A Rough Set Approach to Mining Connections from Information Systems

Ken Kaneiwa
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika, Soraku, Kyoto 619-0289, Japan
kaneiwa@nict.go.jp

ABSTRACT

Mining data changes and connections from information systems (or databases) is made difficult by the different data behaviors and relationships across multiple data sets. When making a decision, such a dynamic and integrated knowledge base can be used to set useful rules (e.g., causality) that differ from the statistical associations in a single resource. In this paper, using techniques based on the rough set theory, we propose 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.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—Data mining

General Terms

Algorithms

Keywords

Rough set theory, change and connection mining, causality

1. INTRODUCTION

As a tool for data analysis, data mining algorithms [3, 7] enable us to discover useful patterns and rules from information systems (or databases). In the algorithms, association rules are generated from the frequent attributes (or itemsets) in an information system. In addition, numerous techniques for extracting association rules (known as decision rules) have been proposed in the field of rough set theory [18, 14, 20, 15]. Unlike conventional algorithms for mining statistical associations, the rough set approach provides an algorithm for logically formulating association rules in a decision table. More specifically, minimal and consistent association rules are computed using the lower approximation of a rough set by checking for any logical combination of condition and decision attributes.

More importantly, the integration of multiple information systems is carried out to create useful data that a single information system does not yield. In the environment of computer systems, sources of data are distributed across multiple sites, contexts, and domains (e.g., data on the web). Therefore, there is an urgent requirement to integrate and analyze distributed data for discovering valuable patterns and rules across multiple information systems. The mining of association rules from multiple information systems has to be realized in a highly sophisticated manner, as it involves analyzing and integrating various data in different contexts of the systems. Existing data mining algorithms [5, 12, 23] effectively and statistically integrate multiple data resources; however, they do not attempt to establish data behaviors and relationships across multiple data sets.

To enable integration and discovery, time and space stamps should be used as references for examining changes and connections in different data sets. The references can be used to indicate a time delay between the time-stamped data of two distributed information systems. The quantitative changes in the temporal data of such information systems 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.

Causality [19, 10] derived from changes and connections is valuable knowledge obtained from the integration of multiple information systems. This is because changing the values of data in one context may affect the values of data in another context. Therefore, over and above the statistical associations in a single information system, such causal knowledge has an advantage in that each of the conditions implies a different effect. However, there are few approaches to mining the changes and connections in data across multiple information systems. As discussed in [11, 21], 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 the rough set theory. We assume

SAC'10 March 22-26, 2010, Sierre, Switzerland.
Copyright 2010 ACM 978-1-60558-638-0/10/03 ...$10.00.

that attribute data in information systems are associated with time stamps. For the purpose of mining, we formalize the quantitative changes that are 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.

This paper is arranged as follows. Section 2 recalls the basic notions of rough sets. In Section 3, we describe a connection method of two temporal information systems for various time delays. In Section 4, we present our algorithm for mining changes and connections from two temporal information systems. The experimental results are reported in Section 5. Finally, we conclude this paper in Section 6.

2. ROUGH SETS

An attribute a is a mapping $a\colon U \to V_a$ 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 $\{d_1, \ldots, d_u\}$ 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 $U = U_1 \cup \cdots \cup U_{|V_d|}$ (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 $P_B(d) = \underline{B}(U_1) \cup \cdots \cup \underline{B}(U_{|V_d|})$. 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 $(a_1 = v_1) \wedge \cdots \wedge (a_n = v_n)$ 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 $[\![(a_1 = v_1) \wedge \cdots \wedge (a_n = v_n)]\!]_T = \{x \in U \mid a_1(x) = v_1, \ldots, a_n(x) = v_n\}$. Let φ be a formula $(a_1 = v_1) \wedge \cdots \wedge (a_n = v_n)$ in T. A decision rule for T is of the form $\varphi \to (d = d_i)$, and it is true if $[\![\varphi]\!]_T \subseteq [\![(d = d_i)]\!]_T (= U_i)$. The accuracy and coverage of a decision rule r of the form $\varphi \to (d = d_i)$ are respectively defined as follows.

$$accuracy(T, r, U_i) = \frac{|U_i \cap [\![\varphi]\!]_T|}{|[\![\varphi]\!]_T|}$$

$$coverage(T, r, U_i) = \frac{|U_i \cap [\![\varphi]\!]_T|}{|U_i|}$$

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

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

&

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

$\Longrightarrow$

time	a_1	a_2	b_1	b_2
1	3	1	2	1
2	3	1	2	1
3	3	2	1	3
4	5	3	3	4
5	16	4	5	5

Figure 1: 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 $\Delta = 1$.

3. CHANGES AND CONNECTIONS IN DATA

3.1 Connecting Information Systems

A temporal information system is an information system $T = (U_{time}, A)$ where 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 $n\colon U_{time} \to Nat_{i,j}$ such that for every $x, y \in U_{time}$, $x < y \Leftrightarrow n(x) < n(y)$, $|U_{time}| = |Nat_{i,j}|$, and $A' = \{a' \mid a \in A \;\&\; \forall x \in U_{time}.a'(n(x)) = a(x)\}$.

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 Δ $(\in Nat)$ is defined as an information system $T = (U_{time}, A)$ such that

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

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

In Figure 1, 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.

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 the 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 $b \in A_2$. 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.

3.2 Changes and Time Delays

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 or not 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). Several quantitative estimation operators for the numeric values $a(x)$ of attributes are defined by the following.

$$\text{(difference) } \pi_d(a(x)) = a(x) - a(x-1)$$

$$\text{(variation rate) } \pi_v(a(x)) = \frac{a(x) - a(x-1)}{a(x)}$$

$$\text{(threshold) } \pi_{\geq k}(a(x)) = \begin{cases} 1 & \text{if } \ a(x) \geq k \\ 0 & \text{otherwise} \end{cases}$$

$$\text{(variation rate of difference) } \pi_{2v}(a(x)) = \pi_v(\pi_d(a(x)))$$

$$\text{(trend) } \pi_{tr(k)}(a(x)) = \frac{a(x{-}k){+}\cdots{+}a(x){+}\cdots{+}a(x{+}k)}{2k+1}$$

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 2 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.

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.

$$con(\pi_1(T_1), \pi_2(T_2[b]), 0), con(\pi_1(T_1), \pi_2(T_2[b]), 1),$$
$$con(\pi_1(T_1), \pi_2(T_2[b]), 2), \ldots, con(\pi_1(T_1), \pi_2(T_2[b]), m)$$

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

$$con(\pi_1(T[a_1]), con(\pi_2(T[a_2]), \cdots con(\pi_{n-1}(T[a_{n-1}]),$$
$$\pi_n(T[a_n]), 0) \cdots, 0), 0)$$

which is simply denoted by $\pi_1(T[a_1]) \circ \cdots \circ \pi_n(T[a_n])$. 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 3 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 1 and Figure 2 are transformed into 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 $\Delta = 2$ and the decision attribute $b_1 \in A_2$.

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

$\Longrightarrow$

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

$\Longrightarrow$

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

Figure 2: 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

$\Longrightarrow$

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 3: 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 $\Delta = 2$.

A decision rule in the rough set theory is called an association rule of changes if it is generated from a cause-effect decision table $con(\pi_1(T_1), \pi_2(T_2[b]), \Delta)$ consisting of the quantitative changes $\pi_1(T_1)$ and $\pi_2(T_2[b])$.

For example, the following association rules of changes are generated from the connected decision table $con(\pi_d(T_1[a_1]) \circ \pi_v(T_1[a_2]), \pi_{\geq 2.5}(T_2[b_1]), 2)$ in Figure 3.

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

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.

3.3 Indiscernibility and Weight

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 $n = 0$.

By the B-indiscernibility I_B^{qc} of quantitative changes, the semantics of the formula $(a_1 = v_1) \wedge \cdots \wedge (a_n = v_n)$ in T is refined by $[\![(a_1 = v_1) \wedge \cdots \wedge (a_n = v_n)]\!]_T^{qc} = \{x \in U_{time} \mid sign(a_1(x)) = sign(v_1), \ldots, sign(a_n(x)) = sign(v_n)\}$. Let $\{s_1, \ldots, s_u\}$ 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 $U = U_1 \cup \cdots \cup U_{|sign(V_d)|}$ (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 $||\ ||\colon R \to R$ is the absolute value function such that $||n|| = n$ if $n \geq 0$, otherwise, $||n|| = -n$. For example, the connected decision table in Figure 3 contains fast changing events because $a_1(4) = +11$ is an intensively higher value than the other values. Therefore, we obtain high weight $w(4) = 12.33\cdots$ 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 $(a_1 = 0) \wedge (a_2 = 0.0) \to (b_1 = 0)$ and $(a_1 = +11) \wedge (a_2 = +1/3) \to (b_1 = 1)$ (shown in Section 3.2). 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 $\varphi \to (d = d(x))$ with $d(x) \in V_d$ are defined as follows.

$$w\_accuracy(T, r, U_i) = \frac{w(U_i \cap [\![\varphi]\!]_T^{qc})}{w([\![\varphi]\!]_T^{qc})}$$

$$w\_coverage(T, r, U_i) = \frac{w(U_i \cap [\![\varphi]\!]_T^{qc})}{w(U_i)}$$

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.

3.4 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 a most consistent association rule for each decision class. Second, we estimate the total value of the maximum consistency evaluations for all the 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.

$$w\_coverage(T, r, U_i) \times \mu_S(w\_accuracy(T, r, U_i))$$

The fuzzy membership function $\mu_S\colon [0, 1] \to [0, 1]$ (similar to the S fuzzy set [6]) is defined by the following.

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

By applying the membership function to each accuracy value, we determine whether or not 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; however, we intend to find consistent rules beyond statistical associations. The reason why we use the fuzzy membership function is 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 $T = (U, A \cup \{d\})$.

$$r_1\colon (a_1 = +1) \wedge (a_2 = -1) \to (d = +1)$$
$$r_2\colon (a_1 = +1) \wedge (a_2 = -1) \to (d = -1)$$

These association rules have identical conditions; however, 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 therefore, 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

$$connectedness(T, B) = \sum_{i=1,\ldots,|sign(V_d)|} max\_eval(T, R_B, U_i)$$

The underlying assumption behind the maximum evaluation $max\_eval(T, R_B,\ U_i)$ is that the connectedness is

Algorithm: Change and Connection Mining

```
Input: temporal information systems $T_1$ and $T_2[b]$,
       maximum time delay $m$ ($\in Nat$)
Output: time delay $t$
1:  $T_1' = \pi_1(T_1[a_1]) \circ \cdots \circ \pi_n(T_1[a_n])$; $T_2' = \pi(T_2[b])$;
2:  for $\Delta = 0$ to $m$ do
3:      $T = con(T_1', T_2', \Delta)$;
4:      $\mathcal{R} = reducts(T)$; (based on $I_B^{qc}$)
5:      for $B \in \mathcal{R}$ do
6:          $C_B = 0$;
7:          for $i = 1$ to $|sign(V_d)|$ do
8:              $max\_eval_i = 0$;
9:              for $x \in U_i$ do
10:                 $r = rule(x, B, T)$;
11:                 $eval_i = eval(T, r, U_i)$; (using $\mu_S$)
12:                 if $max\_eval_i < eval_i$
13:                 then $max\_eval_i = eval_i$;
14:             rof
15:         $C_B = C_B + max\_eval_i$;
16:         rof
17:     rof
18:     $connectedness_\Delta = max(\{C_B \mid B \in \mathcal{R}\})$;
19: rof
20: return $t$ ($connectedness_t =$
21:             $max(\{connectedness_j \mid 0 \leq j \leq m\})$);
```

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 the 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.

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

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, it does not indicate the connectedness of condition and decision attributes.

4. CHANGE AND CONNECTION MINING ALGORITHM

Figure 4 shows a change and connection mining algorithm for two temporal information systems T_1 and $T_2[b]$ and a maximum time delay $m (\in Nat)$.

First of all, in order to analyze hidden behaviors in $T_1 = (U_{time}, \{a_1, \ldots, a_n\})$ and $T_2[b] = (U'_{time}, \{b\})$, quantitative estimation operators $\pi_1, \ldots, \pi_n$ and π are applied to $T_1[a_1]$, $\ldots, T_1[a_n]$ and $T_2[b]$, and the estimated results are stored in the variables T_1' and T_2' (in Line 1). Since we are not certain which time delay constructs a temporal relation suitable for T_1' and T_2', this algorithm functions to connect them in varying time delays Δ from 0 to m (in Lines 2 - 19). In the loop of the time delays, the connected decision table $T = (U''_{time}, \{a_1, \ldots, a_n\} \cup \{b\})$ of T_1' and T_2' is computed by $T = con(T_1', T_2', \Delta)$ (in Line 3). Therefore, the connected decision table T is one of the candidate cause-effect decision tables to generate association rules of changes. Using the rough set theory, minimal association rules $r = rule(x, B, T)$ (as in Line 10) are created for each relative reduct B in $\mathcal{R} = reducts(T)$ [17], i.e., B is a subset of the set $\{a_1, \ldots, a_n\}$ 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 [4]. 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 $U''_{time} = U_1 \cup \cdots \cup U_{|sign(V_b)|}$ 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, the association rules $r = rule(x, B, T)$ for all the time stamps x in U_i are generated for every $i = 1, \ldots, |sign(V_b)|$ (in Lines 7 - 16). 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. For each i from 1 to $|sign(V_b)|$, we obtain the maximum consistency evaluation (stored in the variable max_eval_i) by calculating $eval(T, r, U_i)$ (in Lines 11 - 13). Finally, the connectedness of T_1' and T_2' for each relative reduct B (stored in variable C_B) is calculated (in Line 15). 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 2 - 19), a time delay t with the maximum connectedness is returned by comparing the connectedness for each time delay in $0 \leq j \leq m$.

5. 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 4, by downloading climate data and medical data in Tokyo from the web sites of [2, 1]. On the left-hand side of Figure 4, the climate data set contains the daily observed data of minimum temperature and humidity in Tokyo from August 7 2006 until July 15 2007. On the right-hand side of Figure 4, the medical data set consists of weekly reported numbers of influenza victims per hospital in Tokyo from September 4 2006 until 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. For the climate data, first, to exclude noisy data the long-term behaviors of daily temperatures are estimated by the trends $\pi_{tr(10)}(T_1[a_t])$ of the temperatures in $T_1[a_t]$. Then, the increase-decrease rates of the trends are calculated by the variation rates $\pi_v(\pi_{tr(10)}(T_1[a_t]))$. 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 de-

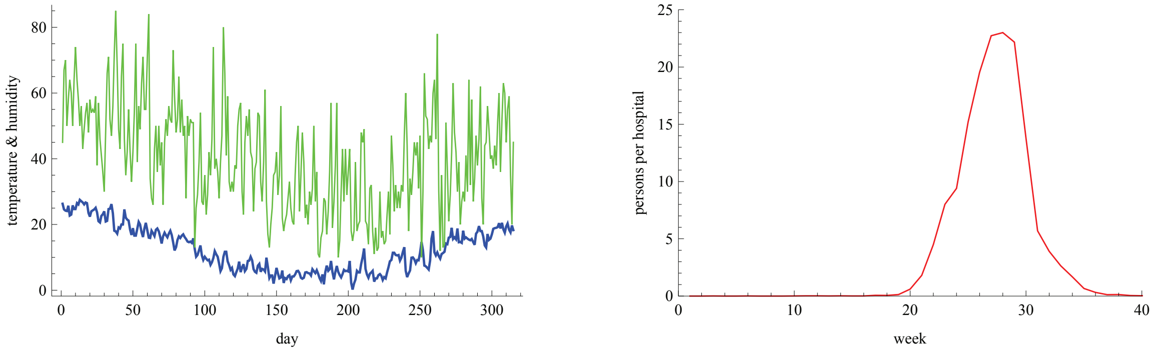

Figure 4: Climate data and medical data in Tokyo.

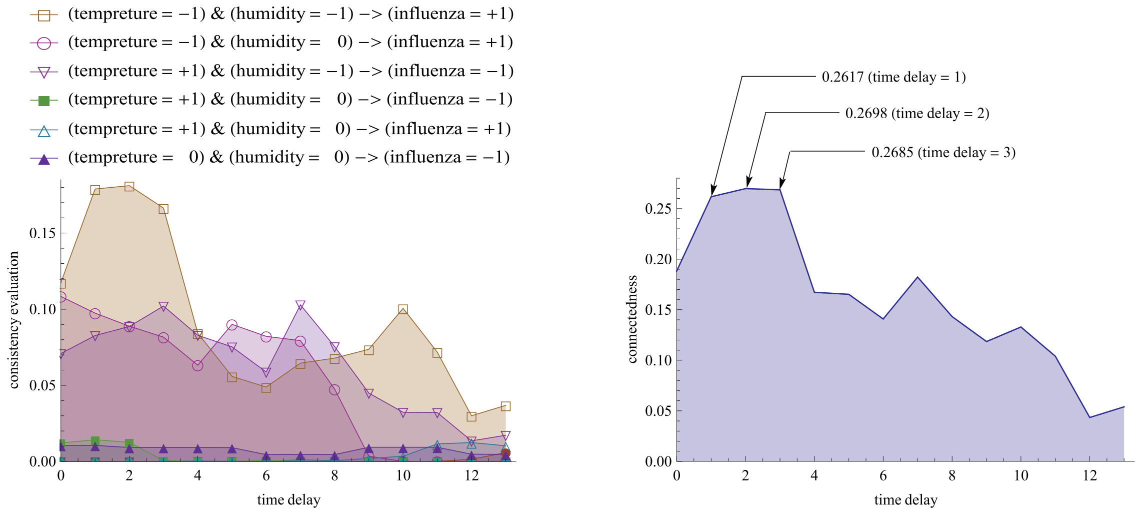

Figure 5: The consistency evaluation and the connectedness in $con(\pi_v(\pi_{tr(10)}(T_1[a_t])) \circ \pi_{\le 35}(T_1[a_h]), exp_7(\pi_d(T_2[b_f]), \Delta))$.

crease to equal to or less than 35%. For the medical data, the differences $\pi_d(T_2[b_f])$ of the numbers of influenza victims in $T_2[b_f]$ are regarded as candidate effects of the quantitative changes $\pi_v(\pi_{tr(10)}(T_1[a_t])) \circ \pi_{\le 35}(T_1[a_h])$. Unlike estimating the variation rates of temperatures, the numbers of victims are absolute values; therefore, the differences should be calculated (since absolute values contain 0). After the estimation, we turn to connecting the climate data and the medical data by using the quantitative changes $\pi_v(\pi_{tr(10)}(T_1[a_t])) \circ \pi_{\le 35}(T_1[a_h])$ 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 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 5 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 $con(\pi_v(\pi_{tr(10)}(T_1[a_t])) \circ \pi_{\le 35}(T_1[a_h]), exp_7(\pi_d(T_2[b_f])), \Delta)$. The left-hand side of Figure 5 shows the consistency evaluations $eval(T, r, U_i)$ for the time delays $\Delta = 0, \ldots, 13$ 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 $(a_t = sign(a_t(x))) \wedge (a_h = sign(a_h(x))) \rightarrow (b_f = sign(b_f(x)))$ for a time stamp $x \in U_i$ where $sign(a_t(x)) \in \{-1, 0, +1\}$, $sign(a_h(x)) \in \{-1, 0\}$, and $sign(b_f(x)) \in \{-1, 0, +1\}$. 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 $(a_t = -1) \wedge (a_h = -1) \rightarrow (b_f = +1)$ has a high consistency evaluation through the whole time delays, compared with the other rules. Note that the value of each generalized association rule in the figure totally 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 values of the temperature decrease and the humidity are equal to or under the 35 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 5, our experimental result indicates the connectedness evaluated for each of the time delays $\Delta = 0, \ldots, 13$. 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 $T_2[b_f]$. Further, high connectedness values 0.2685 and 0.2617 are given for the time delays $\Delta = 3$ and 1 (denoting three days and one day). This result confirms the existence of certain time delays that strongly connect the climate data and the medical data of Tokyo.

6. CONCLUSION

We have proposed a method for mining the changes and time-delayed connections from two temporal information systems in the rough set theory. As a novel approach, we establish a distinction between the indiscernibility and weight of quantitative changes in the rough-set rule generation and evaluation. Our method contains the quantitative estimations for extracting the 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 the climate data and the medical data of Tokyo.

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).

8. REFERENCES

[1] http://www.chieiken.gr.jp/infulurep/pub/rep.do.

[2] http://www.jma.go.jp/jma/indexe.html.

[3] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In Proc. of the 1993 ACM SIGMOD Int. Conf. on Management of Data, pages 207–216, 1993.

[4] 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. Y. Lin, editors, Rough set methods and applications, pages 49–88. Physica-Verlag, 2000.

[5] 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.

[6] J. Galindo. Fuzzy Databases: Modeling, Design, and Implementation. IGI Publishing, 2006.

[7] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, pages 1–12, 2000.

[8] S. Hirano and S. Tsumoto. A clustering method for spatio-temporal data and its application to soccer game records. In Proc. of the 10th Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, pages 612–621, 2005.

[9] M. Inuiguchi. Rule induction based on multi-agent rough sets. In Computational Intelligence and Industrial Applications: Proc. of ISCIIA 2006, pages 320–327, 2006.

[10] D. D. Jensen. Beyond prediction: Directions for probabilistic and relational learning. In Proc. of the 17th Int. Conf. on Inductive Logic Programming (ILP 2007), LNCS 4894, pages 4–21, 2007.

[11] 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.

[12] R. Jin and G. Agrawal. Systematic approach for optimizing complex mining tasks on multiple databases. In Proc. of the 22nd Int. Conf. on Data Engineering (ICDE 2006), page 17, 2006.

[13] K. Karimi and H. J. Hamilton. Distinguishing causal and acausal temporal relations. In Proc. of the 7th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD 2003), LNCS 2637, 2003.

[14] J. Komorowski, L. Polkowski, and A. Skowron. Rough sets: a tutorial. In S. Pal and e. A. Skowron, editors, Rough-Fuzzy Hybridization: A New Method for Decision Making, pages 3–98. Springer-Verlag, 1998.

[15] Y. Kudo and T. Murai. A note on characteristic combination patterns about how to combine objects in object-oriented rough set models. In Third Int. Conf. on Rough Sets and Knowledge Technology (RSKT 2008), pages 115–123, 2008.

[16] 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 Int. Conf. (RSFDGrC 2005), pages 265–274, 2005.

[17] S. K. Pal and P. Mitra. Pattern Recognition Algorithms for Data Mining: Scalability, Knowledge Discovery, and Soft Granular Computing. Chapman & Hall, Ltd., 2004.

[18] Z. Pawlak. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, 1992.

[19] J. Pearl. Causality: models, reasoning, and inference. Cambridge University Press, 2000.

[20] L. Polkowski, S. Tsumoto, and T. Y. Lin. A rough set perspective on knowledge discovery in information systems. In Rough set methods and applications, pages 9–45. Physica-Verlag, 2000.

[21] 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.

[22] M. Tooley. Time, Tense, and Causation. Oxford University Press, 2000.

[23] X. Zhu and X. Wu. Discovering relational patterns across multiple databases. In Proc. of the 23rd Int. Conf. on Data Engineering (ICDE 2007), pages 726–735, 2007.