Advances in Intelligent Systems and Computing 1128

Yukio Ohsawa · Katsutoshi Yada ·
Takayuki Ito · Yasufumi Takama ·
Eri Sato-Shimokawara ·
Akinori Abe · Junichiro Mori ·
Naohiro Matsumura *Editors*

# Advances in Artificial Intelligence

## Selected Papers from the Annual Conference of Japanese Society of Artificial Intelligence (JSAI 2019)

Springer

# Advances in Intelligent Systems and Computing

Volume 1128

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

**** Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink ****

More information about this series at http://www.springer.com/series/11156

Yukio Ohsawa · Katsutoshi Yada ·
Takayuki Ito · Yasufumi Takama ·
Eri Sato-Shimokawara ·
Akinori Abe · Junichiro Mori ·
Naohiro Matsumura
Editors

# Advances in Artificial Intelligence

## Selected Papers from the Annual Conference of Japanese Society of Artificial Intelligence (JSAI 2019)

Springer

*Editors*

Yukio Ohsawa
Department of Systems Innovation
University of Tokyo
Tokyo, Japan

Takayuki Ito
Nagoya Institute of Technology
Nagoya, Japan

Eri Sato-Shimokawara
Department of Information
and Communication
Tokyo Metropolitan University
Tokyo, Japan

Junichiro Mori
School of Engineering
The University of Tokyo
Tokyo, Japan

Katsutoshi Yada
Faculty of Business and Commerce
Kansai University
Osaka, Japan

Yasufumi Takama
Graduate School of System Design
Tokyo Metropolitan University
Tokyo, Japan

Akinori Abe
Faculty of Letters
Chiba University
Chiba, Japan

Naohiro Matsumura
Graduate School of Economics
Osaka University
Toyonaka, Osaka, Japan

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-030-39877-4 ISBN 978-3-030-39878-1 (eBook)
https://doi.org/10.1007/978-3-030-39878-1

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

The JSAI series of Annual Conference is the central event of the Japanese Society for Artificial Intelligence. Every year we collect papers and presentations and have earnest communications toward the progress of sciences, technologies, and philosophies relevant to AI. Traditionally, the JSAI Program Committee has mainly been responsible for organizing the conference program, selecting the invited speakers, and awarding the work. In JSAI2019, with the challenge of partially internationalizing general sessions, the content was organized on the supports by the newly established International Program Committee.

Summarizing the authors' first desired category of oral presentations in the general sessions, machine learning and AI applications increased further among all the submissions, but others decreased in general. One of the reasons is that many presenters who used machine learning technology as the method of each category wanted the machine learning category. The number of participants has increased more than the number of presentations, and there are 2905 participants (the most in history) for 748 presentations, showing the accelerating interest from the industry—the increase in submissions for AI applications is supporting this inference.

This year, for the first time, an international session was held with the same position as the general session. For the international session, five categories, (1) knowledge engineering, (2) machine learning, (3) agents, (4) robots and real worlds, and (5) human interface and education aid, have been chosen considering this as a challenging start of internationalization. Seven International Program Committee members who can cover these categories worked in the selection of papers, so that each paper had two or more reviewers over a peer review period of two months. In consideration of the evaluation by the evaluation committee and chairperson during the conference, we selected the papers to be published by Springer.

We also aimed to improve the quality of the paper. Both Japanese and English clearly indicated to CFP that the outline should include (1) purpose and (2) summary or conclusion of results. The contributors were carefully noted, and 26 of the 80 talks that were finally given were published as selected papers from Springer following the two stages below.

In the first stage, for all the papers presented in the international session (oral or interactive), at least two reviewers from related fields (from Japan or abroad) for each paper for which the program chairperson confirmed the abstract satisfied the conditions specified in the CFP. The reviewers selected by the International Program Committee members (the editors of this book) reviewed them. This peer review was evaluated according to the following criteria, scoring 0–5 for each criterion.

(a) Originality (novelty of the paper)
(b) Significance (impact on sciences and businesses)
(c) Quality (soundness and clarity of the paper)

This first evaluation period was from mid-March to mid-May 2019. Next, there are 80 cases in total, and the award candidates should be 30–45%. Therefore, in each categorized field, first, the top 33% of candidates in the overall score according to the peer reviewers were selected.

In the second stage, the reviewers consisting of all members of the International Program Committee members reassessed the content and the paper points, including the period of the conference.

(1) Contents: Evaluation of research importance and novelty
(2) Paper points: Evaluation of technical quality and clarity of papers
(3) Presentation points: Evaluation of oral presentations

As a result of consideration by the program chair and the head of the International Program Committee based on the second-stage evaluation, 19 long papers and 7 short papers have been chosen for publication in this book. Then, we finally edited this book including these survived papers. In this final step, we reorganized the selected papers into six parts (I) knowledge engineering, (II) agents, (III) education and culture, (IV) natural language processing, (V) machine learning and data mining, and (VI) cyber-physics, considering the balance of the number of papers. Approximately, (I) corresponds to the previously mentioned (1), (II) to (3), (III) to (5), (IV) to the intersection of all (1) through (5), (V) to (2), and (VI) to (4). This change and rearrangement were originally proposed by Yoshiki Kishi and Danilo E. Miura who assisted the program chair in the editing of this book, attending sessions in JSAI2019 and checking details of all the selected papers. Then we, the editors, found this idea realizes a harmony showing out the feature of JSAI where philosophies, i.e., toward a human-centric society with integrating natural and artificial intelligence.

We sincerely express millions of thanks to President Naohiko Uramoto, Vice-President Shusaku Tsumoto (General Chair of JSAI2019) of JSAI, Ex-president Seiji Yamada, Executive Committee Chair Takafumi Koshinaka, Deputy Program Chair Daisuke Katagami, Assistant Program Chair Teruaki Hayashi, Ex-Program Chair Takashi Onoda, Secretary-General Kazuo Sumida, and

Akiko Yamanobe of May Corporation as the Supporting Company, and all the participants of JSAI2019. We would also like to thank Yoshiki Kishi and Danilo E. Miura for their cooperation in the editing of this post-proceedings, including contacting the author, compiling LATEX, and checking the contents of the extension.

November 2019

Yukio Ohsawa
The Program Chair of JSAI2019

Katsutoshi Yada
Takayuki Ito
Yasufumi Takama
Eri Sato-Shimokawara
Akinori Abe
Junichiro Mori
Naohiro Matsumura
The International Program Committee of JSAI, as Editor

# Contents

Knowledge Engineering

Using Sequence Constraints for Modelling Network Interactions . . . . . . 3
Johannes De Smedt, Junichiro Mori, and Masanao Ochi

Prediction of Onset of Lifestyle-Related Diseases Using Regular Health Checkup Data . . . . . . 14
Mitsuru Tsunekawa, Natsuki Oka, Masahiro Araki, Motoshi Shintani, Masataka Yoshikawa, and Takeshi Tanigawa

Feasible Affect Recognition in Advertising Based on Physiological Responses from Wearable Sensors . . . . . . 27
Taweesak Emsawas, Ken-ichi Fukui, and Masayuki Numao

Deep Markov Models for Data Assimilation in Chaotic Dynamical Systems . . . . . . 37
Calvin Janitra Halim and Kazuhiko Kawamoto

Reducing the Number of Multiplications in Convolutional Recurrent Neural Networks (ConvRNNs) . . . . . . 45
Daria Vazhenina and Atsunori Kanemura

Agents

Rewards Visualization System Promotes Information Provision . . . . . . 55
Fujio Toriumi, Hitoshi Yamamoto, and Isamu Okada

An Automated Negotiating Agent that Searches the Bids Around Nash Bargaining Solution to Obtain High Joint Utilities . . . . . . 66
Shan Liu, Shun Okuhara, Ahmed Moustafa, and Takayuki Ito

ANAC 2018: Repeated Multilateral Negotiation League . . . . . . 77
Reyhan Aydoğan, Katsuhide Fujita, Tim Baarslag, Catholijn M. Jonker, and Takayuki Ito

Flexibility of Emulation Learning from Pioneers in Nonstationary Environments . . . . . 90
Moto Shinriki, Hiroaki Wakabayashi, Yu Kono, and Tatsuji Takahashi

Education and Culture

Computerized Adaptive Testing Method Using Integer Programming to Minimize Item Exposure . . . . . 105
Yoshimitsu Miyazawa and Maomi Ueno

Proposal of Context-Aware Music Recommender System Using Negative Sampling . . . . . 114
Jin-cheng Zhang and Yasufumi Takama

The Influence of Story Creating Activities While Appreciating Abstract Artworks . . . . . 126
Kotone Tadaki and Akinori Abe

Natural Language Processing

Unsupervised Joint Learning for Headline Generation and Discourse Structure of Reviews . . . . . 139
Masaru Isonuma, Junichiro Mori, and Ichiro Sakata

Extraction of Online Discussion Structures for Automated Facilitation Agent . . . . . 150
Shota Suzuki, Naoko Yamaguchi, Tomohiro Nishida, Ahmed Moustafa, Daichi Shibata, Kai Yoshino, Kentaro Hiraishi, and Takayuki Ito

Variables Extraction in Natural (English) Language Through Possessive Relationships . . . . . 162
Danilo Eidy Miura, Yukio Ohsawa, and Teruaki Hayashi

GTransE: Generalizing Translation-Based Model on Uncertain Knowledge Graph Embedding . . . . . 170
Natthawut Kertkeidkachorn, Xin Liu, and Ryutaro Ichise

Latent-Space Data Augmentation for Visually-Grounded Language Understanding . . . . . 179
Aly Magassouba, Komei Sugiura, and Hisashi Kawai

Machine Learning and Data Mining

A Community Sensing Approach for User Identity Linkage . . . . . 191
Zexuan Wang, Teruaki Hayashi, and Yukio Ohsawa

Social Influence Prediction by a Community-Based Convolutional Neural Network . . . . . 203
Shao Hsuan Tai, Hao-Shang Ma, and Jen-Wei Huang

**Segment Information Extraction from Financial Annual Reports Using Neural Network** . . . . 215
Tomoki Ito, Hiroki Sakaji, and Kiyoshi Izumi

**One-Shot Learning Using Triplet Network with kNN Classifier** . . . . 227
Mu Zhou, Yusuke Tanimura, and Hidemoto Nakada

**Trees Detection on Google Street View Images Using Deep Learning and City Open Data** . . . . 236
Lieu-Hen Chen, Hao-Ming Hung, Cheng-Yu Sun,
Eric Hsiao-Kuang Wu, Toru Yamaguchi, Eri Sato-Shimokawara,
and Hao Chen

**Cyber Physics**

**Classification of Accessibility and Synergy in the Tokyo Area Multimodal Transportation Network** . . . . 247
Aaron Bramson, Megumi Hori, Bingran Zha, and Hirohisa Inamoto

**Modelling Naturalistic Work Stress Using Spectral HRV Representations and Deep Learning** . . . . 267
Juan Lorenzo Hagad, Ken-ichi Fukui, and Masayuki Numao

**Privacy-Preserving Resident Monitoring System with Ultra Low-Resolution Imaging and the Examination of Its Ease of Installation** . . . . 278
Shogo Murakami, Takumi Kimura, and Ikuko Eguchi Yairi

**Identity Verification Using Face Recognition Improved by Managing Check-in Behavior of Event Attendees** . . . . 291
Akitoshi Okumura, Susumu Handa, Takamichi Hoshino,
Naoki Tokunaga, and Masami Kanda

**Author Index** . . . . 305

# Knowledge Engineering

# Using Sequence Constraints for Modelling Network Interactions

Johannes De Smedt[1], Junichiro Mori[2](✉), and Masanao Ochi[2]

[1] Management Science and Business Economics Group, The University of Edinburgh,
29 Buccleuch Place, Edinburgh EH8 9JS, UK
johannes.desmedt@ed.ac.uk

[2] The University of Tokyo, 7-3-1 Bunkyo-ku, Hongo, Tokyo, Japan
mori@mi.u-tokyo.ac.jp, ochi@ipr-ctr.t.u-tokyo.ac.jp

**Abstract.** This is an extension from a selected paper from JSAI2019. The ubiquitous nature of networks has led to a vast number of works dedicated to the study of capturing their information. Various graph-based techniques exist that report on the characteristics of nodes and edges, e.g., author-citation networks, social interactions, and so on. A significant amount of information can be extracted by summarizing the surrounding network structure of nodes, e.g., by capturing motives, or walk patterns. In this work, we present a new way of capturing the interaction between nodes in a network by making use of the sequence in which they occur. (1) The objective of this paper is to make use of behavioural constraint patterns; a concise but detailed report of node's interactions can be constructed that can be used for various purposes. (2) It is shown how the constraint patterns can be mined form interaction data, and how they can be used for various applications. (3) A case study is presented which uses behavioural constraint patterns to profile user interactions in a communication network.

**Keywords:** Sequence mining · Network analysis

## 1 Introduction

Networks are often formed by the interaction of various actors. For example, social networks grow based on friendship or interested-based relations, forum posts and emails link users according to their communication patterns, and citation networks are formed through authors referencing peers in their field. Typically, the construction of these networks is based on either undirected, or directed edges with weights. Furthermore, many network techniques focus on static relationships, I.e., the evolution over time is not investigated. However, a range of new techniques emerged recently that focus on the time-aspect of a network. Most notably, the use of motifs [11], and streams [8] allow to capture the evolution of a network over time. In this paper, we describe a new approach based on behavioral constraints, I.e., constraints based on sequence patterns that allow to describe the order of the interactions of nodes.

Y. Ohsawa et al. (Eds.): JSAI 2019, AISC 1128, pp. 3–13, 2020.
https://doi.org/10.1007/978-3-030-39878-1_1

We investigate how they can be constructed from a network dataset, and use the various patterns to describe the evolution of the network over time. In particular, we apply the sequence mining method to the question-and-answer interaction-based network. Our preliminary results show that profiling network interactions patterns with sequence mining enables track the behaviour of nodes in a transactional network without relying on the typical partial-order based results.

This paper is structured as follows. In Sect. 2, the methodology is presented to mine constraints from network data. Next, Sect. 3 reports on the application on a real-life dataset. Finally, Sect. 4 concludes the paper and reports on the future directions.

## 2 Behavioural Constraints for Sequence Interactions

In this section, an introduction to behavioural constraints is given, which includes a motivation which ones are suitable for network interactions. Next, the method for mining them is discussed. Finally, a number of potential applications are delineated.

### 2.1 Constraint Set

Behavioural constraint templates have been long used in various areas of computer science. Most notably, a comprehensive set of Linear Temporal Logic (LTL) templates was proposed for the formal verification of program execution [6]. LTL provides an adequate formalism to search for various temporal properties, such as whether something happens eventually, next, and so on, and can be used in conjunction with typical logical operators to construct expressive relations. The initial set was extended to include various other relations, most notably unary ones. These constraints were later adapted towards the case of process modelling [13]. While initially proposed as LTL formulae which are convertible to Büchi automata, finite trace equivalent regular expressions were introduced in [2] and [15]. Models existing of multiple constraints at the same time can be obtained by conjoining the automata to obtain a global language or automaton, over which all constraints hold.

In Table 1, an overview of the most-commonly used constraints in literature. They are organized according to 7 different categories, including unary and binary constraints. Most notably, the binary constraints follow a hierarchy which is reported in [2] and which covers unordered relations up to chain ordered relations (using the next operator) at the top. Besides, the inclusion of negative constraints is unique, as typically only patterns of positive relations are reported. Negative constraints capture behaviour that has not occurred. Including negative behaviour can be used to find relations that are not apparent at first sight, e.g., in Fig. 1, the fact that nodes A and E are both present in the sequence of C, but do not have interactions themselves, still allows the inference of not succession(A, E). While only unary and binary constraints are included, it is also

**Table 1.** An overview of Declare constraint templates with their corresponding LTL formula and regular expression.

| Type | Template | LTL formula [12] | Regular expression [15] |
|---|---|---|---|
| Unary | Existence(A, n) | $\Diamond(A \land \bigcirc(\mathit{existence}(n-1, A)))$ | .*(A.*){n} |
| | Absence(A, n) | $\neg \mathit{existence}(n, A)$ | [^A]*(A?[^A]*){n-1} |
| | Exactly(A, n) | $\mathit{existence}(n, A) \land \mathit{absence}(n+1, A)$ | [^A]*(A[^A]*){n} |
| | Init(A) | $A$ | (A.*)? |
| | Last(A) | $\Box(A \implies \neg X \neg A)$ | .*A |
| Unordered | Responded existence(A, B) | $\Diamond A \implies \Diamond B$ | [^A]*((A.*B.*) \|(B.*A.*))? |
| | Co-existence(A, B) | $\Diamond A \iff \Diamond B$ | [^AB]*((A.*B.*) \|(B.*A.*))? |
| Simple ordered | Response(A, B) | $\Box(A \implies \Diamond B)$ | [^A]*(A.*B)*[^A]* |
| | Precedence(A, B) | $(\neg B\, U\, A) \lor \Box(\neg B)$ | [^B]*(A.*B)*[^B]* |
| | Succession(A, B) | $\mathit{response}(A, B) \land \mathit{precedence}(A, B)$ | [^AB]*(A.*B)*[^AB]* |
| Alternating ordered | Alternate response(A, B) | $\Box(A \implies \bigcirc(\neg A\, U\, B))$ | [^A]*(A[^A]*B[^A]*)* |
| | Alternate precedence(A, B) | $\mathit{precedence}(A, B) \land \Box(B \implies \bigcirc(\mathit{precedence}(A, B))$ | [^B]*(A[^B]*B[^B]*)* |
| | Alternate succession(A, B) | $\mathit{altresponse}(A, B) \land \mathit{precedence}(A, B)$ | [^AB]*(A[^AB]*B[^AB]*)* |
| Chain ordered | Chain response(A, B) | $\Box(A \implies \bigcirc B)$ | [^A]*(AB[^A]*)* |
| | Chain precedence(A, B) | $\Box(\bigcirc B \implies A)$ | [ ^B]*(AB[ ^B]*)* |
| | Chain succession(A, B) | $\Box(A \iff \bigcirc B)$ | [^AB]*(AB[^AB]*)* |
| Negative | Not co-existence(A, B) | $\neg(\Diamond A \land \Diamond B)$ | [^AB]*((A[ ^B]*) \|(B[^A]*))? |
| | Not succession(A, B) | $\Box(A \implies \neg(\Diamond B))$ | [^A]*(A[ ^B]*)* |
| | Not chain succession(A, B) | $\Box(A \implies \neg(\bigcirc B))$ | [^A]*(A+[^AB][^A]*)*A* |
| Choice | Choice(A, B) | $\Diamond A \lor \Diamond B$ | .*[AB].* |
| | Exclusive choice(A, B) | $(\Diamond A \lor \Diamond B) \land \neg(\Diamond A \land \Diamond B)$ | ([ ^B]*A[ ^B]*)<br>\|.*[AB].*([^A]*B[^A]*) |

possible to use target-branched constraints to model interactions between more than 2 nodes [4]. Motifs can be handy to capture various profiles of directed arcs between 2 or more nodes [10], but with the constraint sets it is possible to create even more intricate profiles of arc relations between nodes.

Despite not being useful for capturing interaction effects, the unary constraints can be used for adding information to a node's feature vector in case any exist. I.e., if a particular node is always occurring first in a sequence, this might signify a particular pattern, e.g., a person reporting recently-occurred disasters.

Not every constraint is suitable for binary interaction within a network context, i.e., not chain succession is, in general, not suitable for profiling behavior, as it holds in many situations. Besides, absence is hard to identify unless a particular

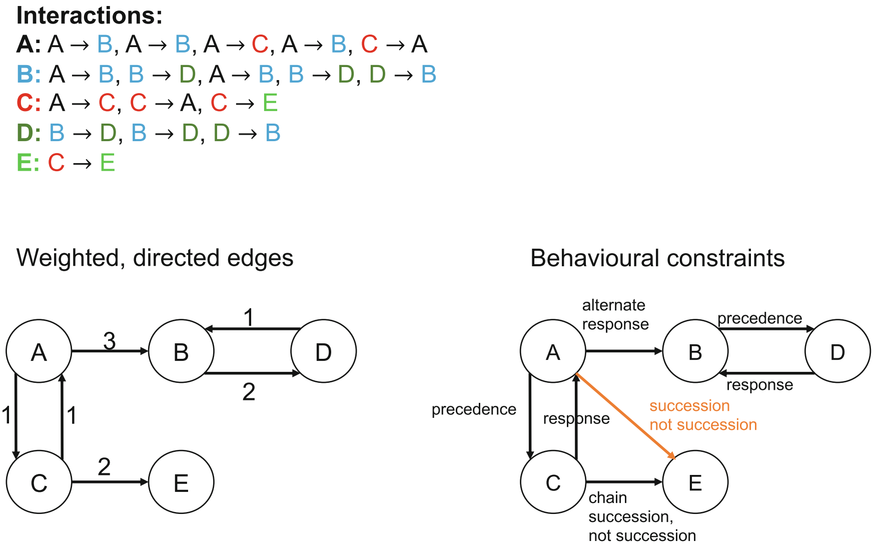

**Fig. 1.** Running example

node is scrutinized for this behaviour in the sequence of another node. Exclusive choice and not co-existence are similar in this respect, where the latter does not require the presence of either. Similar to not chain succession, this might lead to the discovery of many frequently non-occurring pairs.

### 2.2 Mining the Patterns

We define transactional network data as an ordered set of interactions $T$ between nodes from the set $N$, where each transaction is a tuple $(n_1, n_2, ts) \in T$ with $n_1 \in N$ the initiating node, $n_2 \in N$ the receiving node, and $ts \in \mathbb{N}^+$ a timestamp. $T$ can be read sequentially, where each node $n \in N$ has a sequence $s_n \subset 2^{|N|}$ that is extended whenever a transaction $t \in T$ is for that node is witnessed. I.e., $s_n$ gets extended with $\langle n, n_o \rangle$ whenever $n$ is the initiating node, and with $\langle n_o, n \rangle$ when $n$ is on the receiving end given another node $n_o \in N$.

By using the interesting Behavioural Constraint Miner [3], we can mine all patterns in a sequence $s_n$ to obtain a set of constraints $C_{s_n}$. Note, however, that if a given binary constraint $c(n, n_2) \in C_{s_n}$ holds for $n$ in its own sequence, this still has to be verified with the sequence of the other node. If $c(n, n_2)$ is not present in that sequence, the constraints do not hold. Consider for example the interaction in Fig. 1. Despite the evidence in the sequence of A that there exists an alternate succession relationship between A and B due to the alternating ABABAAB pattern, the sequence of B rather indicates that other occurrences of B happen in between (e.g. B→D), breaking the pattern. Hence, a final step is required to recursively ensure that $C_n = \{c \mid c \in C_n \wedge c \in C_{n_i} \forall n_i \in \mathcal{N}(n) \vee c \notin C_n \wedge c \notin C_{n_i} \forall n_i \in \mathcal{N}(n)\}$ where $\mathcal{N}(n) \subseteq N$ denotes the neighbourhood of node

$n$ to check that all constraint pertaining to $n$ are either both in its constraint set and the constraint set of its neighbours to avoid conflict, or that it is present in an unrelated node (e.g. the connection succession(A, E) in Fig. 1). To conclude the discovery of sequence templates from the network interactions, the sets $C_n$ are pruned according to the constraint hierarchy.

### 2.3 Applications

The mining of interactions in a network as sequences has several applications. Most notably, the sequence information can be used for analysing the patterns that exist between nodes, and their evolution over time. By tracking what patterns exist, and whether they return over time gives an overview of how certain relations change and what the underlying sequential behaviour is. In this case, unary constraints might not be useful, but especially the hierarchy of binary constraints can pin down how strong the relationship between two nodes is.

Next, the sequence patterns can be used as features of a node to obtain vector representations of nodes or relations between nodes. The presence of relations can be stored in a binary vector for relations, or the number of relations of each type can be stored for nodes. In this case, also unary constraints help define the node in terms of where in a sequence (init/last), how often (existence/exactly), besides with what other nodes the node is interacting (binary constraints). The features can be used towards node classification [1] or node relation prediction [9].

Given that all constraints have formal semantics in LTL or regular expressions, it is possible to exploit these in order to simulate the sequence behaviour later. These languages support state machine models which can generate strings of nodes by creating a global state machine that models all the constraints' behaviour. This can potentially feed into techniques such as node2vec [7], which typically employ random walks [14] for getting node representations.

Finally, by using the transitivity properties of the constraints, link inference/prediction [9] can also be made.

## 3 Application

In this section, the approach is applied to a real-life dataset in order to retrieve the evolution of constraints over time to profile the network interactions from a behavioural perspective.

### 3.1 Data

We apply the sequence method to the Math Overflow dataset, as used in [11]. On the Overflow web sites, users post questions and receive answers from other users, and users may comment on both questions and answers. We derive a transactional network by creating an edge ($u$, $v$, $t$) if, at time $t$, user $u$: (1) posts an answer to user $v'$s question, (2) comments on user $v'$s question, or

(3) comments on user $v'$s answer. The data contains 24,818 nodes with 506,550 interactions over 2,350 days and deals with question-and-answer data from users regarding mathematical problems.

We retrieve the constraints over the dataset by splitting the interactions into contingent blocks of a varying time length. In this case, we used blocks of 4 h (14,102 blocks), 2 days (1,175 blocks), 100 days (23 blocks), and 1,000 days (2 blocks) in order to track the evolution of the constraints. The evolution is captured by tallying the shift in constraint type between nodes present in subsequent blocks. For this analysis, we limit the constraint set to the 7 most common sequence patterns. The following coding was used:

- 1: not succession
- 2: precedence
- 3: alternate precedence
- 4: chain precedence
- 5: response
- 6: alternate response
- 7: chain response

In the columns, 0 stands for the absence of a constraint in subsequent time blocks between nodes that both reoccur.

In order to understand the difference in behaviour of various user types, we analyse the evolution of various node type and look into nodes that have a degree less than or equal to 3 (low involvement), a degree of 4–10 (medium involvement), and a degree of more than 10 (high involvement - 5,821 nodes). The results can be found in Tables 2 and 3 for incoming and outgoing relations respectively. Besides, the node with the highest authority score [5] is included as well, in order to illustrate how the most important node in terms of the dispersion of information acts within the web site. The results can be found in Table 4. The colours denote the place in the distribution, where red is higher and green is lower. Scores with different colours and equal scores indicate a difference in value behind the significant digits.

### 3.2 Interpretation

**Influence of Degree.** From Tables 2 and 3, we can learn that most relationships between nodes in subsequent blocks vanish, however, this seems to be more of an issue for blocks that are longer (>4 h) especially for nodes with a lower degree ($\leq$10). It seems a majority of the activity is relatively one-off. In general, all alternate relations happen sparsely given their very low share in the overall constraint evolution tally, meaning most nodes only interact once during a particular time block.

More interesting patterns can be observed when looking at how different constraint types evolve over time. A vast majority of chain precedence relations, a result of uninterrupted interaction between nodes, are replaced by not succession (4→1) and vice versa (1→4), and a shift from not succession to chain response

**Table 2.** An overview of the proportion of incoming constraints that shift from one sequence pattern into another for nodes with different degrees between different time blocks of varying lengths.

| | | 4 hours | | | | | | | | 2 days | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **IN** | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| ≤3 | **1** | 0.072 | 0.025 | 0.025 | 0.000 | 0.100 | 0.007 | 0.000 | 0.086 | 0.214 | 0.022 | 0.028 | 0.000 | 0.040 | 0.013 | 0.000 | 0.028 |
| | **2** | 0.063 | 0.013 | 0.011 | 0.000 | 0.011 | 0.001 | 0.000 | 0.012 | 0.144 | 0.018 | 0.022 | 0.000 | 0.004 | 0.008 | 0.000 | 0.004 |
| | **3** | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.001 | 0.001 | 0.000 | 0.005 | 0.000 | 0.000 | 0.001 |
| | **4** | 0.076 | 0.064 | 0.032 | 0.001 | 0.038 | 0.008 | 0.001 | 0.034 | 0.079 | 0.038 | 0.030 | 0.001 | 0.010 | 0.014 | 0.001 | 0.015 |
| | **5** | 0.058 | 0.023 | 0.004 | 0.001 | 0.008 | 0.007 | 0.000 | 0.011 | 0.116 | 0.018 | 0.007 | 0.000 | 0.006 | 0.021 | 0.000 | 0.003 |
| | **6** | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **7** | 0.050 | 0.042 | 0.011 | 0.000 | 0.012 | 0.015 | 0.000 | 0.075 | 0.027 | 0.021 | 0.006 | 0.000 | 0.004 | 0.008 | 0.000 | 0.019 |
| 4–10 | **1** | 0.086 | 0.018 | 0.037 | 0.000 | 0.047 | 0.011 | 0.000 | 0.059 | 0.229 | 0.014 | 0.038 | 0.001 | 0.011 | 0.021 | 0.000 | 0.016 |
| | **2** | 0.106 | 0.013 | 0.028 | 0.000 | 0.011 | 0.007 | 0.000 | 0.021 | 0.201 | 0.018 | 0.027 | 0.000 | 0.005 | 0.011 | 0.000 | 0.008 |
| | **3** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.003 | 0.000 | 0.003 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 |
| | **4** | 0.062 | 0.023 | 0.031 | 0.000 | 0.018 | 0.010 | 0.000 | 0.035 | 0.042 | 0.010 | 0.014 | 0.000 | 0.003 | 0.011 | 0.000 | 0.008 |
| | **5** | 0.091 | 0.010 | 0.010 | 0.000 | 0.007 | 0.014 | 0.000 | 0.013 | 0.164 | 0.014 | 0.007 | 0.000 | 0.003 | 0.019 | 0.000 | 0.006 |
| | **6** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **7** | 0.062 | 0.019 | 0.012 | 0.000 | 0.008 | 0.021 | 0.000 | 0.103 | 0.038 | 0.007 | 0.006 | 0.000 | 0.001 | 0.011 | 0.000 | 0.019 |
| >10 | **1** | 0.143 | 0.028 | 0.028 | 0.001 | 0.024 | 0.017 | 0.001 | 0.027 | 0.218 | 0.014 | 0.018 | 0.001 | 0.002 | 0.017 | 0.000 | 0.002 |
| | **2** | 0.146 | 0.024 | 0.024 | 0.001 | 0.008 | 0.011 | 0.000 | 0.013 | 0.274 | 0.017 | 0.023 | 0.001 | 0.001 | 0.007 | 0.001 | 0.002 |
| | **3** | 0.003 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.007 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **4** | 0.046 | 0.028 | 0.015 | 0.000 | 0.010 | 0.010 | 0.000 | 0.015 | 0.016 | 0.002 | 0.002 | 0.000 | 0.001 | 0.002 | 0.000 | 0.001 |
| | **5** | 0.142 | 0.022 | 0.008 | 0.000 | 0.006 | 0.021 | 0.001 | 0.009 | 0.278 | 0.017 | 0.006 | 0.001 | 0.001 | 0.024 | 0.001 | 0.002 |
| | **6** | 0.003 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.006 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **7** | 0.050 | 0.026 | 0.009 | 0.000 | 0.005 | 0.019 | 0.000 | 0.050 | 0.021 | 0.002 | 0.002 | 0.000 | 0.000 | 0.004 | 0.000 | 0.003 |

| | | 100 days | | | | | | | | 1000 days | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| ≤3 | **1** | 0.330 | 0.022 | 0.024 | 0.001 | 0.001 | 0.017 | 0.000 | 0.000 | 0.277 | 0.022 | 0.027 | 0.000 | 0.000 | 0.018 | 0.000 | 0.000 |
| | **2** | 0.283 | 0.020 | 0.028 | 0.001 | 0.001 | 0.002 | 0.000 | 0.000 | 0.312 | 0.031 | 0.025 | 0.001 | 0.000 | 0.003 | 0.000 | 0.000 |
| | **3** | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **4** | 0.011 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **5** | 0.212 | 0.013 | 0.008 | 0.001 | 0.000 | 0.013 | 0.000 | 0.000 | 0.230 | 0.022 | 0.008 | 0.001 | 0.000 | 0.011 | 0.000 | 0.000 |
| | **6** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **7** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 4–10 | **1** | 0.309 | 0.023 | 0.032 | 0.001 | 0.000 | 0.029 | 0.000 | 0.000 | 0.256 | 0.024 | 0.040 | 0.002 | 0.000 | 0.032 | 0.001 | 0.000 |
| | **2** | 0.264 | 0.021 | 0.027 | 0.001 | 0.000 | 0.005 | 0.000 | 0.000 | 0.265 | 0.027 | 0.039 | 0.002 | 0.000 | 0.006 | 0.001 | 0.000 |
| | **3** | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.008 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **4** | 0.003 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **5** | 0.231 | 0.017 | 0.007 | 0.000 | 0.000 | 0.015 | 0.000 | 0.000 | 0.231 | 0.022 | 0.011 | 0.001 | 0.000 | 0.023 | 0.001 | 0.000 |
| | **6** | 0.003 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **7** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| >10 | **1** | 0.194 | 0.024 | 0.033 | 0.001 | 0.000 | 0.034 | 0.001 | 0.000 | 0.160 | 0.032 | 0.046 | 0.001 | 0.000 | 0.047 | 0.001 | 0.000 |
| | **2** | 0.252 | 0.032 | 0.047 | 0.001 | 0.000 | 0.012 | 0.001 | 0.000 | 0.216 | 0.043 | 0.062 | 0.001 | 0.000 | 0.019 | 0.001 | 0.000 |
| | **3** | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.004 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **4** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **5** | 0.254 | 0.033 | 0.012 | 0.001 | 0.000 | 0.048 | 0.001 | 0.000 | 0.222 | 0.043 | 0.017 | 0.001 | 0.000 | 0.068 | 0.002 | 0.000 |
| | **6** | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **7** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

**Table 3.** An overview of the proportion of outgoing constraints that shift from one sequence pattern into another for nodes with different degrees between different time blocks of varying lengths.

| | | 100 days | | | | | | | | 1000 days | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OUT** | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| ≤3 | **1** | 0.136 | 0.058 | 0.017 | 0.000 | 0.059 | 0.014 | 0.002 | 0.068 | 0.133 | 0.021 | 0.030 | 0.000 | 0.017 | 0.016 | 0.001 | 0.030 |
| | **2** | 0.080 | 0.041 | 0.013 | 0.000 | 0.005 | 0.015 | 0.001 | 0.006 | 0.216 | 0.025 | 0.020 | 0.000 | 0.004 | 0.009 | 0.000 | 0.007 |
| | **3** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **4** | 0.013 | 0.086 | 0.009 | 0.000 | 0.013 | 0.008 | 0.000 | 0.010 | 0.018 | 0.019 | 0.004 | 0.000 | 0.004 | 0.004 | 0.000 | 0.006 |
| | **5** | 0.082 | 0.021 | 0.004 | 0.000 | 0.005 | 0.011 | 0.001 | 0.008 | 0.236 | 0.012 | 0.004 | 0.000 | 0.003 | 0.025 | 0.001 | 0.005 |
| | **6** | 0.002 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **7** | 0.022 | 0.114 | 0.001 | 0.000 | 0.002 | 0.029 | 0.000 | 0.048 | 0.028 | 0.040 | 0.003 | 0.000 | 0.002 | 0.027 | 0.001 | 0.023 |
| 4–10 | **1** | 0.167 | 0.060 | 0.018 | 0.001 | 0.028 | 0.013 | 0.001 | 0.026 | 0.152 | 0.013 | 0.021 | 0.000 | 0.007 | 0.017 | 0.001 | 0.008 |
| | **2** | 0.116 | 0.038 | 0.017 | 0.000 | 0.007 | 0.007 | 0.001 | 0.007 | 0.253 | 0.022 | 0.018 | 0.001 | 0.003 | 0.009 | 0.002 | 0.004 |
| | **3** | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **4** | 0.032 | 0.063 | 0.010 | 0.000 | 0.008 | 0.011 | 0.000 | 0.010 | 0.022 | 0.014 | 0.003 | 0.000 | 0.002 | 0.007 | 0.001 | 0.002 |
| | **5** | 0.119 | 0.036 | 0.003 | 0.000 | 0.007 | 0.017 | 0.001 | 0.007 | 0.268 | 0.022 | 0.004 | 0.001 | 0.003 | 0.023 | 0.002 | 0.005 |
| | **6** | 0.003 | 0.001 | 0.001 | 0.000 | 0.001 | 0.001 | 0.000 | 0.001 | 0.006 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **7** | 0.033 | 0.061 | 0.005 | 0.000 | 0.004 | 0.023 | 0.001 | 0.034 | 0.033 | 0.015 | 0.004 | 0.000 | 0.001 | 0.018 | 0.002 | 0.009 |
| >10 | **1** | 0.138 | 0.026 | 0.029 | 0.001 | 0.026 | 0.017 | 0.000 | 0.030 | 0.220 | 0.014 | 0.018 | 0.001 | 0.002 | 0.017 | 0.000 | 0.002 |
| | **2** | 0.144 | 0.023 | 0.025 | 0.001 | 0.008 | 0.010 | 0.000 | 0.014 | 0.272 | 0.017 | 0.023 | 0.001 | 0.001 | 0.008 | 0.000 | 0.002 |
| | **3** | 0.003 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.007 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **4** | 0.048 | 0.026 | 0.017 | 0.000 | 0.011 | 0.010 | 0.000 | 0.017 | 0.017 | 0.002 | 0.003 | 0.000 | 0.001 | 0.002 | 0.000 | 0.001 |
| | **5** | 0.139 | 0.021 | 0.008 | 0.000 | 0.006 | 0.020 | 0.001 | 0.010 | 0.275 | 0.016 | 0.006 | 0.001 | 0.001 | 0.024 | 0.001 | 0.002 |
| | **6** | 0.002 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.006 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **7** | 0.052 | 0.023 | 0.010 | 0.000 | 0.005 | 0.019 | 0.000 | 0.054 | 0.021 | 0.002 | 0.002 | 0.000 | 0.000 | 0.004 | 0.000 | 0.003 |

| | | 100 days | | | | | | | | 1000 days | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| ≤3 | **1** | 0.180 | 0.010 | 0.014 | 0.000 | 0.000 | 0.017 | 0.002 | 0.001 | 0.211 | 0.007 | 0.023 | 0.000 | 0.000 | 0.028 | 0.002 | 0.000 |
| | **2** | 0.297 | 0.012 | 0.023 | 0.000 | 0.000 | 0.008 | 0.000 | 0.000 | 0.265 | 0.016 | 0.022 | 0.000 | 0.000 | 0.010 | 0.000 | 0.000 |
| | **3** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **4** | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **5** | 0.359 | 0.016 | 0.003 | 0.000 | 0.000 | 0.037 | 0.001 | 0.001 | 0.338 | 0.020 | 0.002 | 0.000 | 0.000 | 0.043 | 0.001 | 0.000 |
| | **6** | 0.002 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.004 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **7** | 0.007 | 0.001 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.001 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 4–10 | **1** | 0.167 | 0.009 | 0.015 | 0.000 | 0.000 | 0.018 | 0.001 | 0.000 | 0.179 | 0.014 | 0.022 | 0.000 | 0.000 | 0.027 | 0.001 | 0.000 |
| | **2** | 0.305 | 0.021 | 0.029 | 0.000 | 0.000 | 0.009 | 0.000 | 0.000 | 0.271 | 0.024 | 0.032 | 0.000 | 0.000 | 0.015 | 0.000 | 0.000 |
| | **3** | 0.004 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **4** | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **5** | 0.332 | 0.024 | 0.007 | 0.000 | 0.000 | 0.039 | 0.001 | 0.000 | 0.302 | 0.029 | 0.008 | 0.000 | 0.000 | 0.050 | 0.002 | 0.000 |
| | **6** | 0.007 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.011 | 0.001 | 0.001 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 |
| | **7** | 0.004 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| >10 | **1** | 0.198 | 0.025 | 0.034 | 0.001 | 0.000 | 0.034 | 0.001 | 0.000 | 0.168 | 0.033 | 0.047 | 0.001 | 0.000 | 0.048 | 0.001 | 0.000 |
| | **2** | 0.251 | 0.032 | 0.047 | 0.001 | 0.000 | 0.012 | 0.001 | 0.000 | 0.215 | 0.044 | 0.063 | 0.001 | 0.000 | 0.018 | 0.001 | 0.000 |
| | **3** | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **4** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **5** | 0.251 | 0.032 | 0.012 | 0.001 | 0.000 | 0.047 | 0.001 | 0.000 | 0.214 | 0.043 | 0.017 | 0.001 | 0.000 | 0.066 | 0.001 | 0.000 |
| | **6** | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **7** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

**Table 4.** An overview of the proportion of constraints that shift from one sequence pattern into another, both for incoming and outgoing constraints of the node with the highest authority score in the network.

| | | 4 hours | | | | | | | | 2 days | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **IN** | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| | **1** | 0.172 | 0.020 | 0.019 | 0.000 | 0.007 | 0.019 | 0.000 | 0.009 | 0.195 | 0.020 | 0.026 | 0.001 | 0.001 | 0.025 | 0.001 | 0.001 |
| | **2** | 0.219 | 0.021 | 0.026 | 0.001 | 0.003 | 0.008 | 0.001 | 0.005 | 0.267 | 0.028 | 0.035 | 0.001 | 0.001 | 0.011 | 0.001 | 0.002 |
| | **3** | 0.006 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.006 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **4** | 0.040 | 0.007 | 0.006 | 0.000 | 0.003 | 0.005 | 0.000 | 0.005 | 0.010 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **5** | 0.224 | 0.020 | 0.007 | 0.000 | 0.003 | 0.026 | 0.001 | 0.004 | 0.264 | 0.028 | 0.009 | 0.001 | 0.001 | 0.036 | 0.001 | 0.002 |
| | **6** | 0.002 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **7** | 0.062 | 0.010 | 0.008 | 0.000 | 0.002 | 0.012 | 0.001 | 0.015 | 0.014 | 0.001 | 0.001 | 0.000 | 0.000 | 0.002 | 0.000 | 0.001 |

| **OUT** | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | 0.195 | 0.018 | 0.026 | 0.001 | 0.009 | 0.025 | 0.000 | 0.010 | 0.215 | 0.020 | 0.029 | 0.001 | 0.001 | 0.030 | 0.001 | 0.001 |
| | **2** | 0.201 | 0.016 | 0.021 | 0.001 | 0.005 | 0.010 | 0.001 | 0.008 | 0.249 | 0.025 | 0.033 | 0.001 | 0.001 | 0.011 | 0.001 | 0.001 |
| | **3** | 0.006 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.007 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **4** | 0.046 | 0.006 | 0.011 | 0.000 | 0.002 | 0.011 | 0.000 | 0.004 | 0.009 | 0.001 | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **5** | 0.203 | 0.017 | 0.007 | 0.001 | 0.005 | 0.027 | 0.002 | 0.006 | 0.263 | 0.025 | 0.007 | 0.001 | 0.001 | 0.038 | 0.001 | 0.001 |
| | **6** | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| | **7** | 0.051 | 0.005 | 0.008 | 0.000 | 0.001 | 0.011 | 0.000 | 0.014 | 0.008 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |

| | | 100 days | | | | | | | | 1000 days | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **IN** | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| | **1** | 0.099 | 0.023 | 0.040 | 0.001 | 0.000 | 0.051 | 0.002 | 0.001 | 0.011 | 0.027 | 0.013 | 0.000 | 0.000 | 0.047 | 0.001 | 0.001 |
| | **2** | 0.185 | 0.043 | 0.087 | 0.001 | 0.000 | 0.032 | 0.001 | 0.001 | 0.077 | 0.059 | 0.049 | 0.000 | 0.003 | 0.121 | 0.003 | 0.001 |
| | **3** | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **4** | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **5** | 0.209 | 0.053 | 0.023 | 0.001 | 0.000 | 0.121 | 0.003 | 0.001 | 0.138 | 0.101 | 0.039 | 0.000 | 0.000 | 0.285 | 0.005 | 0.000 |
| | **6** | 0.005 | 0.001 | 0.003 | 0.000 | 0.000 | 0.003 | 0.000 | 0.000 | 0.001 | 0.003 | 0.003 | 0.000 | 0.000 | 0.008 | 0.000 | 0.000 |
| | **7** | 0.002 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |

| **OUT** | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | 0.175 | 0.059 | 0.069 | 0.001 | 0.000 | 0.060 | 0.000 | 0.000 | 0.070 | 0.026 | 0.219 | 0.006 | 0.000 | 0.100 | 0.000 | 0.001 |
| | **2** | 0.182 | 0.056 | 0.083 | 0.002 | 0.000 | 0.017 | 0.001 | 0.000 | 0.061 | 0.025 | 0.269 | 0.015 | 0.000 | 0.035 | 0.001 | 0.000 |
| | **3** | 0.003 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.003 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **4** | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **5** | 0.152 | 0.049 | 0.026 | 0.001 | 0.000 | 0.053 | 0.000 | 0.000 | 0.019 | 0.015 | 0.060 | 0.001 | 0.000 | 0.061 | 0.000 | 0.000 |
| | **6** | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | **7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.003 | 0.000 | 0.003 | 0.000 | 0.000 |

(1→7) mainly for a degree lower or equal to 3 and 4 h blocks and incoming relationships. The same trend holds for 2–1000 days, but here there is a higher rate of precedence to precedence (2→2). This indicates that in the short term, mostly the absence of response of the node (1) is eventually replaced by a very close interaction (4) in the short run, and that for the longer run (100–1000 days), the 2→2 indicates that the other party initiates a message often without there being a strict conversion towards other constraints such as response (5).

For nodes with a degree between 4 and 10, the evolutions look different with a strong persistence of chain response (7→7) in the short run (4 h), and a similar

(1/2→1/2) in the longer run. The former indicates that nodes tend to maintain a close interaction over short time spans, which is substituted by more sporadic interchange for longer time spans.

For nodes with a high degree, 7→7 for 4 h, and 5→5 relations are present for a large share of the constraint evolutions. Hence, these nodes tend to maintain close contact at first, and afterwards remain the consequent in a response relationship indicate being especially forthcoming in terms of responding to other users.

Overall, users with a high degree tend to be more involved in mutually positive relationships (not 1), and reciprocate by being mostly involved in response relationships (5–7).

For outgoing constraints, we see a different picture, where incoming chain responses (7) are often converted into not succession (1) by the receiving party, which persists for 4 h and 2 days at a degree of 3. For longer periods, 5→5 or, response to response is prevalent.

For nodes with a degree between 4 and 10, most initial relations end in not succession, later converging to a similar profile as nodes with a lower degree, and a high proportion of 5→5. The same holds for nodes with a high degree, where initially 7→7 is prevalent.

These observations clarify which nodes tend to be the source of cutting of contact in short time blocks (mostly with degree ≤10), and explain the interpretation of the incoming nodes earlier. In this respect, the evolution of constraints can be tracked between different types of nodes in terms of degree

**Authority.** The node with the highest authority score which has a degree of 2,680 does not follow the trends of the other nodes discussed above. Again, the high proportion of 5→5 evolution is present for incoming relationships, but instead of materialising later on, this happens even for shorter time blocks (4 h). This indicates that the authority will respond regardless of any established relationship.

## 4 Conclusion and Future Work

In this paper, we have shown how mining network interaction patterns can be profiled using sequence mining techniques. We apply the sequence mining method to the question-and-answer interaction-based network. Our preliminary results show that employing sequence patterns enables us track the behaviour of nodes in a transactional network and summarize their interactions without relying on the typical partial-order based results that are offered in sequence mining, while still going beyond the typical general nature of motifs that focus on directed arcs between 2 or 3 actors [11] . In a small experimental evaluation, we demonstrate the usefulness of the approach in the context of message board analysis.

For future work, we envision to focus on testing the patterns in the context of feature engineering, and link inference.

## References

1. Bhagat, S., Cormode, G., Muthukrishnan, S.: Node classification in social networks. In: Social Network Data Analytics, pp. 115–148. Springer (2011)
2. Ciccio, C.D., Mecella, M.: A two-step fast algorithm for the automated discovery of declarative workflows. In: IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013, Singapore, 16–19 April, 2013, pp. 135–142 (2013)
3. De Smedt, J., Deeva, G., De Weerdt, J.: Behavioral constraint template-based sequence classification. In: ECML/PKDD (2). LNCS, vol. 10535, pp. 20–36. Springer (2017)
4. Di Ciccio, C.D., Maggi, F.M., Mendling, J.: Efficient discovery of target-branched declare constraints. Inf. Syst. **56**, 258–283 (2016)
5. Ding, C.H.Q., Zha, H., He, X., Husbands, P., Simon, H.D.: Link analysis: hubs and authorities on the world wide web. SIAM Rev. **46**(2), 256–268 (2004)
6. Dwyer, M.B., Avrunin, G.S., Corbett, J.C.: Patterns in property specifications for finite-state verification. In: ICSE, pp. 411–420. ACM (1999)
7. Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864. ACM (2016)
8. Latapy, M., Viard, T., Magnien, C.: Stream graphs and link streams for the modeling of interactions over time. Social Netw. Analys. Min. **8**(1), 61:1–61:29 (2018)
9. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inform. Sci. Technol. **58**(7), 1019–1031 (2007)
10. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science **298**(5594), 824–827 (2002)
11. Paranjape, A., Benson, A.R., Leskovec, J.: Motifs in temporal networks. In: WSDM, pp. 601–610. ACM (2017)
12. Pesić, M.: Constraint-based work on management systems: shifting control to users. Ph.D. thesis, Eindhoven University of Technology (2008). 26
13. Pesic, M., van der Aalst, W.M.P.: A declarative approach for flexible business processes management. In: Business Process Management Workshops. LNCS, vol. 4103, pp. 169–180. Springer (2006)
14. Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: International Symposium on Computer and Information Sciences, pp. 284–293. Springer (2005)
15. Westergaard, M., Stahl, C., Reijers, H.A.: Unconstrainedminer: efficient discovery of generalized declarative process models. BPM Center Report BPM-13-28, BPMcenter.org, p. 28 (2013)

# Prediction of Onset of Lifestyle-Related Diseases Using Regular Health Checkup Data

Mitsuru Tsunekawa[1](✉), Natsuki Oka[1], Masahiro Araki[1], Motoshi Shintani[2], Masataka Yoshikawa[3], and Takeshi Tanigawa[4]

[1] Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto, Japan
m-tsune@ii.is.kit.ac.jp

[2] SG Holdings Group Health Insurance Association,
637, Suiginnya-cho, Shimogyo-ku, Kyoto, Japan

[3] Japan System Techniques Co., Ltd., 2-3-18, Nakanoshima, Kita-ku, Osaka, Japan

[4] Juntendo University, 2-1-1, Hongo, Bunkyo-ku, Tokyo, Japan

**Abstract.** This is an extension from a selected paper from JSAI2019. This study proposes a method to predict the onset of lifestyle-related diseases using periodical health checkup data. In this study, we carefully examine insurance claims data to identify onset of diseases and use the data for supervised learning. We aim to predict whether lifestyle-related diseases, except cancer, will develop within a year. We adopt the undersampling and bagging approach to address the class imbalance problem. The precision and recall of the proposed method are found to be 0.32 and 0.89, respectively. Compared with a baseline that sets thresholds for each examination item and considers their logical sum, our method achieves much higher precision while maintaining the recall; this allows suppression of the number of targets for health guidance, without increasingly neglecting those who are likely to become severely ill.

**Keywords:** Machine learning · Class imbalance · Medical informatics · Undersampling · Bagging

## 1 Introduction

Many people have recently begun using Internet mail orders, which has greatly increased the number of deliveries. Consequently, social interest in the work environment and health management of courier drivers is increasing. If appropriate health guidance can reduce the occurrence of lifestyle-related diseases (except cancer) among drivers and prevent severe idiopathic illnesses during driving, medical expenses and traffic accidents can be decreased. We, therefore, aim to use regular health checkup data of employees from a home delivery company to predict the onset of lifestyle-related diseases within a year as accurately as possible so that appropriate health guidance can be provided.

Y. Ohsawa et al. (Eds.): JSAI 2019, AISC 1128, pp. 14–26, 2020.
https://doi.org/10.1007/978-3-030-39878-1_2

There have been many studies on the prediction of diseases using medical data. For example, [1] created a risk equation using the Cox proportional hazards model to estimate colorectal cancer risk using the results of electronic medical records of a unified database in a specific country or region. Statistical analysis techniques have been used for cohort studies that predict the occurrence probability of myocardial infarction and cerebral infarction from health examination results [2]. The prediction target of [2] overlaps with this study, but the data and prediction methods are different. [3] proposed a model that predicts pneumonia hospitalization using the Lasso logistic regression of regular health checkup data; similar to our study, periodical health checkup data of healthy people are used for prediction here. However, our study differs in that it uses machine learning techniques according to data characteristics. Recently, many studies have used machine learning and data mining techniques to predict disease onset from medical data. [4] highlighted the superiority of machine learning techniques to predict cardiovascular risk from routine clinical data. [5] developed a model that uses a neural network to predict ventricular tachycardia one hour before occurrence. Because ventricular tachycardia can lead to sudden cardiac death, predicting the occurrence is useful to treat it immediately. There was also a study that developed a probability-based artificial neural network model, called NORF, for predicting risks of oyster norovirus outbreaks [6]. Similarly, prediction research using deep learning has been actively conducted [7,8]. In this study, to emphasize the explainability of prediction results, we did not use deep learning because its grounds for judgment tend to be black boxes. The data used in this study includes data on occupations other than drivers.

## 2 Target Data

### 2.1 Data Overview

We used insurance claim and regular health checkup data of employees from the SG Holdings Group. Health insurance claim data is created when a person is injured or ill and visits a medical institution, whereas health checkup data is created regularly (typically once a year). These two datasets were anonymized and linked with a hash code for unique identification of patients. In this study, disease onsets extracted from insurance claim data were used as correct answers for onset prediction; health checkup data was used as input for onset prediction. Health insurance claim data includes basic patient information such as sex, age, medical treatment date, diagnosed disease, medical procedure, and prescribed medicine. On the other hand, health checkup data includes information such as patient height, weight, blood pressure, and red blood cell count.

In this study, we analyzed insurance claim data from 2006 to 2018 and health checkup data from 1996 to 2017 for individuals aged 15–74 years. There were 961,906 sheets of health checkup data for 156,145 people, and 1,617,078 sheets of insurance claim data for 108,581 people.

## 2.2 Disease Names as Prediction Targets

An individual's diagnosed disease names were obtained by searching through the disease name codes in the health insurance claims data and cross-referencing with those in the ICD-10, which is the international statistical classification code of diseases and related health problems created by the World Health Organization. Table 1 presents the ICD-10 codes and the corresponding disease names to be predicted (hereinafter, referred to as "severe disease names").

**Table 1.** ICD-10 codes of severe disease names.

| ICD-10 | Disease name |
|---|---|
| E10 | Insulin dependent diabetes mellitus |
| E11 | Non-insulin dependent diabetes mellitus |
| E14 | Diabetes mellitus other than the above |
| I20 | Angina pectoris |
| I21, I22 | Acute myocardial infarction |
| I42 | Cardiomyopathy |
| I44 I49 | Arrhythmia, Conduction defects |
| I60, I690 | Subarachnoid hemorrhage |
| I61, I691 | Intracerebral hemorrhage |
| I63, I693 | Cerebral infarction |

## 2.3 Feature Values for Prediction

The following examination items from health checkup data were used as feature values for prediction. Health examination data included not only the numerical data of inspection results, but also the results of a questionnaire on lifestyle habits and the judgment results of six levels derived from the examination data by medical institutions. The six levels were normal, almost normal, guidance required, reexamination required, close examination required, and treatment required. Factors of abdominal girth and visual acuity judgment, heart rate, visual acuity judgment, fundus judgment, and metabolic judgment were removed because 50% or more values were missing. Health examination data also included findings freely described by doctors; however, we excluded them because natural language understanding is required for using free description.

Health examination data items used as input features are as follows:

Sex; Age; Height; Weight; Body fat percentage; Systolic blood pressure; Diastolic blood pressure; Number of red blood cells; Hemoglobin; Hematocrit; Platelet count; GOT; GPT; $\gamma$-GTP; Total cholesterol; HDL cholesterol; LDL cholesterol; Neutral fat; Uric acid; Creatinine; eGFR; HbA1c; Questions regarding insulin injections or medicines to lower blood sugar, medicines to lower blood pressure, ameliorate dyslipidemia, stroke, chronic renal failure, and anemia; Lifestyle questions related to smoking habits, weight change from the age of 20 years, exercise habits, walking habits, walking speed, weight change over the past year, eating speed, meals just before going to bed, after dinner snacks, skipping breakfast, drinking habits, amount of alcohol consumed, sleeping time, willingness to improve lifestyle habits, and willingness to receive health guidance; Judgments on urinary protein and urine sugar; Representative judgment; Judgments on physical measurements, hearing ability, blood pressures, anemia, liver function, renal function, uric acid and gout, blood sugar, sugar metabolism, and urinalysis; Examination judgment.

### 2.4 Characteristics of Data

The data typically had two characteristics. First is considerable imbalance; for example, in 2017, the proportion of people diagnosed with severe diseases was only 4.5%. Learning of such unbalanced data may be greatly affected by the properties of many negative examples, i.e., persons not diagnosed with severe diseases. Therefore, a method that can successfully learn this imbalanced data must be adopted.

Second, classifying data as positive or negative is not easy. In this study, we aimed to predict whether a person who is healthy at the time of a regular health checkup will be diagnosed with a severe diseases within a year of the checkup. Consequently, data was classified as positive if the person would fall sick within a year and as negative if not.

The point at which the target disease name first appeared in an employee's insurance claim data is not necessarily the point when he/she first developed the disease. It is not uncommon for individuals with previously diagnosed diseases to join a health insurance association, especially, in an industry with large personnel flow. However, because the data used in this study was from a health insurance association, it was only for the period of joining the association; the insurance claim data before entering the association could not be confirmed. The next section describes a way to address this problem.

## 3 Data Selection and Machine Learning

### 3.1 Data Selection

We address the classification problem of predicting whether individuals will suffer severe illness within one year by using medical examination data. The selection method for positive and negative data was as follows.

First, to address the previously mentioned data availability problem, we determined whether an individual's first-time diagnosis of a severe disease as per the insurance claim data was actually the first. We first calculated the hospital visit interval for the disease after the diagnosis. If the hospital visit interval was shorter than the interval between the day of joining the health insurance association and the day of first-time diagnosis of a severe disease, the diagnosis was considered to be the first; alternatively, if the visit interval was greater, it was considered not to be first. The visit interval was calculated using three interval data, which was considered as sufficient. The specific procedure was as follows (see also Fig. 1):

1. Extract the oldest data item (*) containing a severe disease name from the individual's insurance claim data.
2. Select three consecutive data items with the same disease name that are newer than the extracted data and calculate the hospital visit intervals.
3. Retrieve the individual's oldest insurance claim data (**) and calculate the interval between data (**) and data (*).
4. If the maximum of the three values calculated in "2." is smaller than that calculated in "3.," regard data (*) as the first-diagnosis data of the disease.

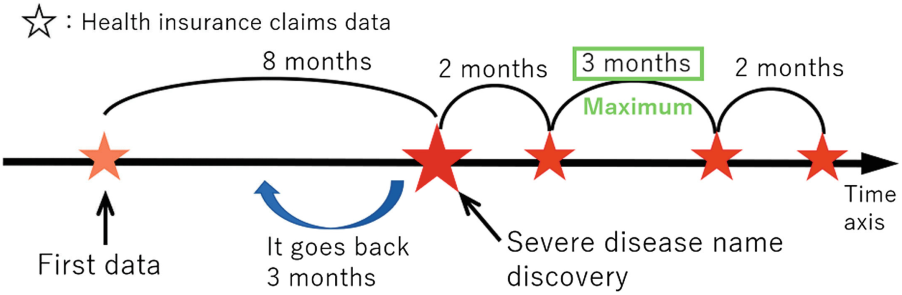

**Fig. 1.** Judging the genuineness of the first diagnosis.

Next, positive health checkup data was chosen from the data included within one year or less before the first appearance of the severe disease. When there were multiple data items in the range, we considered the oldest one. We also considered the amount of changes in the health checkup data, calculated the differences between the chosen data and the previous data and between the chosen data and the previous two data, and added them to the feature set (Fig. 2). Because some items in the health checkup data would have changed with the development of the disease, we considered that the discrimination accuracy would improve by explicitly adding the change to the feature set.

For negative data, we excluded data of individuals who had been diagnosed with a severe disease even once; only the remaining data was used. Furthermore, if there is no insurance claim data more than one year after a health checkup of an individual, we excluded the health heckup data from negative data because

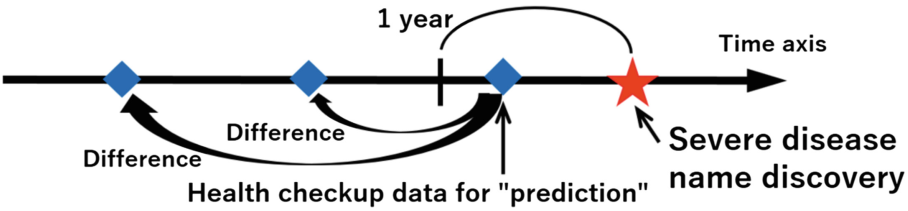
1 year
Time axis
Difference
Difference
Health checkup data for "prediction"
Severe disease name discovery

**Fig. 2.** Selection of positive data.

the person may have left his job and may have been diagnosed with a severe disease within a year. As with positive data, we calculated the differences using three consecutive health examination data and added them to the feature set (Fig. 3).

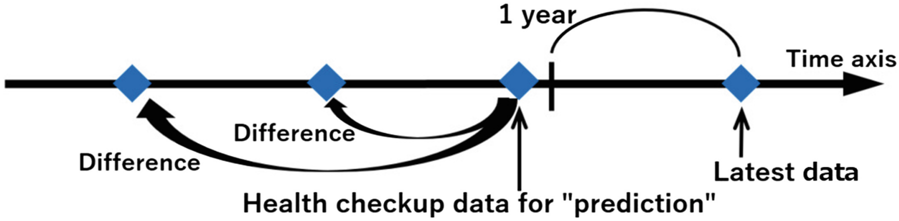

**Fig. 3.** Selection of negative data.

There were cases in which an individual had multiple medical examination data that fit the selection criteria. This was true for both positive and negative data. However, we used data for only one medical examination per individual to prevent data imbalance. If we did not select the oldest data for positive data, we used data after diagnoses of one of the severe diseases; however, negative data could be selected for any point in time.

Thus, we obtained 1255 positive data and 37664 negative data, with 132 features. The missing values were filled with median values.

### 3.2 Machine Learning for Imbalanced Data

This study deals with imbalanced data with only approximately 3% positive cases. Many methods for learning imbalanced data have been proposed such as cost-based strategies, undersampling, and oversampling. Cost-based strategies penalize misclassifications of the minority class more heavily than those of the majority class. Undersampling considers equal number of examples from both positive and negative classes by discarding training data from the majority class; whereas, oversampling achieves this by introducing duplicate instances from the minority class. Because duplication tends to cause overfitting, Synthetic Minority

Over-sampling Technique (SMOTE) [9] was proposed, which creates synthetic examples of the minority class by combining neighboring examples.

Wallace, Small, Brodley, and Trikalinos [10] argued that "undersampling + bagging" is the best strategy for handling imbalance. From the probabilistic interpretation of class imbalance, undersampling should be used to handle an imbalance in most scenarios; further, bagging should be used to reduce the variance of the approach. They also showed that "undersampling + bagging" should outperform cost-based strategies and SMOTE on datasets with high dimensionality, low prevalence, or small training set size. Based on their claim, we adopted "undersampling + bagging", and used decision trees, without pruning, as weak learners because they were unstable; combining them led to better performance.

# 4 Results and Discussion

## 4.1 Comparison with a Baseline Method

Five hundred weak learners were used in the proposed method. If the number of weak learners was changed to anywhere between 100 and 500, there was only slight change in recall and precision; however, if the number was less than 100, the precision reduced. Recall and precision did not change even when the number of classifiers was increased beyond 500; hence, the number of weak learners was set to 500. Because the decision tree used for the weak learners was an algorithm that is not affected by scale, data scaling was not performed. We adopted stratified 10-fold cross-validation as an evaluation method. Table 2 presents the confusion matrix of the proposed method. The positive precision and recall were 0.32 and 0.89, respectively.

**Table 2.** Confusion matrix of the proposed method.

| | | Predicted class | |
|---|---|---|---|
| | | Positive | Negative |
| Actual class | Positive | 1120 | 135 |
| | Negative | 2366 | 35298 |

We used a judgment category table[1] officially released by the Japan Society of Ningen Dock as the baseline method. This is because notifications to the examinees of regular health checkup are usually made based on similar standards in Japan. A threshold value for each item was set and discrimination was carried out by the logical OR operation on each item. We only used the items that were common to the input features of this study. The total number of items used was

[1] https://www.ningen-dock.jp/wp/wp-content/uploads/2013/09/Dock-Hantei2018-20181214.pdf.