming hyper-heuristics (GPHH).



YAN WANG received the Ph.D. degree in control theory and control engineering from Nanjing University of Science and Technology, Nanjing, China, in 2006. From 2013 to 2014, she was a Visiting Researcher with the School of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA, USA. She is currently a Professor of control science with Jiangnan University and the Head of the Provincial Excellent Team of Innovative Research. She has authored more than 80 articles. She holds 11 invention patents and provided one industry standard. Her research interests include intelligent manufacturing using perception and collaboration technologies. She was a recipient of the Yangtse River Scholar of the Ministry of Education of China and the Provincial Science Fund for Distinguished Young Scholars, Jiangsu, China.

ZHICHENG JI received the Ph.D. degree in power electronics and drives from China University of Mining and Technology, Xuzhou, China, in 2004. From 2003 to 2004, he was a Visiting Researcher with the University of Toronto, Canada. He is currently a Professor of control science with Jiangnan University and the Deputy Director of the Information Department, 7th Science Technology Committee, Ministry of Education of China. He has authored one book and more than 200 articles. He holds more than ten invention patents. His research interests include intelligent manufacturing, new energy, and control techniques for the Internet of Things. His awards and honors include, three times, the First Class Award in teaching achievement, Jiangsu, China, in 2007, 2009, and 2013; twice, the First Class Award in research achievements (science and technology), Ministry of Education, China, in 2011 and 2016; and the First Class Award in teaching achievement, Ministry of Education, China, in 2018.