A Clustering-based Sequence Variants Analysis Method for Electronic Medical Records of Multimedical Institutions

Hieu Hanh Le
Ochanomizu University
Tokyo, Japan
le@is.ocha.ac.jp

Yuki Yasumitsu
Tokyo Institute of Technology
Tokyo, Japan
yasumitsu@li.c.titech.ac.jp

Ryosuke Matsuo, Tomoyoshi Yamazaki
Ochanomizu University
Tokyo, Japan
matsuo@ldi.or.jp,yamazaki.cp@gmail.com

Haruo Yokota
Josai University
Tokyo, Japan
yokota.h.aa@gmail.com

***Abstract*—A sequence variant (SV) containing branches is considered an extension of a sequence widely used to represent an ordered list of items. Although comparing such SVs is vital in practical applications, an efficient method to compare more than three SVs efficiently has yet to be studied. When the number of SVs increases, there is a high possibility that common parts do not exist. Hence, we cannot thoughtfully understand the similarities and differences among the SVs that were compared. In this paper, we first develop a method to exclude general items that have high frequency because they appear in almost every sequence and then cluster the SVs into several groups using a defined SV distance. Finally, we calculate the longest common SV in each group and generate a merged SV to visualize the commonality and differences of the target SVs efficiently. The proposed method is shown to be effective when applied to a real medical dataset from 23 medical institutions.**

***Index Terms*—Sequential Pattern Mining, Electronic Medical Records, Sequence Variant, Medical Support.**

I. Introduction

A sequence variant (SV) is an extension of a sequence widely used to represent an ordered list of items. SV is used to describe a list of partially ordered items where branches exist. Figure 1 shows an SV example describing a clinical treatment with branches at "Body Test" and "Surgery". It means that there would be whether a "CT Test" or an "X-Ray Test" after the "Body Test", and a "Prescription" or an "Injection" after the "Surgery". Several works have aimed to analyze such SV by quantitatively evaluating [1], [2] or studying the reasons that led to the branches [3]. The concept of SV has been formulated, and algorithms to visualize the common and the different parts of the two SVs have been proposed [4]. The algorithms have been applied to real electronic medical records (EMR) data and were found to be effective in comparing two SVs [4], [5].

However, in these works, the proposed algorithms only work on comparing two SVs, causing limitations in the practical world. In 2015, the Government of Japan initiated the Millennium Medical Record Project to manage and reuse medical records nationwide. Because of the increasing number of medical institutes participating in the project, useful methods to analyze multiple hospitals' data are highly desired. Comparing

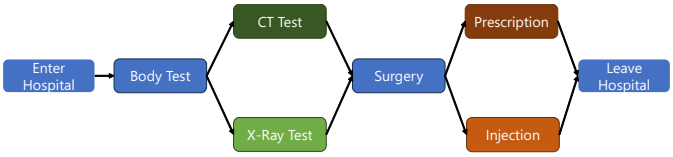

Fig. 1. An illustration of a Sequence Variant (SV)

frequent medical order patterns across various medical institutions and elucidating commonalities and differences in medical order patterns among medical institutions is expected. Such comparison will confirm the characteristics of one's medical institution and lead to improvements in medical practices by referencing treatment patterns from other medical institutions.

Efficiently comparing more than three SVs that were extracted from more than three sequence databases (SDBs) is highly needed but challenging. As the number of medical institutions for comparison increases, the commonalities decrease, making it difficult to grasp the characteristics of medical institutions accurately. Moreover, as we focus on the data from multiple sources, there are general items that occur in almost all sequences; hence, their occurrence will be high. However, such items are usually optional in the final results. For example, body or blood tests are generally performed in almost all the treatment processes but do not give useful information in the extracted frequent patterns. If such items are included in the calculation, other essential items with less frequency may be excluded in the final results.

This paper proposes a clustering-based comparison method for more than three SDBs. First, we suggest a calculation to exclude general items during running sequential pattern mining (SPM) to extract SV from each SDB. In detail, we consider the total occurrence of the items and the number of sequences in which the item appears. The items that have high frequency but appear in fewer sequences will have a higher probability of being extracted in the final frequency patterns. For each SDB, we extract its SV and propose an SV distance (SVD) to calculate the distance among SVs for hierarchical clustering. Finally, we propose an algorithm for

each cluster to calculate the longest common SV (LCSV) and merged SV (MSV) for more than three SVs. The proposed method is then evaluated using a real EMR dataset from 23 medical institutions. It is verified by the medical staff that the proposed methods are successful in clustering the medical institutions and visualizing the commonality and differences in the typical treatment processes of COVID-19 of compared institutions.

The remainder of this paper is organized as follows. Related works are summarized in Section II. The proposed methods and experimental evaluation are described in Sections III and IV. Conclusion and future works are discussed in Section V.

II. Related Works

This section gives a brief review of related works on SPM and analyzing SVs.

A. Sequential Pattern Mining

SPM is a very active research topic including numerous extension approaches for specific needs. This section introduces major approaches related to this work: frequency-based SPM and time-interval SPM. A well-known SPM algorithm is the Apriori-based frequent-pattern mining algorithm [6]. However, it is very time-consuming with large datasets and generates many irrelevant patterns among its results. To exclude patterns, PrefixSpan [7] was proposed to mine the complete set of patterns while reducing the effort of candidate pattern generation by exploring prefix projection. To improve the efficiency further, CSpan [8] was proposed for mining closed sequential patterns. For a sequence seq, if there is no supersequence with the same support as it and containing it, seq is a closed sequence.

Initially, the method proposed by Agrawal et al. [6] did not consider the time interval between items. T-PrefixSpan [9] is a method of extracting frequent sequential patterns from EMR data that considers time intervals and the efficacy of medicines. T-CSpan [10] further improves the speed performance by applying the idea of mining only closed patterns.

B. Sequential Variant Analysis

Honda et al. detected the common part of the closed frequent pattern with the same number of items for each relative treatment day [11]. Then, Le et al. proposed methods to evaluate the reasons why variants might appear by using multivariate analysis [1], [3]. There has been a method to compare two SVs by suggesting the concept of common SVs (CSVs), LCSV, and the MSV and describing the algorithms to calculate these sequences [4]. Based on the idea in this paper, Zhao et al. have proposed a method to analyze the transitions in differences between medical orders for the treatment of COVID-19 [5].

III. Proposal

In this section, we describe the proposal that is presented in Figure 2. For each SDB, we perform the SPM algorithm while excluding general items to extract SVs. Then, we conduct

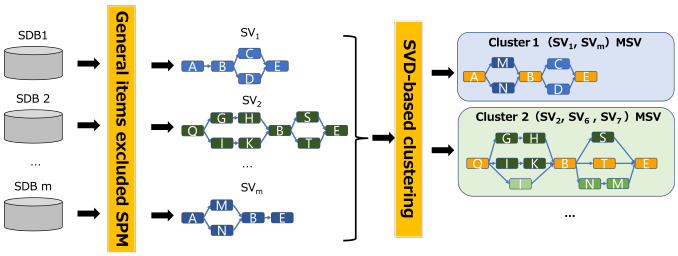

Fig. 2. An illustration of the proposed method

SVD-based clustering based on the calculation of distance among SVs to cluster the SDBs. Finally, for each cluster, to generate LCSV and MSV, we perform algorithms that work with more than three SVs.

A. SPM with Excluding General Items

Only considering an item's occurred frequency during mining for frequency patterns is not enough as it may exclude other less frequent but more important items. For instance, medical orders contained in EMR data are generally categorized into test items and treatment items. Treatment items mainly include medication, and when analyzing treatment patterns for a specific disease, treatment items are crucial medical orders. However, because medical order sequences often contain numerous test items, frequent medical order patterns obtained through SPM tend to include many high-frequency test items, making it difficult to extract treatment items. This problem arises from high-frequency medical orders that appear in many sequences and occur frequently within sequences. This is because SVs are created using the longest frequent patterns obtained through SPM. Therefore, to extract treatment items, it is necessary to exclude these test items while conducting SPM.

In determining which items to exclude, an indicator evaluating the frequency of occurrences is required. Thus, we propose the term occurrence per sequence (TO/S) as the average number of times an item appears in a sequence.

$$TO/S(item) = \frac{number\_of\_occurrences(item)}{number\_of\_sequences} \quad (1)$$

TO/S is calculated for each SDB regarding a specific item. When conducting experiments involving multiple SDBs, the excluded items must be determined based on the TO/S of each SDB collectively. Therefore, we calculate the maximum value ($Max(TO/S)$) of TO/S within the target group of SDBs. $Max(TO/S)$ considers items with high frequency. Therefore, items with low $Max(TO/S)$ have low occurrence frequencies across all SDBs, mitigating the problem of treatment item extraction during SPM.

In excluding general items with high $Max(TO/S)$, a threshold needs to be determined. As the objective is to extract characteristic items, we progressively exclude general items with high $Max(TO/S)$, setting the threshold as the maximum value at which the extraction of characteristic items remains

unchanged for general items with $Max(TO/S)$ lower than a certain value.

B. SV Clustering

After extracting the SV of each SDB from the target SDB group, we compare them using MSV. As the number of target SDBs increases, the LCSV, which is the common part of the SV, ceases to exist. MSV is used to show both commonalities and differences between SVs, and even if an MSV is created from a group of SVs that have no common parts, it will be similar to simply arranging SVs. Therefore, it cannot be said that the SV is compared well.

To solve this problem, clustering is used to classify SDB into several clusters, and MSVs are created within the clusters. During clustering, it is necessary to define the distance between SVs. In this paper, we propose an SVD calculated by Equation(2).

$$SVD(SV_1, SV_2) = 1 - \frac{2 \times |LCSV(SV_1, SV_2)|}{|SV_1| + |SV_2|} \quad (2)$$

Here, $|LCSV(SV_1, SV_2)|$ and $|SV|$ describe the number of items in $LCSV(SV_1, SV_2)$ and SV, respectively.

The clustering method employed is hierarchical clustering. Although other clustering methods like DBSCAN [12] were considered, it resulted in numerous clusters with only one element. Clusters are created based on the dendrogram produced by hierarchical clustering. The threshold distance for splitting clusters is determined such that the LCSV exists within the SVs of the cluster and that the number of clusters is minimized. Five hierarchical clustering methods are utilized: single linkage, complete linkage, centroid linkage, average linkage, and Ward's method. To extend the proposed method of creating MSVs for three or more medical institutions to a larger number of medical institutions, a method is adopted where clusters containing three or more elements become the most common.

C. LCSV and MSV Generation Algorithms for More Than Three SVs

We extend the algorithms [4] that only generate LCSV and MSV from two SVs to work on more than three SVs. As the number of SVs increases, the method to generate an MSV must be carefully designed to visualize the common and different parts of the compared SVs efficiently. In this paper, we propose two algorithms to obtain the MSV from three or more SVs.

a) Direct Merging Method: This method involves obtaining LCSV for N SVs and directly merging them with N SVs (Algorithm 1).

b) Distance-based Merging Method: This method involves merging SVs in order of proximity on the dendrogram obtained during clustering (Algorithm 2).

Algorithm 1 MSV (Direct Merging Method)

input: $SV_1, SV_2, \ldots, SV_N$
output: MSV
$LCSV = LCSV(SV_1, SV_2, \ldots, SV_N)$
$MSV = MSV(SV_1, SV_2, LCSV)$
for $i = 3, \ldots, N$ **do**
    $MSV = MSV(SV_i, MSV, LCSV)$
end for

Algorithm 2 MSV (Distance-based Merging Method)

input: $Tree(SV_1, SV_2, \ldots, SV_N)$
output: MSV
$Tree$=$Tree(SV_1, SV_2, \ldots, SV_N)$
$Tree_{left}$ is left subtree
$Tree_{right}$ is right subtree
if $Tree_{left} != Null$ **then**
    $MSV_{left} = MSV\_TREE(Tree_{left})$
    $MSV_{right} = MSV\_TREE(Tree_{right})$
    $LCSV_{Tree} = LCSV(SV\ in\ Tree)$
    return $MSV(MSV_{left}, MSV_{right}, LCSV_{Tree})$
else
    Tree has one SV, SV_{leaf}
    return SV_{leaf}
end if

IV. Experimental Evaluation

A. Dataset

In this experiment, we focused on medical instruction data from actual EMR pertaining to the fifth wave of COVID-19 (July 1, 2021, to September 30, 2021) provided by 23 medical institutions (Medical Institution A, B, ..., W), which are research collaborators for the study on "Investigation of the Impact of COVID-19 Infections on Medical Care and Development of Prediction Models." Tables I and II present the information on the medical instruction sequences for each medical institution.

B. Experimental Environment

To ensure the protection of personal information, a limited cloud environment managed by Miyazaki University Faculty of Medicine Hospital (Amazon WorkSpaces) is used as the experimental environment. This research has obtained approval from the ethics committee of Kyoto University. To evaluate the correct occurrence frequency of medical orders with different notations, drugs with different names but the same components, and tests with different methodologies but the same purpose, preprocessing is conducted to treat these medical orders as identical before extracting frequent patterns using SPM. We chose T-CSpan [10] as a basic SPM algorithm.

C. Experimental Results

1) Medical Orders Exclusion: We calculated $Max(TO/S)$ for medical orders appearing in frequent sequence patterns, excluding highly common test items and medical orders with low relevance to COVID-19 treatment (Table III).

TABLE I
STATISTIC INFORMATION OF SEQUENCES IN THE DATASET (PART 1/2)

Institution	A	B	C	D	E	F	G	H	I	J	K
#sequences	130	23	11	102	89	79	126	36	29	104	104
#ave_length (raw)	181.7	255.6	45.3	153.7	242.6	284.3	90.2	127.4	211.4	317.5	112.2
#ave_length (after exclusion)	18.7	56.6	5.3	22.0	47.2	31.8	23.2	15.6	37.2	51.3	35.3

TABLE II
STATISTIC INFORMATION OF SEQUENCES IN THE DATASET (PART 2/2)

Institution	L	M	N	O	P	Q	R	S	T	U	V	W
#sequences	30	179	228	182	29	57	283	55	69	22	28	22
#ave_length (raw)	275.7	282.4	272.5	181.0	73.7	310.0	139.2	174.5	189.4	340.0	220.3	137.1
#ave_length (after exclusion)	57.8	40.6	46.2	23.0	16.4	66.0	13.3	21.3	36.2	51.1	38.6	18.4

TABLE III
A PART OF $Max(TO/S)$ AND EXCLUDED TEST MEDICAL ORDERS

Medical order	Max(TO/S)
SpO2	11.0
Blood glucose measurement	10.2
Respiratory monitoring	9.5
X-ray test	9.5
Cold evaporator	9.3
Isotonic sodium chloride solution syringe	8.3
Peripheral blood general test	7.5
Glucose	6.8
Total bilirubin	6.5
C-reactive protein	6.4
Intravenous drip	6.2
Ferritin	5.6
Blood gas analysis	5.2
D-dimer	5.1
Prothrombin time (PT)	4.0
Sodium and chlorine	3.7
Direct measurement of arterial pressure	3.6
Arterial blood sampling	3.3
Activated partial thromboplastin time (APTT)	3.3
Lactic acid	3.2
Total protein	3.2
Central venous injection	3.0
Central venous pressure monitoring	3.0
Narcotic analgesic	2.9
Bacterial culture and identification test	2.7
End tidal Co2 monitoring	2.6
Inspiratory distribution	2.4
Bacterial microscope test	2.4
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	2.3
von Willebrand factor antigen	2.2
Hepatitis screening	2.1

Medical order	Max(TO/S)
Urinalysis	1.9
Syphilis test	1.6
Electrocardiogram	1.6
Computed tomography scan	1.5
Procalcitonin	1.5
Blood urea nitrogen	1.4
Creatinine	1.4
Aspartate aminotransferase	1.4
Nasopharyngeal swab collection	1.3
Krebs von den Lungen-6	1.0
Brain natriuretic peptide	0.9
Interleukin	0.9
ABO blood type	0.9
Hemoglobin A1c (HbA1c)	0.9
Free thyroxine (FT4)	0.8
Thyroid-stimulating hormone (TSH)	0.8
Thymus and activation-regulated chemokine (TARC)	0.7
LDL cholesterol	0.6
Creatine kinase MB (CK-MB)	0.6
Anti-virus antibody in each globurlin class	0.6
Interferon-gamma release assays	0.5
Pulmonary Surfactant Protein	0.5
Indwelling catheter	0.5
Soluble interleukin-2 receptor	0.5
Fibrinogen	0.2

Order excluded by Max(TO/S)

Order excluded by low relationship with Covid-19 treatment

Regarding the exclusion of common test items, the threshold for $Max(TO/S)$ was set based on hepatitis virus tests (i.e., 2.1). Excluding test items with $Max(TO/S)$ lower than that of hepatitis virus tests did not lead to the inability to extract treatment items. In addition, medical orders with low relevance to COVID-19 treatment (blue highlighted ones) were selected by healthcare workers, e.g., ABO blood type and nasopharyngeal swab collection.

2) SV Extraction: After performing the identification and exclusion of medical instructions as described in the preceding section for the 23 medical institutions, SPM was applied to extract frequent medical instruction patterns denoted as SV. For medical institutions A through V, a minimum support $minsup$ of 0.2 was used. However, for medical institution W, setting $minsup$ to 0.2 did not yield any SV patterns, so the $minsup$ was adjusted to 0.1.

3) Medical Institution Clustering: We conducted hierarchical clustering using SVD (Equation 2) as the distance function for the extracted SV patterns using the single linkage method, complete linkage method, centroid method, average linkage method, and Ward's method. As the average linkage method (Figure 3) and Ward's method (Figure 4) contained the highest number of medical institutions with three or more elements in each cluster, we adopted these two methods for the later experiments.

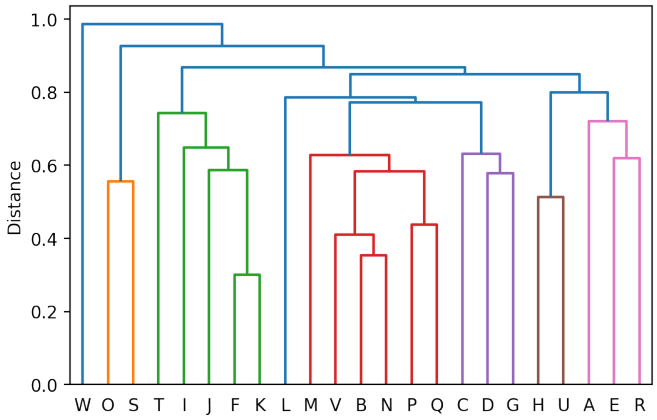

Fig. 3. Dendrogram (Average linkage method)

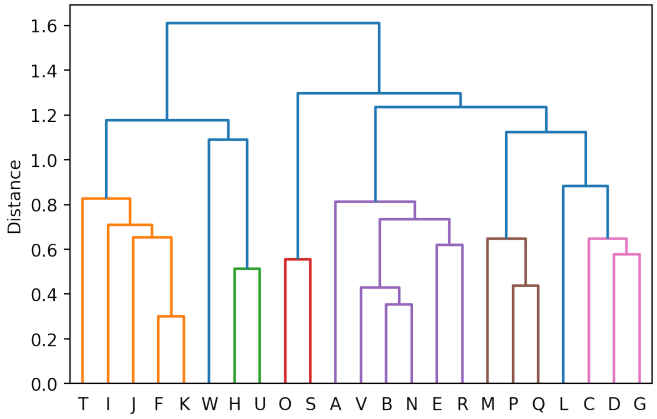

Fig. 4. Dendrogram (Ward's method)

The difference between the average linkage method and Ward's method lies in whether medical institutions B, N, and V were classified into clusters with DM or HP as common elements. In the average linkage method, B, N, and V were clustered in the same group with M, P, and Q. By contrast, in

TABLE IV
ABBREVIATION OF MEDICAL ORDERS

Abbreviation	Medical order name	Description
CT	Computed tomography scan	A test to capture the inside of the body
ECG	Electrocardiogram	A test to check the condition of the heart
UE	Urinalysis	A test to analyze the components of urine
TP	Syphilis test	A test to check for syphilis infection
BNP	Brain natriuretic peptide	A test to assess cardiac stress
BUN	Blood urea nitrogen	A test to evaluate kidney function
GB	Antivirus antibody in each globulin class	A test to examine antibodies for various viruses
QFT	Interferon-gamma release assays	A test to check for tuberculosis infection
AST	Aspartate aminotransferase	A test to examine abnormalities in the liver or heart
HbA1c	Hemoglobin A1c	A test to evaluate blood sugar levels
OT	Oxygen administration	Oxygen administration
DM	Dexamethasone	Steroid medication
PD	Prednisolone	Steroid medication
HP	Heparin	Anticoagulant
AA	Acetaminophen	Antipyretic analgesic
IN	Insulin	Medication for diabetes treatment
SEN	Sennoside	Laxative treatment

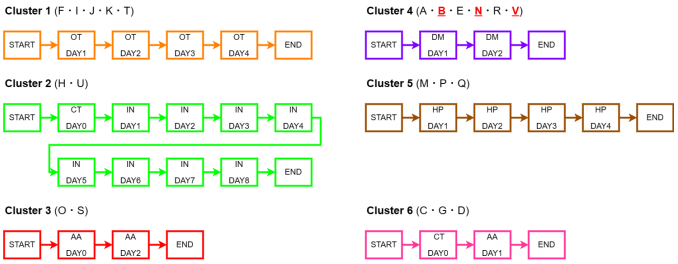

Fig. 5. LCSV in the clusters (average linkage method)

Ward's method, B, N, and V were with A, E, and R.

4) LCSV Generation: At first, we generated the LCSVs in each intracluster (Figures 5 and 6) for the average linkage and Ward's methods. Each node is presented by the medical order name and the relative day since entering the hospital. Table IV shows the abbreviations for medical instructions. The common items in each cluster include oxygen administration (OT), insulin (IN), acetaminophen (AA), dexamethasone (DM), heparin (HP), and computed tomography scan (CT).

These results show that several important treatment orders appeared in the clustered SVs. DM, a steroid agent suppressing inflammation in the lungs, heparin for preventing thrombosis, and OT are particularly crucial treatment orders in COVID-19 care. The presence or absence of these treatment items correlates with the severity of patients and reflects the characteristics of each medical institution. It is well noticed that the LCSV in cluster 2 contains an IN order that shows the medication for diabetes treatment. It suggests that the two institutions H and U served a large number of COVID-19 patients who have diabetes as a complicating disease.

5) MSV Generation: We generated MSVs for clusters 4 and 5 of Ward's method clustering results using two different methods (direct merging and distance-based methods) proposed in Section III-C. The results are presented in Figures 7 to 10. By assigning colors to nodes and edges corresponding to labels, we were able to visualize the SVs of three or more medical institutions and their common elements.

The direct merging method makes individual SVs easier to discern, but it results in a larger number of nodes, making it more redundant. By contrast, the distance-based merging method effectively represents common nodes for a subset of compared institutions' SVs. For instance, in Figure 8, DM on

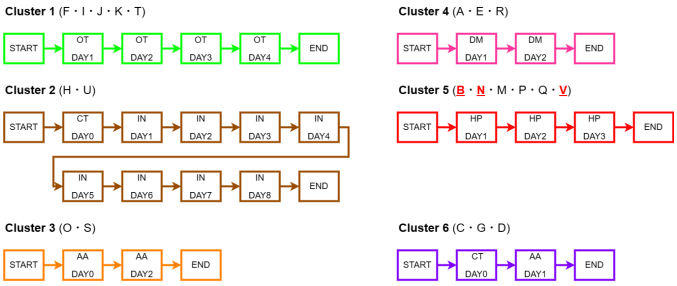

Fig. 6. LCSV in the clusters (Ward's method)

Days 3 and 4 are the common orders for B, E, N, R, and V; HP on Days 1, 2, and 3 are the common ones for B, N, and V; and HP on Days 0 and 4 are the common ones for B and N. However, having more diverse labels increases the complexity of MSV. The choice between the two methods depends on whether clarity or redundancy reduction is prioritized.

Several medical staff indicated that these results were useful in grasping the common and different medical orders in the treatment process at similar medical institutions. The same comment was obtained when we applied the proposed method to heart-related disease treatment that uses percutaneous coronary intervention to place a stent to open up blood vessels in the heart.

V. CONCLUSION AND FUTURE WORK

In this paper, to understand the commonalities and differences in frequent medical order patterns among three or more medical institutions, we proposed a method to exclude general but not important items during SPM. The SVs were extracted from the EMR data of multiple institutions. Hierarchical clustering was then performed using a defined distance metric between SVs, and MSVs were obtained by combining LCSVs and SVs within clusters using two methods, i.e., direct merging and distance-based methods. The proposed method was applied to real EMR datasets obtained from 23 medical institutions relating to the fifth wave of COVID-19. From the results, medical institutions were classified based on important treatment items, and commonalities and differences in frequent medical order patterns within clusters were identified.

In the future, guidelines for determining appropriate $minsup$ values need to be established. Furthermore, reconsideration of the SVD distance calculation and the clustering threshold are needed, e.g., taking into account factors such as the maximum number of nodes in each cluster to ensure visibility.

ACKNOWLEDGEMENT

The research is partially supported by Grants-in-Aid from JSPS (#20H04192, #21K1774, #24K02943), and uses EMR data provided by collaborating organizations in the project of "Survey of the Impact of COVID-19 Infection on Medical Practice and Development of a Predictive Model" supported by the Kansai Economic Federation.

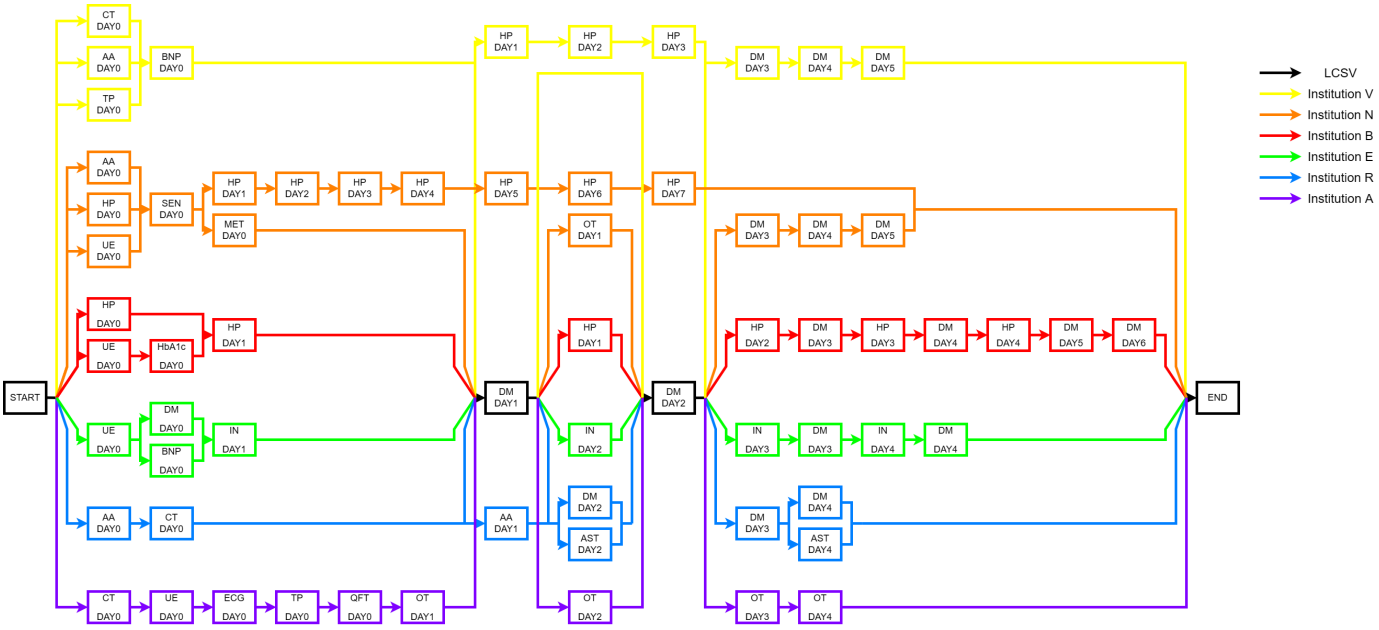

Fig. 7. MSV of cluster 4 (Direct merging method)

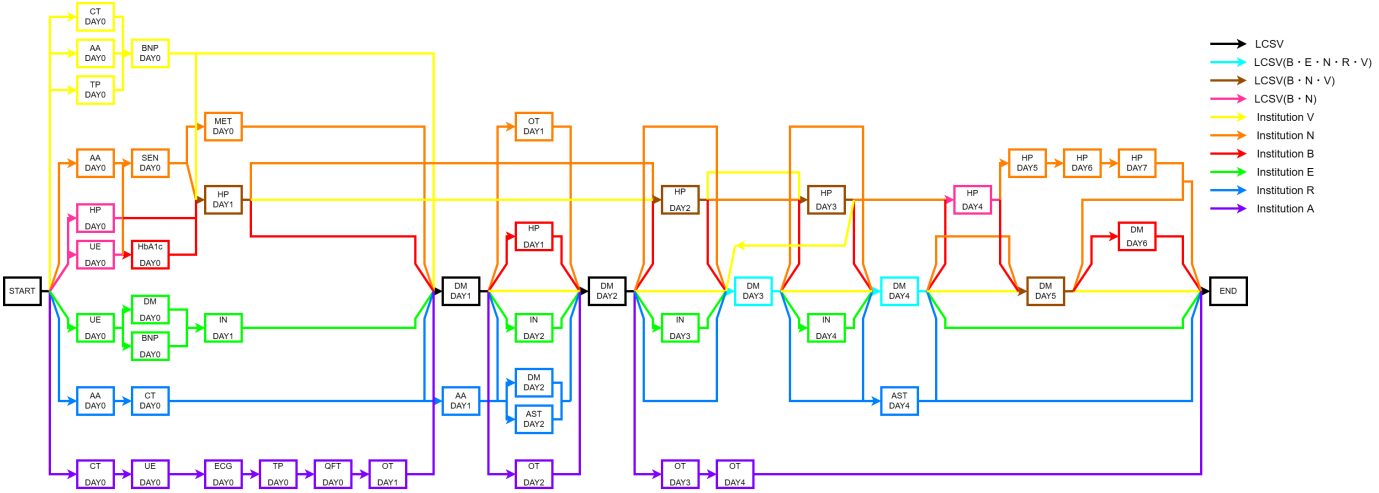

Fig. 8. MSV of cluster 4 (Distance-based method)

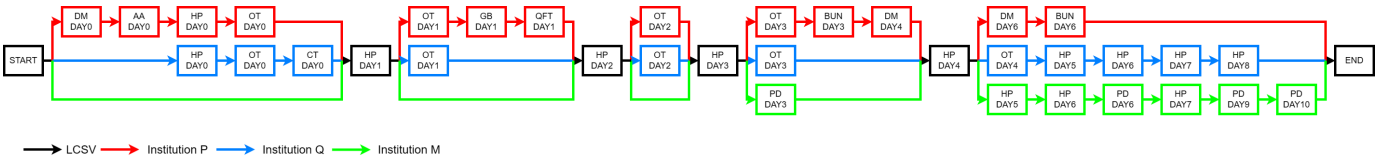

Fig. 9. MSV of cluster 5 (Direct merging method)

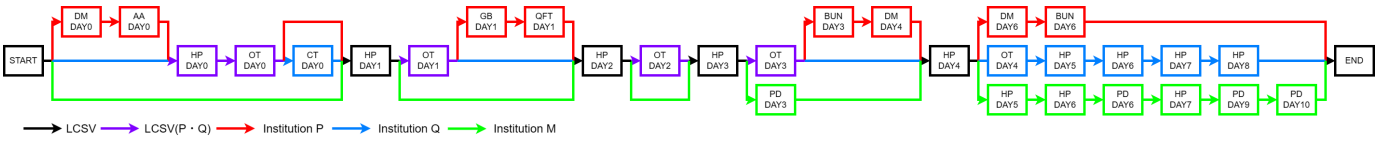

Fig. 10. MSV of cluster 5 (Distance-base method)

REFERENCES

[1] H. H. Le, T. Yamada, Y. Honda, M. Kayahara, M. Kushima, K. Araki, and H. Yokota, "Analyzing Sequence Pattern Variants in Sequential Pattern Mining and Its Application to Electronic Medical Record Systems," in *Proc. the 30th International Conference on Database and Expert Systems Applications*, ser. DEXA'19. Springer, 2019, pp. 393–408.

[2] H. H. Le, T. Yamada, Y. Honda, M. Kayahara, M. Kushima, K. Araki, and H. Yokota, "Effects of Mining Parameters on the Performance of the Sequence Pattern Variants Analyzing Method Applied to Electronic Medical Record Systems," in *Proc. the 21st International Conference on Information Integration and Web-based Applications & Services*, ser. iiWAS'21. ACM, 2021, pp. 127–135.

[3] H. H. Le, T. Yamada, Y. Honda, T. Sakamoto, R. Matsuo, T. Yamazaki, K. Araki, and H. Yokota, "Methods for Analyzing Medical-order Sequence Variants in Sequential Pattern Mining for Electronic Medical Record Systems," *ACM Transaction on Computing for Healthcare*, vol. 4, no. 1, pp. 3:1–28, 2023.

[4] Y. Li, H. H. Le, R. Matsuo, T. Yamazaki, K. Araki, and H. Yokota, "Comparison of sequence variants and the application in electronic medical records," in *Proc. the 33rd Database and Expert Systems Applications, Part 2*, ser. DEXA'22, 2022, pp. 117–130.

[5] Z. Zhao, Y. Yasumitsu, H. H. Le, T. Yamazaki, K. Araki, and H. Yokota, "Analysis of Transitions in Differences between Frequent Medical-order Sequences for COVID-19," in *Proc. the 36th International Symposium on Computer-Based Medical Systems*. IEEE, 2023, pp. 666–671.

[6] A. Rakesh and S. Ramakrishnan, "Mining Sequential Patterns," in *Proc. the 11th International Conference on Data Engineering*, ser. ICDE'95. IEEE, 1995, pp. 3–14.

[7] H. Jiawei, P. Jian, M.-A. Behzad, P. Helen, C. Qiming, D. Umeshwar, and H. MC, "Prefixspan: Mining Sequential Patterns Efficiently by Prefix-projected Pattern Growth," in *Proc. the 17th International Conference on Data Engineering (ICDE)*, 2001, pp. 215–224.

[8] R. V. Purushothama and V. G. Saradhi, "Mining Closed Sequential Patterns in Large Sequence Databases," *International Journal of Database Management Systems*, vol. 7, no. 1, pp. 29–39, 2015.

[9] K. Uragaki, T. Hosaka, Y. Arahori, M. Kushima, T. Yamazaki, K. Araki, and H. Yokota, "Sequential Pattern Mining on Electronic Medical Records with Handling Time Intervals and the Efficacy of Medicines," in *Proc. 2016 IEEE Symposium on Computers and Communication (ISCC)*. IEEE, 2016, pp. 20–25.

[10] H. H. Le, H. Edman, Y. Honda, M. Kushima, T. Yamazaki, K. Araki, and H. Yokota, "Fast Generation of Clinical Pathways Including Time Intervals in Sequential Pattern Mining on Electronic Medical Record Systems," in *Proc. the 4th International Conference on Computer Science and Computational Intelligent*, ser. CSCI'21, 2017, pp. 1726–1731.

[11] Y. Honda, M. Kushima, T. Yamazaki, K. Araki, and H. Yokota, "Detection and Visualization of Variants in Typical Medical Treatment Sequences," in *Proc. the 3rd VLDB Workshop on Data Management and Analytics for Medicine and Healthcare*, ser. DMAH'17. Springer, 2017, pp. 88–101.

[12] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proc. the second International Conference on Knowledge Discovery and Data Mining*, vol. 96, no. 34, 1996, pp. 226–231.