

A Rough Set Approach to Multiple Dataset Analysis

Ken Kaneiwa

*Department of Electrical Engineering and Computer Science, Iwate University
4-3-5 Ueda, Morioka, Iwate 020-8551, Japan
kaneiwa@cis.iwate-u.ac.jp*

Abstract

In the area of data mining, the discovery of valuable changes and connections (e.g., causality) from multiple data sets has been recognized as an important issue. This issue essentially differs from finding statistical associations in a single data set because it is complicated by the different data behaviors and relationships across multiple data sets. Using rough set theory, this paper proposes a change and connection mining algorithm for discovering a time delay between the quantitative changes in the data of two temporal information systems and for generating the association rules of changes from their connected decision table. We establish evaluation criteria for the connectedness of two temporal information systems with varying time delays by calculating weight-based accuracy and coverage of the association rules of changes, adjusted by a fuzzy membership function.

Keywords: rough set theory, fuzzy theory, multiple datasets, causality

1. Introduction

Data mining algorithms [2, 3] enable us to find useful patterns and rules generated from the frequent itemsets in databases. Alternatively, numerous techniques for extracting decision rules have been proposed in the field of rough set theory [4, 5, 6, 7]. In the environment of computer networks, there is a further requirement to integrate and analyze distributed data for discovering valuable patterns and rules across distributed multiple databases. Ex-

¹This paper is an extended version of [1].

isting data mining algorithms [8, 9, 10] effectively and statistically integrate multiple data sets, but do not deal with data behaviors and relationships across multiple data sets.

To enable integration and discovery, we have to analyze and integrate various data in a highly sophisticated manner due to the different contexts of multiple databases. In particular, time and space stamps should be used for examining changes and connections among different data sets. In this way, we can find a time delay between the time-stamped data of two distributed databases. The *quantitative* changes in the temporal data of such time stamped databases can be interpreted as events; therefore, a time delayed connection implies that one quantitative change causes another. From this perspective, a candidate causality is analyzed and obtained by connecting the quantitative changes in the distributed data.

The causality [11, 12] derived from changes and connections is valuable knowledge obtained from the integration of multiple databases. Such causal knowledge has an advantage in that each of the conditions implies a different effect. Yet there are few approaches to mining the changes and connections in data across multiple data sets. As discussed [13, 14], most knowledge-discovery algorithms capture only statistical associations that are substantially different from causality.

In this paper, we propose a change and connection mining algorithm based on the notions of attribute reduction and minimal rule generation in rough set theory. We assume that attribute data in information systems are associated with time stamps. For the purpose of causality mining, we formalize the quantitative changes estimated using different operators. These results are used to derive the association rules of changes by slidingly connecting two *temporal* information systems for varying time delays. In order to measure the changes (including *fast-changing events*), we propose *weight-based* accuracy and coverage of the association rules with respect to the indiscernibility of changes. These evaluations identify the connectedness of the two temporal information systems for each *time delay*.

Our algorithm for mining changes and connections has the following interesting features.

- **Minimal rules:** The minimal rule generation in rough set theory enables us to obtain minimal association rules using attribute reduction. Minimal causes provide better explanations of effects, as employed in abductive reasoning [15].

- **Consistent rules:** The discernibility of decision classes in rough set theory is used to derive consistent rules; i.e., logically, inconsistent rules $\varphi \rightarrow \alpha$ and $\varphi \rightarrow \neg\alpha$ are excluded. As a result, the consistency is suitable for establishing the connectedness of distributed data due to the small conflict among the consistent rules.
- **Connections between quantitative changes:** In order to uncover a strong connection between information systems, we evaluate the association rules of changes that are generated from the quantitative changes in estimating the values of data.
- **Time delays:** Mining of changes and connections is realized by considering differing time delays between the data of two temporal information systems. This is temporally consistent because of adherence to the assumption in the causal theory [16]; i.e., “if A causes B , then A occurs earlier than B .”

These features follow the three classical conditions under which A causes B (as defined in [13]): (i) statistical associations between the values of A and B , (ii) direction of causality, and (iii) no common causes of A and B .

This paper is arranged as follows. In Section 2, we describe a connection method of two temporal information systems for various time delays. As part of the method, we propose the notions of quantitative changes in attribute values and the association rules of changes. Then we establish the connectedness of two temporal information systems for each time delay evaluated by weight-based accuracy and coverage. In Section 3, we present our algorithm for mining changes and connections from two temporal information systems. The experimental results are reported in Section 4. Finally, we discuss the related work to our approach in Section 5 and conclude the paper in Section 6. In addition, we briefly describe the basic notions of rough sets in Appendix A.

2. Changes and Connections in Data

In this section, we present a connection method for two temporal information systems with various time delays. We establish the quantitative changes in the attribute values of temporal information systems and formulate the association rules of changes in a connected decision table. For each time delay, the connectedness of two temporal information systems is evaluated by our proposed weight-based accuracy and coverage.

2.1. Connecting Temporal Information Systems

A considerable number of databases are distributed across multiple sites and domains in computer networks. This is because each database is developed for a different purpose by a different person and/or enterprise. If we integrate and analyze these distributed databases and their relationships, we can obtain very useful information for the discovery of valuable patterns and rules from various combinations of databases. However, data mining across differently constructed datasets is not an easy task because we do not know whether a relationship exists among the datasets.

As an approach to the task, we focus on finding useful patterns and rules by analyzing time series data in many different datasets. Time series data contain time information, which we can use for identifying a temporal relationship between databases. Consider the climate data and medical data on Tokyo in Figure 1. From these data, experts can conclude that the number of influenza patients increases when the temperature decreases and the humidity is below 45 percent. They can also find that changes in the medical data are observed two to three days after changes in the climate data.

To enable the data analysis above, we first connect two temporal databases by sliding their time delays. Then we find a causal connection of the two temporal databases by evaluating the consistency of each time-delayed connection. Roughly speaking, we attempt to discover a cause-effect relationship between each pair of multiple databases.

2.2. Time Delays between Two Temporal Information Systems

In rough set theory, each database is represented by an information system $T = (U, A)$ where U is a non-empty finite set of objects and A is a non-empty finite set of attributes (see the basic concept of rough set theory in Appendix A). In order to represent time series data (each of climate data, medical data, etc.) in an information system, we originally define a specific information system with time stamps.

A *temporal* information system is an information system $T = (U_{time}, A)$ wherein the objects x of U_{time} denote time stamps (e.g., dates and weeks) and for each attribute $a \in A$, $a(x)$ maps the value of a at time stamp x . Each temporal information system $T = (U_{time}, A)$ is normalized such that all the elements of U_{time} are replaced by natural numbers. Let $Nat_{i,j}$ denote a finite set of natural numbers such that $\{x \in Nat \mid i \leq x \leq j\}$. Then, we define the normalization of a temporal information system $T = (U_{time}, A)$ by $N(T) = (Nat_{i,j}, A')$ with a bijection $n: U_{time} \rightarrow Nat_{i,j}$ such

that for every $x, y \in U_{time}$, $x < y \Leftrightarrow n(x) < n(y)$, $|U_{time}| = |Nat_{i,j}|$, and $A' = \{a' \mid a \in A \ \& \ \forall x \in U_{time}. a'(n(x)) = a(x)\}$. For example, $T_1 = (\{2001-1-1, 2001-1-2, 2001-1-3\}, A_1)$ is normalized by the natural number set $Nat_{1,3} = \{1, 2, 3\}$; i.e., $N(T_1) = (Nat_{1,3}, A'_1)$ where $n(2001-1-1) = 1$, $n(2001-1-2) = 2$, and $n(2001-1-3) = 3$.

Lemma 1. *Every temporal information system is translated into a normalized temporal information system.*

In the remainder of this paper, we will focus on normalized temporal information systems because of this lemma.

We define a connection of two normalized temporal information systems with a time delay by matching and arranging time stamps.

Definition 1 (Connections with a Time Delay). *Let $T_1 = (Nat_{i,j}, A_1)$ and $T_2 = (Nat_{h,m}, A_2)$ be two normalized temporal information systems such that T_1 and T_2 have no common attributes ($A_1 \cap A_2 = \emptyset$). The connection $con(T_1, T_2, \Delta)$ of T_1 and T_2 with a time delay $\Delta \in Nat$ is defined as an information system $T = (U_{time}, A)$ such that*

- $U_{time} = Nat_{i,j} \cap Nat_{h',m-\Delta}$ and
- $A = \{a \upharpoonright U_{time} \mid a \in A_1\} \cup \{b' \mid b \in A_2 \ \& \ \forall x \in U_{time}. b'(x) = b(x - \Delta)\}$

where $a \upharpoonright U_{time}$ denotes the attribute a restricted to domain U_{time} , and $h' = i$ if $h - \Delta < i$, otherwise, $h' = h - \Delta$.

In Figure 2, the tables on the left-hand side present two normalized temporal information systems $T_1 = (Nat_{1,7}, A_1)$ and $T_2 = (Nat_{2,6}, A_2)$, and the table on the right-hand side shows the connection $con(T_1, T_2, 1)$ of T_1 and T_2 . The time stamps of attributes b_1 and b_2 in T_2 are decreased by the time delay $\Delta = 1$. For example, time stamp 3 of values 2 and 1 of b_1 and b_2 in T_2 is changed into time stamp 2 in $con(T_1, T_2, 1)$. As a result of the connection, the lowest two rows in T_1 are deleted to adjust for the size of T_2 . This connection method can also be applied to more than two information systems. For example, the connection $con(con(T_1, T_2, \Delta_1), T_3, \Delta_2)$ is obtained when we integrate three information systems T_1 , T_2 , and T_3 with time delays Δ_1 and Δ_2 .

Let $T_1 = (U_{time}, A_1)$ and $T_2 = (U'_{time}, A_2)$ be two temporal information systems. Their connection builds a decision table $T = con(T_1, T_2, \Delta)$ in rough set theory if A_2 is a singleton as a decision attribute. If A_2 is not a

singleton, then T_2 is reduced to an information system $T_2[b] = (U'_{time}, \{b\})$ for an attribute $b \in A_2$. Alternatively, we can translate from A_2 into a singleton without losing the information. Let $A_2 = \{b_1, \dots, b_n\}$. Then the translation is defined such that the attributes b_1, \dots, b_n are expressed by an n -tuple; i.e., $T_2^f = (U'_{time}, \{f_{(b_1, \dots, b_n)}\})$ where for every $x \in U'_{time}$, $f_{(b_1, \dots, b_n)}(x) = (b_1(x), \dots, b_n(x))$. Therefore, we obtain the two decision tables $con(T_1, T_2[b], \Delta)$ and $con(T_1, T_2^f, \Delta)$. If T_1 and T_2 are regarded as cause-and-effect information systems, then the connection of T_1 and T_2 with a time delay leads to a cause-effect decision table. However, if the data behavior cannot be interpreted in terms of a decision table, a cause-effect decision table cannot be generated. In addition to the varying of time delays, we have to extract data behaviors by estimating the changes in the data values of both cause-and-effect information systems.

2.3. Quantitative Changes

In order to classify data behaviors such as increase and decrease, we define the *quantitative changes* in the data values of temporal information systems. Given two temporal information systems T_1 and T_2 , their quantitative changes may represent a connection between the candidate data of causes and effects. Through the diversity of time delays, numerous connections of the quantitative changes will be analyzed to determine whether changing the values of attributes in an information system T_1 affects the values of attributes in another information system T_2 .

Definition 2 (Quantitative Estimation Operators).

Based on [17], several quantitative estimation operators for the numeric values $a(x)$ of attributes are defined by the following.

$$\begin{aligned}
 (\text{difference}) \quad \pi_d(a(x)) &= a(x) - a(x - 1) \\
 (\text{variation rate}) \quad \pi_v(a(x)) &= \frac{a(x) - a(x - 1)}{a(x)} \\
 (\text{threshold}) \quad \pi_{\geq k}(a(x)) &= \begin{cases} 1 & \text{if } a(x) \geq k \\ 0 & \text{otherwise} \end{cases} \\
 (\text{variation rate of difference}) \quad \pi_{2v}(a(x)) &= \pi_v(\pi_d(a(x))) \\
 (\text{trend}) \quad \pi_{tr(k)}(a(x)) &= \frac{a(x-k) + \dots + a(x) + \dots + a(x+k)}{2k + 1}
 \end{aligned}$$

These operators are (sometimes compositionally) used to estimate the quantitative changes in temporal information systems T as follows.

Definition 3 (Quantitative Changes). *Let $T = (Nat_{i,j}, A)$ be a normalized temporal information system. The quantitative change of T obtained by a quantitative estimation operator $\pi \in \{\pi_d, \pi_v, \pi_{\geq k}, \pi_{\leq k}, \pi_{2v}, \pi_{tr(k)}\}$ with $h, m \in Nat$ is an information system $\pi(T) = (U_{time}, A')$ such that*

- $U_{time} = Nat_{i+h, j-m}$ with $i + h \leq j - m$ and
- $A' = \{a' \mid a \in A \ \& \ \forall x \in U_{time}. a'(x) = \pi(a(x))\}$

where $h = 1$ and $m = 0$ if $\pi = \pi_d$ or π_v , $h = m = 0$ if $\pi = \pi_{\geq k}$ or $\pi_{\leq k}$, $h = 2$ and $m = 0$ if $\pi = \pi_{2v}$, and $h = k$ and $m = k$ if $\pi = \pi_{tr(k)}$.

The natural numbers h and m indicate that $Nat_{i,j}$ is reduced to $Nat_{i+h, j-m}$ because the number of values obtained by some operators decreases. For example, Figure 3 shows the differences $\pi_d(T_1) = (Nat_{2,7}, A'_1)$ and the results of the high threshold $\pi_{\geq 2.5}(T_2) = (Nat_{2,6}, A'_2)$ estimated from $T_1 = (Nat_{1,7}, A_1)$ and $T_2 = (Nat_{2,6}, A_2)$, respectively.

In the process of data analysis, we are not certain about the temporal relationship between two information systems; therefore, they should be combined exhaustively within the scope of time delays. By increasing the time delay Δ from 0, the quantitative changes $\pi_1(T_1)$ and $\pi_2(T_2[b])$ of two temporal information systems T_1 and $T_2[b]$ are slidingly connected to achieve numerous cause-effect decision tables as follows.

$$\begin{aligned} & con(\pi_1(T_1), \pi_2(T_2[b]), 0), con(\pi_1(T_1), \pi_2(T_2[b]), 1), \\ & con(\pi_1(T_1), \pi_2(T_2[b]), 2), \dots, con(\pi_1(T_1), \pi_2(T_2[b]), m) \end{aligned}$$

Let $T = (U_{time}, A)$ with $A = \{a_1, \dots, a_n\}$ and let quantitative estimation operators π_1, \dots, π_n be applied to each attribute a_j in T . Then, the decomposed and estimated information systems $\pi_1(T[a_1]), \dots, \pi_n(T[a_n])$ are reconnected by

$$\begin{aligned} & con(\pi_1(T[a_1]), con(\pi_2(T[a_2]), \dots con(\pi_{n-1}(T[a_{n-1}]), \\ & \pi_n(T[a_n]), 0) \dots , 0), 0) \end{aligned}$$

which is simply denoted by $\pi_1(T[a_1]) \circ \dots \circ \pi_n(T[a_n])$. For example, let π_d be a difference operator, π_v be a variation rate operator, and $\pi_{\geq 2.5}$ be

a high threshold operator. Figure 4 shows that the quantitative changes $\pi_d(T_1[a_1]) \circ \pi_v(T_1[a_2])$ and $\pi_{\geq 2.5}(T_2[b_1])$ of T_1 and $T_2[b_1]$ in Figure 2 and Figure 3 are transformed into the connected decision table $\text{con}(\pi_d(T_1[a_1]) \circ \pi_v(T_1[a_2]), \pi_{\geq 2.5}(T_2[b_1]), 2)$ with time delay $\Delta = 2$ and the decision attribute $b_1 \in A_2$.

2.4. Association Rules of Changes

We define association rules with respect to quantitative changes, which are employed to evaluate the connectedness of two temporal information systems. Let T_1 and $T_2[b]$ be two temporal information systems and π_1 and π_2 be quantitative estimation operators. A decision rule in rough set theory is called an *association rule of changes* if it is generated from a cause-effect decision table $\text{con}(\pi_1(T_1), \pi_2(T_2[b]), \Delta)$ consisting of the quantitative changes $\pi_1(T_1)$ and $\pi_2(T_2[b])$. It should be noted that the association rules of changes $(a_1 = v_1) \wedge \dots \wedge (a_n = v_n) \rightarrow (b = v)$ adhere to the natures of causality as follows.

- (i) The cause-effect decision table is built by the quantitative changes in the data of two temporal information systems T_1 and $T_2[b]$ that correspond to the data behaviors of causes and effects.
- (ii) A time delay exists such that the time stamps of the condition attributes $a_1 = v_1, \dots, a_n = v_n$ in $\pi_1(T_1)$ occur earlier than those of the decision attribute $b = v$ in $\pi_2(T_2[b])$.

For example, the following association rules of changes are generated from the connected decision table $\text{con}(\pi_d(T_1[a_1]) \circ \pi_v(T_1[a_2]), \pi_{\geq 2.5}(T_2[b_1]), 2)$ in Figure 4.

$$\begin{aligned} &(\text{difference}) \wedge (\text{variation}) \rightarrow (\text{threshold}) \wedge (\text{time delay}) \\ &(a_1 = 0) \wedge (a_2 = 0.0) \rightarrow (b_1 = 0) \wedge (\Delta = 2) \\ &(a_1 = +11) \wedge (a_2 = +1/3) \rightarrow (b_1 = 1) \wedge (\Delta = 2) \end{aligned}$$

In these rules, the condition attributes are difference π_d and variation rate π_v and the decision attribute is threshold $\pi_{\geq 2.5}$ with time delay $\Delta = 2$. The first rule implies that if the values of attributes a_1 and a_2 are neither decreased nor increased, then the value of attribute b_1 does not exceed threshold 2.5 in the next two time slots. The second rule means that if the value of attribute a_1 is increased by +11 and the variation rate of the value of attribute a_2 is +1/3, then the value of attribute b_1 exceeds threshold 2.5 in the next two time slots.

2.5. Indiscernibility and Weight

To interpret the association rules of changes, we would like to distinguish between the indiscernibility and weight of quantitative changes by improving the evaluation of decision rules in rough set theory. In this study, indiscernibility captures the increase and decrease of values, and weight measures the quantity of data behaviors.

Let $T = (U_{time}, A \cup \{d\})$ be a decision table and B be a relative reduct of T . The B -indiscernibility relation of quantitative changes is defined by an equivalence relation I_B^{qc} on U_{time} such that

$$I_B^{qc} = \{(x, y) \in U_{time}^2 \mid \forall a \in B. sign(a(x)) = sign(a(y))\}.$$

The sign function $sign(n)$ is defined by $sign(n) = 1$ if $n > 0$, $sign(n) = -1$ if $n < 0$, and $sign(n) = 0$ if $n = 0$ (as in [18]).

By the B -indiscernibility I_B^{qc} of quantitative changes, the semantics of the formula $(a_1 = v_1) \wedge \dots \wedge (a_n = v_n)$ in T is refined by $\llbracket (a_1 = v_1) \wedge \dots \wedge (a_n = v_n) \rrbracket_T^{qc} = \{x \in U_{time} \mid sign(a_1(x)) = sign(v_1), \dots, sign(a_n(x)) = sign(v_n)\}$. Let $\{s_1, \dots, s_u\}$ denote $sign(V_d) = \{sign(d_j) \mid d_j \in V_d\}$. For each value s_i of $sign(V_d)$ of the decision attribute d , we define a decision class on quantitative changes $U_i = \{x \in U \mid sign(d(x)) = s_i\}$ where $U = U_1 \cup \dots \cup U_{|sign(V_d)|}$ (i.e., $u = |sign(V_d)|$) and for every $x, y \in U_i$, $sign(d(x)) = sign(d(y))$.

We consider scanning not only quantitative changes but also those events that drastically change attribute values (which we refer to as *fast changing events*). As a measuring method, the weight $w(x)$ of changes for each time stamp $x \in U_{time}$ is defined by the following.

$$w(x) = \sum_{a \in A \cup \{d\}} ||a(x)||$$

where $|| \cdot ||: R \rightarrow R$ is the absolute value function such that $||n|| = n$ if $n \geq 0$, otherwise $||n|| = -n$. For example, the connected decision table in Figure 4 contains fast-changing events because $a_1(4) = +11$ is an intensively higher value than the other values. Therefore, we obtain high weight $w(4) = 12.33 \dots$ for time stamp 4 but low weight $w(1) = 0$ for time stamp 1. Weights $w(1)$ and $w(4)$ can be used to respectively evaluate the two association rules $(a_1 = 0) \wedge (a_2 = 0.0) \rightarrow (b_1 = 0)$ and $(a_1 = +11) \wedge (a_2 = +1/3) \rightarrow (b_1 = 1)$ (shown in Section 2.4). The second association rule importantly includes a fast-changing event because of high weight $w(4)$.

Definition 4. (Weight-Based Accuracy and Coverage on I_B^{qc}) Let U_i be a decision class on quantitative changes and let $x \in U_i$. The weight-based accuracy $w_accuracy$ and weight-based coverage $w_coverage$ of a decision rule r of the form $\varphi \rightarrow (d = d(x))$ with $d(x) \in V_d$ are defined as follows.

$$w_accuracy(T, r, U_i) = \frac{w(U_i \cap \llbracket \varphi \rrbracket_T^{qc})}{w(\llbracket \varphi \rrbracket_T^{qc})}$$

$$w_coverage(T, r, U_i) = \frac{w(U_i \cap \llbracket \varphi \rrbracket_T^{qc})}{w(U_i)}$$

where $w(X) = \sum_{x \in X} w(x)$.

The accuracy and coverage unify similar data behaviors on indiscernibility and measure the quantity of data behaviors on the weight (by extending the standard accuracy and coverage in Appendix A).

2.6. Connectedness

The *connectedness* of two temporal information systems is significantly evaluated by weight-based accuracy and coverage. First, we define a consistency evaluation that finds the most consistent association rule for each decision class. Second, we estimate the total value of the maximum consistency evaluations for all decision classes.

The consistency evaluation $eval(T, r, U_i)$ is defined by using the weight-based coverage and accuracy of a decision rule r with a fuzzy membership function μ_S as follows.

$$eval(T, r, U_i) = w_coverage(T, r, U_i) \times \mu_S(w_accuracy(T, r, U_i))$$

The fuzzy membership function $\mu_S: [0, 1] \rightarrow [0, 1]$ (similar to the S fuzzy set [19]) is defined by the following.

$$\mu_S(v) = \begin{cases} 0 & \text{if } 0 \leq v \leq \frac{1}{2} \\ 8(v - \frac{1}{2})^2 & \text{if } \frac{1}{2} < v \leq \frac{3}{4} \\ 1 - 8(v - 1)^2 & \text{if } \frac{3}{4} < v \leq 1 \end{cases}$$

By applying the membership function to each accuracy value, we determine whether each association rule is suitable for connecting two temporal information systems. In other words, association rules with low accuracy values

should be eliminated even if their coverage values are high. This is because such rules are *inconsistent* with some other rules; though we intend to find *consistent* rules beyond statistical associations. We use the fuzzy membership function to emphasize the accuracy values of (in)consistent rules.

For example, consider the following association rules r_1 and r_2 in a connected decision table $T = (U, A \cup \{d\})$.

$$\begin{aligned} r_1: (a_1 = +1) \wedge (a_2 = -1) &\rightarrow (d = +1) \\ r_2: (a_1 = +1) \wedge (a_2 = -1) &\rightarrow (d = -1) \end{aligned}$$

These association rules have identical conditions; though their decisions contradict each other. Let $U_1 = \{x \in U \mid d(x) = +1\}$ and $U_2 = \{x \in U \mid d(x) = -1\}$. If we have $w_accuracy(T, r_1, U_1) = 0.45$ and $w_accuracy(T, r_2, U_2) = 0.55$, then these rules are inconsistent with each other. Therefore, the membership function results in $\mu_S(0.45) = 0$ and $\mu_S(0.55) = 0.02$, and the consistency evaluations $eval(T, r_1, U_1) = coverage(T, r_1, U_1) \times 0.02$ and $eval(T, r_2, U_2) = coverage(T, r_2, U_2) \times 0$ return very low values. In contrast, if we have the other values $w_accuracy(T, r_1, U_1) = 0.89$ and $w_accuracy(T, r_2, U_2) = 0.1$, then the first rule should be highly evaluated due to the low conflict between r_1 and r_2 . In this case, the membership function yields $\mu_S(0.89) = 0.9032$ and $\mu_S(0.1) = 0$. Thus, the accuracy value of the first association rule will affect the consistency evaluation by calculating $eval(T, r_1, U_1) = coverage(T, r_1, U_1) \times 0.9032$.

Let \mathcal{R} be the family of subsets B of A in a decision table $T = (U_{time}, A \cup \{d\})$ such that B is a relative reduct of T . For each reduct $B \in \mathcal{R}$, we can obtain the set R_B of minimal association rules for every decision class U_i ; i.e., each reduct B provides a minimal set of condition attributes in A . The maximum consistency evaluation in the set R_B is defined by $max_eval(T, R_B, U_i) = eval(T, r, U_i)$ if $r \in R_B$ and for every rule $r' \in R_B$, $eval(T, r, U_i) \geq eval(T, r', U_i)$.

Definition 5 (Connectedness on Association Rules of Changes).

The *connectedness of condition and decision attributes for each relative reduct B of T* is defined by

$$connectedness(T, B) = \sum_{i=1, \dots, |sign(V_d)|} max_eval(T, R_B, U_i)$$

The underlying assumption behind the maximum evaluation $max_eval(T, R_B, U_i)$ is that the connectedness is strengthened by the existence of one-sided

association rules. That is, it is required that a rule has a high consistency evaluation in U_i rather than the total value of consistency evaluations for all the association rules in U_i . From that assumption, a maximum value of one is selected from the association rules of different conditions implying the same decision. For example, consider the above rule r_1 and the following rule r_3 .

$$r_3: (a_1 = -1) \wedge (a_2 = +1) \rightarrow (d = +1)$$

A higher value of one is selected from the consistency evaluations $eval(T, r_1, U_1)$ and $eval(T, r_3, U_1)$ rather than their sum. This is because the two rules have opposite conditions $(a_1 = +1) \wedge (a_2 = -1)$ and $(a_1 = -1) \wedge (a_2 = +1)$ deriving the same decision $(d = +1)$. The sum of their consistency evaluations aggregates different behaviors and therefore does not indicate the connectedness of condition and decision attributes.

We calculate the connectedness for each connected decision table such that the association rules of changes are generated and evaluated by sliding the connection of two temporal information systems. From various time delays, a maximum connectedness is discovered as a strong connection of the systems.

3. Change and Connection Mining Algorithm

In this section, we describe a change and connection mining algorithm for two temporal information systems T_1 and $T_2[b]$ and a maximum time delay $m(\in Nat)$, as shown in Figure 5. For the input data, this algorithm returns a time delay as a result of generating and evaluating association rules of changes in the connections of T_1 and $T_2[b]$ for various time delays (from 0 to m). The algorithm is outlined as follows.

1. **Change and connection mining:** For each time delay Δ from 0 to m , two temporal information systems T_1 and $T_2[b]$ are connected in order to analyze changes and connections in their connected decision table by the following sub-processes.
 - (a) **Quantitative change calculation:** Two temporal information systems T_1 and $T_2[b]$ are decomposed into $T_1[a_1], \dots, T_1[a_n]$ and $T_2[b]$ for each attribute data. Then the quantitative changes $\pi_1(T_1[a_1]), \dots, \pi_n(T_1[a_n])$ and $\pi(T_2[b])$ of the data in the decomposed information systems are calculated.

- (b) **Connection of two information systems:** After the quantitative change calculation, the two information systems $T'_1 = \pi_1(T_1[a_1]) \circ \dots \circ \pi_n(T_1[a_n])$ and $T'_2 = \pi(T_2[b])$ that correspond to the data behaviors of causes and effects are connected into $T = con(T'_1, T'_2, \Delta)$ for each time delay Δ .
 - (c) **Rule generation:** By using the rough set rule generation method, association rules of changes are generated from the connected cause-effect decision table $T = con(T'_1, T'_2, \Delta)$.
2. **Connectedness evaluation:** The consistency evaluation $eval(T, r, U_i)$ is calculated by using the weight-based accuracy and coverage of the association rules of changes. Then the connectedness of the two temporal information systems T_1 and $T_2[b]$ is evaluated by the sum of maximum consistency evaluations. Finally, the value of the connectedness decides which time delay implies the strongest connection of the two information systems.

3.1. Change and Connection Mining

The process of change and connection mining is carried out as follows.

In Line 2 of Figure 5, in order to analyze hidden behaviors in $T_1 = (U_{time}, \{a_1, \dots, a_n\})$ and $T_2[b] = (U'_{time}, \{b\})$, quantitative estimation operators π_1, \dots, π_n and π are applied to the decomposed information systems $T_1[a_1], \dots, T_1[a_n]$ and $T_2[b]$, and the estimated results $\pi_1(T_1[a_1]) \circ \dots \circ \pi_n(T_1[a_n])$ and $\pi(T_2[b])$ are stored in the variables T'_1 and T'_2 .

Since we are not certain which time delay constructs a temporal relation suitable for T'_1 and T'_2 , in Lines 3 - 19, this algorithm functions to connect them in varying time delays Δ from 0 to m . In the loop of the time delays, the connected decision table $T = (U''_{time}, \{a_1, \dots, a_n\} \cup \{b\})$ of T'_1 and T'_2 is computed by $T = con(T'_1, T'_2, \Delta)$ (in Line 4). Therefore, the connected decision table T is one of the candidate cause-effect decision tables to generate association rules of changes.

In Line 11, using rough set theory, minimal association rules $r = rule(x, B, T)$ are created for each relative reduct B in $\mathcal{R} = reducts(T)$ [20], i.e., B is a subset of the set $\{a_1, \dots, a_n\}$ in T that supplies a minimal set of condition attributes of rules. Computing the set $\mathcal{R} = reducts(T)$ of relative reducts of $T = con(T'_1, T'_2, \Delta)$ is based on the B -indiscernibility relation of quantitative changes by extending the standard reduct set computation in [21]. In other words, the reducts are computed by the standard reduct set algorithm that is extended by the B -indiscernibility relation of quantitative changes.

Let U_i be a decision class on quantitative changes where $U''_{time} = U_1 \cup \dots \cup U_{|sign(V_b)|}$ in the connected decision table T . Each decision class on quantitative changes U_i corresponds to one of the values of $sign(V_b)$ of decision attribute b . Then, in Lines 8 - 16, the association rules $r = rule(x, B, T)$ for all time stamps x in U_i are generated for every $i = 1, \dots, |sign(V_b)|$. The function $rule(x, B, T)$ simply determines a decision rule for each item x (in U) in information system T using each reduct of B .

3.2. Connectedness Evaluation

In the process of connectedness evaluation, in Line 12 the consistency evaluation $eval(T, r, U_i)$ is calculated by using the weight-based coverage of each association rule r in decision classes U_i . In Lines 12 - 13, for each i from 1 to $|sign(V_b)|$, we obtain the maximum consistency evaluation (stored in the variable max_eval_i) by the evaluation $eval(T, r, U_i)$. Finally, in Line 15, the connectedness of T'_1 and T'_2 for each relative reduct B (stored in variable C_B) is calculated. For every current time delay Δ , we select the maximum connectedness denoted by $connectedness_\Delta$ from all the relative reducts B in \mathcal{R} (in Line 18). After the loop of the time delays (Lines 3 - 19), a time delay t with the maximum connectedness is returned by comparing the connectedness for each time delay in $0 \leq j \leq m$.

4. Experimental Results

We implemented the change and connection mining algorithm in Java. In order to evaluate the proposed mining algorithm on real-world data, two time-series data sets were built as shown in Figure 1, by downloading climate data and medical data on Tokyo from web sites [22, 23]. On the left-hand side of Figure 1, the climate data set contains the daily observed data on minimum temperature and humidity in Tokyo from August 7, 2006 to July 15, 2007. On the right-hand side of Figure 1, the medical data set consists of weekly reported numbers of influenza victims per hospital in Tokyo from September 4, 2006 to June 17, 2007. These data sets can be represented by the two normalized temporal information systems $T_1 = (N_{1,315}, \{temperature, humidity\})$ and $T_2 = (N_{5,45}, \{influenza\})$. The sizes of data sets and attributes are obtained from information systems, e.g., $T_1 = (N_{1,315}, \{temperature, humidity\})$ implies (daily) 315 items and two attributes. We simply denote the attributes *temperature*, *humidity*, and *influenza* by a_t , a_h , and b_f , respectively.

Let π_v , $\pi_{tr(10)}$, $\pi_{\leq 35}$, and π_d be quantitative estimation operators. Figures 6 - 8 demonstrate the results of applying the quantitative estimation operators to climate data and medical data on Tokyo.

For the climate data, first, to exclude noisy data the long-term behaviors of daily observed temperatures are estimated by the trends $\pi_{tr(10)}(T_1[a_t])$ of the temperatures in $T_1[a_t]$ on the left-hand side in Figure 6. Then, the increase-decrease rates of the trends are calculated by the variation rates $\pi_v(\pi_{tr(10)}(T_1[a_t]))$ from the data of the trends $\pi_{tr(10)}(T_1[a_t])$ on the right-hand side. Second, we set the threshold such that the changes $\pi_{\leq 35}(T_1[a_h])$ of humidity in $T_1[a_h]$ are denoted by -1 if their values decrease to equal to or less than 35% at the time stamp. Figure 7 presents the daily observed humidity levels in the temporal information system $T_1[a_h]$ on the left-hand side. Then, the quantitative changes $\pi_{\leq 35}(T_1[a_h])$ are estimated from the data of $T_1[a_h]$ on the right-hand side.

For the medical data, the increase or decrease $\pi_d(T_2[b_f])$ of the numbers of influenza victims in $T_2[b_f]$ are regarded as candidate effects of the quantitative changes $\pi_v(\pi_{tr(10)}(T_1[a_t])) \circ \pi_{\leq 35}(T_1[a_h])$. Figure 8 shows the number of influenza victims for each week in the temporal information system $T_1[b_f]$ on the left-hand side and the increase or decrease $\pi_d(T_2[b_f])$ estimated from the data of $T_1[b_f]$ on the right-hand side. Unlike estimating the variation rates of temperatures, the numbers of victims are absolute values; therefore, the increase or decrease should be calculated (since absolute values contain 0).

After the estimation, we turn to connecting the climate data and medical data by using the quantitative changes $\pi_v(\pi_{tr(10)}(T_1[a_t])) \circ \pi_{\leq 35}(T_1[a_h])$ and $exp_7(\pi_d(T_2[b_f]))$ of T_1 and $T_2[b_f]$ with an expansion function exp_7 . Since $T_2[b_f]$ consists of weekly data, we expand it to daily data denoting the differences of values per week. Let $T = (Nat_{i,j}, A)$ and $k > 1$. Then the expansion function is defined by $exp_k(T) = (U_{time}, A')$ such that

- $U_{time} = Nat_{(i-1) \cdot k + 1, j \cdot k}$ and
- $A' = \{a' \mid a \in A \ \& \ \forall x \in U_{time}. a'(x) = a(quotient(x + k - 1, k))\}$.

Figure 9 represents the outcomes of applying the algorithm to the two temporal information systems T_1 and $T_2[b_f]$. Let T be the connected decision table $con(\pi_v(\pi_{tr(10)}(T_1[a_t])) \circ \pi_{\leq 35}(T_1[a_h]), exp_7(\pi_d(T_2[b_f])), \Delta)$. The left-hand side of Figure 9 shows the consistency evaluations $eval(T, r, U_i)$ for the time delays $\Delta = 0, \dots, 13$ where r is an association rule $(a_t = a_t(x)) \wedge (a_h = a_h(x)) \rightarrow (b_f = b_f(x))$ for a time stamp $x \in U_i$ where

$a_t(x) \in V_{a_t}$, $a_h(x) \in V_{a_h}$, and $b_f(x) \in V_{b_f}$. In order to overview the association rules with respect to the B -indiscernibility I_B^{qc} , each association rule is generalized by $(a_t = \text{sign}(a_t(x))) \wedge (a_h = \text{sign}(a_h(x))) \rightarrow (b_f = \text{sign}(b_f(x)))$ for a time stamp $x \in U_i$ where $\text{sign}(a_t(x)) \in \{-1, 0, +1\}$, $\text{sign}(a_h(x)) \in \{-1, 0\}$, and $\text{sign}(b_f(x)) \in \{-1, 0, +1\}$. In this figure above, the *generalized* association rules are listed by indexing the plot styles for their consistency evaluations. In particular, it can be seen that the generalized association rule $(a_t = -1) \wedge (a_h = -1) \rightarrow (b_f = +1)$, compared with the other rules, has a high consistency evaluation through all the time delays. Note that the value of each generalized association rule in the figure in total indicates the evaluation obtained by measuring the weights of association rules with similar data behaviors. The high evaluation of the generalized rule reports that if the temperature values decrease, and humidity is equal to or under the percent limit, then the number of influenza victims increases. Without the consistency evaluation and various time delays, it would be difficult to select a time-delayed decision table that consistently connects the different contexts of distributed data.

On the right-hand side of Figure 9, our experimental result indicates the connectedness evaluated for each of the time delays $\Delta = 0, \dots, 13$. As indicated by the result in the figure, the algorithm returns the time delay $\Delta = 2$ (denoting two days) that has the maximum connectedness 0.2698 of T_1 and $T_2[b_f]$. High connectedness values 0.2685 and 0.2617 are also given for the time delays $\Delta = 3$ and 1 (denoting three days and one day, respectively). This result confirms the existence of certain time delays that strongly connect climate data and medical data on Tokyo.

5. Related Work

Many data-mining algorithms for multiple databases have been proposed in the area of data mining and knowledge discovery. Cheung et al [8] developed a distributed algorithm for Distributed Mining of Association (DMA) rules.. This algorithm is efficient because the number of candidate itemsets is reduced by locally pruning them into multiple databases. Jin and Agrawal [9] established new operators for querying frequent patterns across multiple databases and presented an algorithm for finding an optimized query plan across multiple databases. These approaches can find frequent patterns in multiple databases, but they cannot analyze quantitative changes and temporal relationships in data across multiple databases. Zhu and Wu [10]

developed an algorithm for discovering relational patterns where multiple databases are jointed together to construct a hybrid frequent pattern tree. The relational patterns are constraints described by relationships and operators on support values in multiple databases. Unlike their algorithm, our algorithm for mining changes and connections with some time delays is closely related to the mining of causality in time-series data, essentially differing from the frequent relational patterns.

As described in [24], several causal discovery systems have been developed, such as CaMML, TETRAD, and TimeSleuth. Their algorithms and systems infer causal relationships based on Bayesian learning ideas. Silverstein et al.[14] proposed a constraint-based algorithm for discovering causality. They distinguished causality mining from traditional association rule mining as simply finding a statistical relationship between itemsets. Our method is related to such causal discovery research because the minimal and consistent rules with time delays follow some notions of the causal theories. In comparison, their methods basically analyze the values of original data, but our work attempts to estimate quantitative changes (including fast-changing events) and time-delayed connections in the data of two temporal databases. In other words, our algorithm formulates association rules under the indiscernibility of unifying similar data behaviors and the weight of measuring quantities. These rules are significantly used to evaluate the connectedness of two databases for each different time delay.

In the area of rough set theory, several studies have been undertaken on mining temporal data. Hirano and Tsumoto [25] presented a method for finding patterns from spatio-temporal data using rough-set-based clustering. This approach can group sequences from a single spatio-temporal information system wherein data are associated with time and spatial positions. In comparison, our work proposes to distinguish between the indiscernibility and weight in the rule generation so that fast changing events and time-delayed connections can be detected from the quantitative changes in distributed data.

As rough set research, Milton, Maheswari, and Siromoney [26] proposed simply combining multiple information systems on the axis of common attributes. Their work is based on the relational learning of multiple relations in inductive logic programming. However, in the research, no evaluation criteria for connecting multiple information systems were presented and no algorithm for analyzing quantitative changes and time delays between two information systems was considered. In addition, Inuiguchi [27] extended a

rule induction method for multi-agent rough sets. His method can analyze a conflict tolerance of multiple decision tables that have identical objects and attributes. In contrast, we consider multiple information systems with different attributes for analyzing their changes and time-delayed connections. By using the time stamps of data, a connected decision table is constructed such that one information system is used as condition attributes and the other is used as decision attributes.

In the discussion of the above related works, we summarize that our approach is based on a new combination of the following three aspects of causality mining:

- (i) Data mining for the qualitative changes in time series data
- (ii) Discarding inconsistent rules as causal rules using rough set theory
- (iii) Analyzing various time delays between the data of two temporal information systems

No other approach covers all of the three aspects in (i), (ii), and (iii) that enable us to discover causal relations across multiple databases. The combination is implemented in our algorithm that cannot be achieved by simply combining existing works, as discussed below.

First, related works on rough set theory do not analyze the qualitative changes in data for (i) and various time delays between two temporal information systems for (iii). For causality mining, these are very important for exploring differently time-delayed data behaviors across two temporal information systems.

Second, frequency-based data mining algorithms or Bayesian-based algorithms can find a correlation between data, yet they do not consider logical inconsistency of data for (ii) and various time delays in causal relations for (iii). The consistency analysis is necessary to exclude some redundant or inconsistent causal rules when evaluating a large amount of generated rules for various time delays.

Third, the discovery of a time delay that implies a strong connection between multiple data sets is considered as an original problem. To the best of my knowledge, existing works on causality mining have not attempted to find a time delay wherein the data behaviors in two temporal information systems strongly affect each other in consistent causal rules.

6. Conclusion and Future Work

We have proposed a method for mining the changes and time-delayed connections from two temporal information systems in rough set theory. As a novel approach, we establish a distinction between the indiscernibility and weight of quantitative changes in rough-set rule generation and evaluation. Our method contains quantitative estimations for extracting changes in numeric values from different data sets. The proposed mining algorithm slidingly connects the quantitative changes in distributed data to generate (candidate) cause-effect decision tables for various time delays. We devise an evaluation method for the consistency in the association rules of changes by adjusting weight-based accuracy and coverage in order to compute the connectedness between two information systems. The experimental result indicates that our method can discover certain time delays that causally connect climate data and medical data on Tokyo.

Our future research is concerned with establishing an algebra for complex queries of quantitative estimation operators in order to discover changes and connections across multiple information systems for *users' requirements*. In practical applications, users could indicate such complex queries that help decide the direction to capture the implicit essence of data when data mining relies on domain knowledge.

7. Acknowledgments

This research has been partially supported by the Japanese Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (20700147).

References

- [1] K. Kaneiwa. A rough set approach to mining connections from information systems. In *Proceedings of the 25th ACM Symposium on Applied Computing (ACM SAC 2010), Track on Data Mining*, pages 990–996, 2010.
- [2] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

- [3] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 1–12, 2000.
- [4] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, 1992.
- [5] J. Komorowski, L. Polkowski, and A. Skowron. Rough sets: a tutorial. In S.K. Pal and editors A. Skowron, editors, *Rough-Fuzzy Hybridization: A New Method for Decision Making*, pages 3–98. Springer-Verlag, 1998.
- [6] L. Polkowski, S. Tsumoto, and T. Young Lin. A rough set perspective on knowledge discovery in information systems. In *Rough set methods and applications*, pages 9–45. Physica-Verlag, 2000.
- [7] Y. Kudo and T. Murai. A note on characteristic combination patterns about how to combine objects in object-oriented rough set models. In *Third International Conference on Rough Sets and Knowledge Technology (RSKT 2008)*, pages 115–123, 2008.
- [8] D. W. Cheung, V. T. Y. Ng, A. W. Fu, and Y. Fu. Efficient mining of association rules in distributed databases. *IEEE Trans. Knowl. Data Eng.*, 8(6):911–922, 1996.
- [9] R. Jin and G. Agrawal. Systematic approach for optimizing complex mining tasks on multiple databases. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006)*, page 17, 2006.
- [10] X. Zhu and X. Wu. Discovering relational patterns across multiple databases. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007)*, pages 726–735, 2007.
- [11] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000.
- [12] D. D. Jensen. Beyond prediction: Directions for probabilistic and relational learning. In *Proceedings of the 17th International Conference on Inductive Logic Programming (ILP 2007)*, LNCS 4894, pages 4–21, 2007.

- [13] D. D. Jensen, A. S. Fast, B. J. Taylor, and M. E. Maier. Automatic identification of quasi-experimental designs for discovering causal knowledge. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 372–380, 2008.
- [14] C. Silverstein, S. Brin, R. Motwani, and J. D. Ullman. Scalable techniques for mining causal structures. *Data Min. Knowl. Discov.*, 4(2/3):163–192, 2000.
- [15] K. Inoue. Automated abduction. In A. C. Kakas and F. Sadri, editors, *Computational Logic: Logic Programming and Beyond*, LNCS 2408, pages 311–341. Springer, 2002.
- [16] M. Tooley. *Time, Tense, and Causation*. Oxford University Press, 2000.
- [17] M. J. Crawley. *Statistics: An Introduction using R*. Wiley, 2005.
- [18] O. Aberth. *Computable Calculus*. Academic Press, 2001.
- [19] J. Galindo. *Fuzzy Databases: Modeling, Design, and Implementation*. IGI Publishing, 2006.
- [20] S. K. Pal and P. Mitra. *Pattern Recognition Algorithms for Data Mining: Scalability, Knowledge Discovery, and Soft Granular Computing*. Chapman & Hall, Ltd., 2004.
- [21] J. G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, and J. Wróblewski. Rough set algorithms in classification problem. In L. Polkowski, S. Tsumoto, and T. Young Lin, editors, *Rough set methods and applications*, pages 49–88. Physica-Verlag, 2000.
- [22] <http://www.jma.go.jp/jma/indexe.html>.
- [23] <http://www.chieiken.gr.jp/infulurep/pub/rep.do>.
- [24] K. Karimi and H. J. Hamilton. Distinguishing causal and acausal temporal relations. In *Proc. of the 7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2003)*, LNCS 2637, pages 234–240, 2003.

- [25] S. Hirano and S. Tsumoto. A clustering method for spatio-temporal data and its application to soccer game records. In *Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, pages 612–621, 2005.
- [26] R. S. Milton, V. U. Maheswari, and A. Siromoney. Studies on rough sets in multiple tables. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 10th International Conference (RSFDGrC 2005)*, pages 265–274, 2005.
- [27] M. Inuiguchi. Rule induction based on multi-agent rough sets. In *Computational Intelligence and Industrial Applications: Proceedings of ISCIIA 2006*, pages 320–327, 2006.
- [28] L. Polkowski. *Rough Sets: Mathematical Foundations*. Physica-Verlag, Heidelberg ; New York, 2002.

Appendix

A. Rough Sets, Information Systems, and Decision Rules

In this appendix, we introduce the basic concept of rough set theory, based on the mathematical foundations in [28, 5].

An attribute a is a mapping $a: U \rightarrow V_a$ where U is a non-empty finite set of objects (called the universe) and V_a is the value set of a . An information system is a pair $T = (U, A)$ of the universe U and a non-empty finite set A of attributes. Let B be a subset of A . The B -indiscernibility relation is defined by an equivalence relation I_B on U such that $I_B = \{(x, y) \in U^2 \mid \forall a \in B. a(x) = a(y)\}$. The equivalence class of I_B for each object $x (\in U)$ is denoted by $[x]_B$. Let X be a subset of U . We define the lower and upper approximations of X by $\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\}$ and $\overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \emptyset\}$. A subset B of A is a reduct of T if $I_B = I_A$ and there is no subset B' of B with $I_{B'} = I_A$ (i.e., B is a minimal set of attributes without losing discernibility).

A decision table is an information system $T = (U, A \cup \{d\})$ such that each $a \in A$ is a condition attribute and $d \notin A$ is a decision attribute. Let V_d be the value set $\{d_1, \dots, d_u\}$ of the decision attribute d . For each value $d_i \in V_d$, we obtain a decision class $U_i = \{x \in U \mid d(x) = d_i\}$ where $U = U_1 \cup \dots \cup U_{|V_d|}$ (i.e., $u = |V_d|$) and for every $x, y \in U_i$, $d(x) = d(y)$. The B -positive region of

d is defined by $P_B(d) = \underline{B}(U_1) \cup \dots \cup \underline{B}(U_{|V_d|})$. A subset B of A is a relative reduct of T if $P_B(d) = P_A(d)$ and there is no subset B' of B with $P_{B'}(d) = P_A(d)$.

We define a formula $(a_1 = v_1) \wedge \dots \wedge (a_n = v_n)$ in T (denoting the condition of a rule) where $a_j \in A$ and $v_j \in V_{a_j}$ ($1 \leq j \leq n$). The semantics of the formula in T is defined by $\llbracket (a_1 = v_1) \wedge \dots \wedge (a_n = v_n) \rrbracket_T = \{x \in U \mid a_1(x) = v_1, \dots, a_n(x) = v_n\}$. Let φ be a formula $(a_1 = v_1) \wedge \dots \wedge (a_n = v_n)$ in T . A decision rule for T is of the form $\varphi \rightarrow (d = d_i)$, and it is true if $\llbracket \varphi \rrbracket_T \subseteq \llbracket (d = d_i) \rrbracket_T (= U_i)$. The accuracy and coverage of a decision rule r of the form $\varphi \rightarrow (d = d_i)$ are respectively defined as follows.

$$\text{accuracy}(T, r, U_i) = \frac{|U_i \cap \llbracket \varphi \rrbracket_T|}{|\llbracket \varphi \rrbracket_T|}$$

$$\text{coverage}(T, r, U_i) = \frac{|U_i \cap \llbracket \varphi \rrbracket_T|}{|U_i|}$$

In the evaluations, $|U_i|$ is the number of objects in a decision class U_i and $|\llbracket \varphi \rrbracket_T|$ is the number of objects in the universe $U = U_1 \cup \dots \cup U_{|V_d|}$ that satisfy condition φ of rule r . Therefore, $|U_i \cap \llbracket \varphi \rrbracket_T|$ is the number of objects satisfying the condition φ restricted to a decision class U_i .

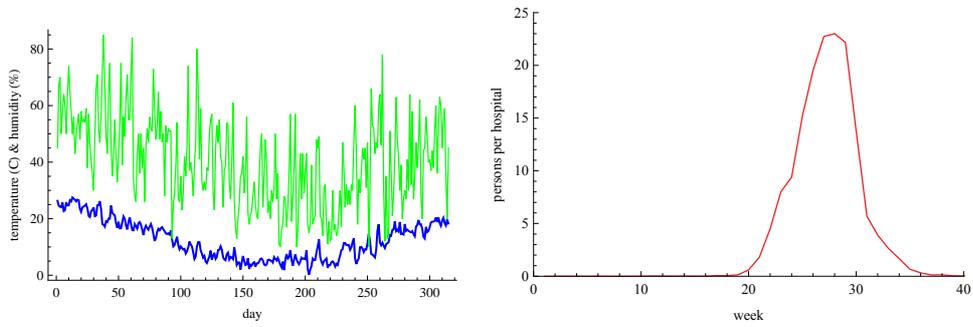


Figure 1: Climate data and medical data in Tokyo.

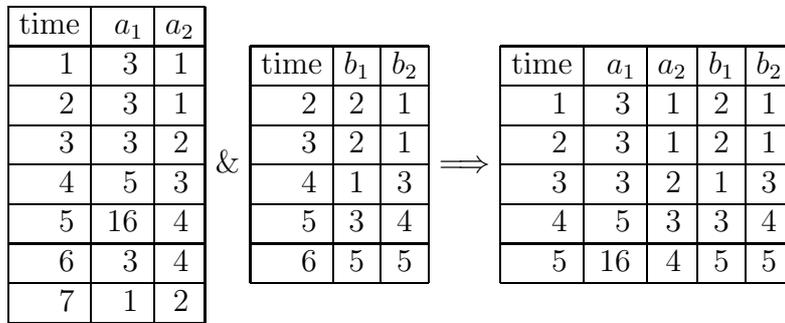


Figure 2: The connection $con(T_1, T_2, 1)$ of $T_1 = (Nat_{1,7}, A_1)$ and $T_2 = (Nat_{2,6}, A_2)$ with time delay $\Delta = 1$.

time	a_1	a_2
1	3	1
2	3	1
3	3	2
4	5	3
5	16	4
6	3	4
7	1	2

 \Rightarrow

time	a_1	a_2
2	0	0
3	0	+1
4	+2	+1
5	+11	+1
6	-13	0
7	-2	-2

time	b_1	b_2
2	2	1
3	2	1
4	1	3
5	3	4
6	5	5

 \Rightarrow

time	b_1	b_2
2	0	0
3	0	0
4	0	1
5	1	1
6	1	1

Figure 3: The quantitative changes $\pi_d(T_1)$ and $\pi_{\geq 2.5}(T_2)$.

time	a_1	a_2
1	0	0.0
2	0	+1.0
3	+2	+0.5
4	+11	+1/3
5	-13	0.0
6	-2	-0.5

 $\&$

time	b_1
2	0
3	0
4	0
5	1
6	1

 \Rightarrow

time	a_1	a_2	b_1
1	0	0	0
2	0	+1.0	0
3	+2	+0.5	1
4	+11	+1/3	1

Figure 4: The connected decision table $con(\pi_d(T_1[a_1]) \circ \pi_v(T_1[a_2]), \pi_{\geq 2.5}(T_2[b_1]), 2)$ with time delay $\Delta = 2$.

Algorithm: *Change and Connection Mining*

Input: temporal information systems T_1 and $T_2[b]$,
maximum time delay m ($\in \text{Nat}$)

Output: time delay t

```
1: begin
2:    $T'_1 = \pi_1(T_1[a_1]) \circ \dots \circ \pi_n(T_1[a_n]); T'_2 = \pi(T_2[b]);$ 
3:   for  $\Delta = 0$  to  $m$  do
4:      $T = \text{con}(T'_1, T'_2, \Delta);$ 
5:      $\mathcal{R} = \text{reducts}(T);$  (based on  $I_B^{qc}$ )
6:     for  $B \in \mathcal{R}$  do
7:        $C_B = 0;$ 
8:       for  $i = 1$  to  $|\text{sign}(V_d)|$  do
9:          $\text{max\_eval}_i = 0;$ 
10:        for  $x \in U_i$  do
11:           $r = \text{rule}(x, B, T);$ 
12:           $\text{eval}_i = \text{eval}(T, r, U_i);$  (using  $\mu_S$ )
13:          if  $\text{max\_eval}_i < \text{eval}_i$  then  $\text{max\_eval}_i = \text{eval}_i;$ 
14:        rof
15:         $C_B = C_B + \text{max\_eval}_i;$ 
16:      rof
17:    rof
18:     $\text{connectedness}_\Delta = \max(\{C_B \mid B \in \mathcal{R}\});$ 
19:  rof
20:  return  $t$  ( $\text{connectedness}_t = \max(\{\text{connectedness}_j \mid 0 \leq j \leq m\});$ )
21: end;
```

Figure 5: The change and connection mining algorithm for two temporal information systems T_1 and $T_2[b]$.

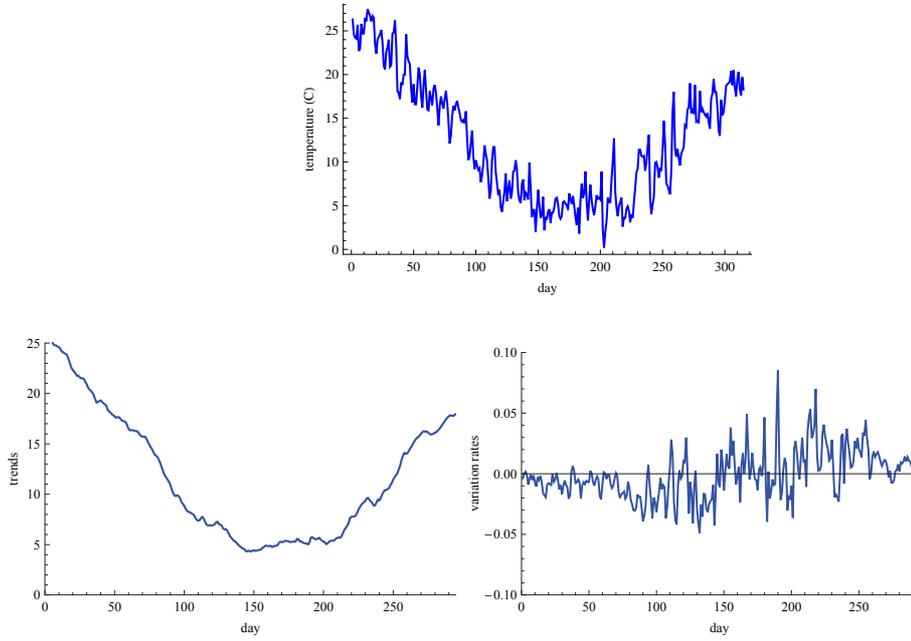


Figure 6: Temperature data in $T_1[a_t]$, and the quantitative changes $\pi_{tr(10)}(T_1[a_t])$ and $\pi_v(\pi_{tr(10)}(T_1[a_t]))$.

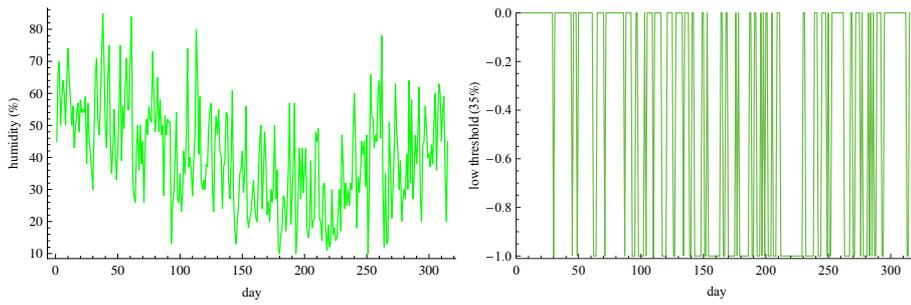


Figure 7: Humidity data in $T_1[a_h]$ and the quantitative change $\pi_{\leq 35}(T_1[a_h])$.

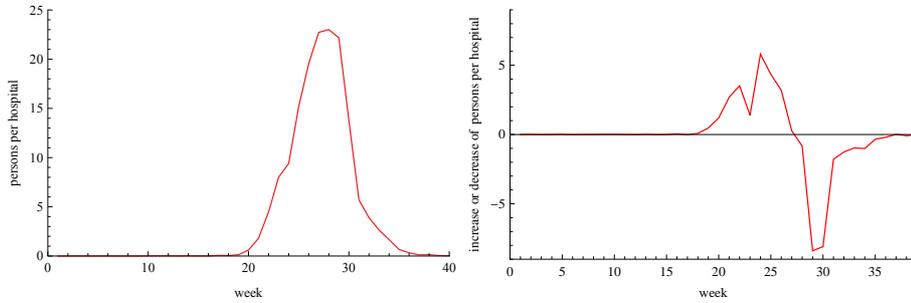


Figure 8: Influenza data in $T_1[b_f]$ and the quantitative change $\pi_d(T_2[b_f])$.

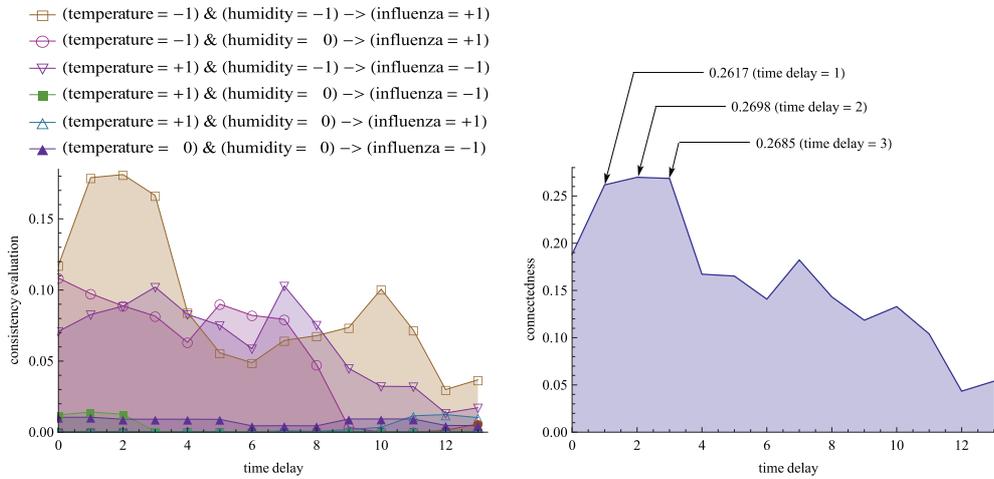


Figure 9: The consistency evaluation and the connectedness in $con(\pi_v(\pi_{tr(10)}(T_1[a_t])) \circ \pi_{\leq 35}(T_1[a_h]), exp_7(\pi_d(T_2[b_f]), \Delta))$.