Mining Correlated Association Rules from Multi-Relational Data using Interval Patterns

Hirohisa Seki and Masahiro Nagao*

Dept. of Computer Science, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
seki@nitech.ac.jp, m.nagao.738@nitech.jp

Abstract. In this paper, we consider the problem of mining numerical association rules (ARs) from a multi-relational database (MRDB). More specifically, we examine the effectiveness of numerical ARs with *interval patterns (IPs)* proposed by Kaytoue *et al.* in FCA (Formal Concept Analysis), and show that the *MinIntChange* algorithm by Kaytoue *et al.* can be readily extended to mine correlated interval-based ARs with the maximal significance in terms of the χ^2 measure, by incorporating into the algorithm a pruning technique by Morishita *et al.* Moreover, since the search space for computing closed IPs becomes larger as the number of numerical attributes increases, we utilize *Super CWC*, an off the shelf feature selection algorithm to reduce the number of attributes to use. Our approach is experimentally evaluated and compared with the conventional methods such as a discretization-based approach or an optimization-based approach.

1 Introduction

Numerical data arise prevalently in databases, including business and scientific databases. Handling numerical (or quantitative) data in data mining has attracted much attention since the work on mining quantitative association rules by Srikant and Agrawal [16]. Conventionally, *data discretization* is commonly used to handle numerical data; for a quantitative attribute which can have continuous values, it reduces the number of values by dividing the range of the attribute into intervals. The other approaches to handling numerical data have also been proposed, including a statistical distribution-based approach and an optimization-based approach (see the survey in [14]).

Kaytoue *et al.* [8] proposed an FCA-based approach to handling quantitative attributes, and introduced the notions of *closed interval patterns (CIPs)* as well as *generators*. The notion of IPs is an instance of the general framework of *pattern structures* studied by Ganter and Kuznetsov [5]. Recently, some methods have been proposed to use CIPs for mining association rules [6,13]. In particular, the approach in [13] handles multi-relational data mining (MRDM); it uses CIPs for mining *relational* quantitative association rules of the form $A \to C$, where

[*] Current Affiliation: Fujitsu Limited, Japan.

© Dmitry I. Ignatov, Lhouari Nourine (Eds.): CLA 2018, pp. 47–58,
ISBN 978–80–244–5328–6, Department of Computer Science, Palacký University
Olomouc, 2018.

A and C are relational patterns (i.e., logical conjunctions), and they consist of categorical attributes and quantitative ones from a given multi-relational data. In both work, although closed interval patterns allow us to represent intervals concisely, the numbers of generated patterns (i.e., rules) are still large, which makes the computation expensive and imposes a significant burden on the user's understanding.

In this paper, we study the problem of mining optimal relational association rules that have the maximum χ^2 value between the assumption and the conclusion of the rule. To find such rules, we use the original *MinIntChange* algorithm by Kaytoue *et al.*, and incorporate into it a pruning technique by Morishita *et al.* [11]. Moreover, since the search space for computing closed IPs becomes larger as the number of numerical attributes increases, we utilize *Super CWC* [15], an off the shelf feature selection algorithm to reduce the number of attributes to use. We give some experimental results, which show the effectiveness of the proposed method.

The organization of the rest of this paper is as follows. We first summarize some basic notations and definitions of relational association rule mining and interval patterns in Sect. 2. We then explain our approach to mining quantitative association rules from multi-relational data in Sect. 3, and show some experimental results in Sect. 4. Finally, we give a summary of this work in Sect. 5.

2 Relational Association Rules with Quantitative Attributes

2.1 Relational Pattern Mining and Interval Patterns

We use some basic notions of MRDM in [3]. To represent data and patterns, we use a class of first-order logical formulas. An *atom* is an expression of the form $p(t_1, \ldots . t_n)$, where p is a *predicate* and each t_i is a *term* (i.e., a constant or a variable). A substitution $\theta = \{X_1/t_1, \ldots, X_n/t_n\}$ is an assignment of terms to variables. The result of applying a substitution θ to a formula (i.e., an atom or a conjunction in this case) F is the formula $F\theta$, where all occurrences of variables V_i have been simultaneously replaced by the corresponding terms t_i in θ. The set of variables occurring in a formula F is denoted by $Var(F)$. A *pattern* is expressed as a conjunction $l_1 \wedge \cdots \wedge l_n$ of atoms, denoted simply by $l_1, \ldots, l_n$.

A database DB is a set of ground atoms. For a pattern C, let $answerset(C; DB)$ be the set of substitutions θ such that $C\theta$ is logically entailed by a database DB, denoted by $DB \models C\theta$.

In MRDM, we often specify one of the predicates as a *key* (e.g., [2,1]), which determines the entities of interest and what is to be counted. The key (target) is thus to be present in all patterns considered. Given a database DB and a conjunction C containing a key atom $key(X)$, the *support* (or *frequency*) of C, denoted by $supp(C)$, is defined to be the number of different keys that answer C divided by the total number of keys. C is said to be *frequent*, if $supp(C)$ is no less than some user defined threshold *minsup*.

$\mathit{customerId}(c_1)$.	$\mathit{marriedTo}(c_1, c_2)$.	$\mathit{income}(c_1, 200)$.
$\mathit{customerId}(c_2)$.	$\mathit{marriedTo}(c_2, c_1)$.	$\mathit{income}(c_2, 120)$.
$\mathit{customerId}(c_3)$.	$\mathit{marriedTo}(c_3, c_4)$.	$\mathit{income}(c_3, 50)$.
...		
$\mathit{age}(c_1, 30)$.	$\mathit{bigSpender}(c_1)$.	
$\mathit{age}(c_2, 25)$.	$\mathit{bigSpender}(c_2)$.	
$\mathit{age}(c_3, 55)$.		
...		

Fig. 1. An Example Database *DB* with *customerId* as a key: adapted from [3].

An association rule we consider in this paper is an existentially quantified implication of the form: $A \rightarrow C$, where A (resp., C) is a conjunction of the form: $a_1, \ldots, a_m$ (resp., a single atom) ($m \geq 1$). We call A (C) the *antecedent* (*conclusion*) of the rule, respectively. The *support* of a (relational) association rule is defined as the support of $A \wedge C$, while the *confidence* of an association rule is defined as the support of C divided by the support of the antecedent A. Following [7], we call a rule *strong*, if it satisfies both a minimum support threshold (*minsup*) and a minimum confidence (*minconf*).

Example 1. Consider a toy example of a multi-relational database *DB* in Fig. 1, which is adapted and simplified from [3]. Predicate *customerId* is assumed to be a key. Let P be a pattern of the form: $\mathit{customerId}(X)$, $\mathit{age}(X, Q_1)$, $\mathit{marriedTo}(X, Y)$, $\mathit{income}(Y, Q_2)$, whose meaning is obvious. Then, $\mathit{answerset}(P; DB)$ contains substitutions $\{X/c_1, Q_1/30, Y/c_2, Q_2/120\}$ and $\{X/c_2, Q_1/25, Y/c_1, Q_2/200\}$, for example.

The following rule is an example of association rules:

$$\mathit{customerId}(X), \mathit{age}(X, Q_1), \mathit{marriedTo}(X, Y), \mathit{income}(Y, Q_2) \rightarrow \mathit{bigSpender}(X). \tag{1}$$

In the above, Q_1 and Q_2 are quantitative attributes, while the others are considered to be categorical ones. □

We call a variable corresponding to a quantitative (resp., categorical) attribute a *quantitative* (resp., *categorical*) *variable*. We also call a variable occurring in a key predicate a *key variable*.

Relational Association Rules with Interval Patterns We use interval patterns to specify constraints on quantitative variables in an association rule. In the aforementioned association rule (1), for example, we consider the following association rule with constraints consisting of interval patterns:

$$\begin{aligned} &\mathit{customerId}(X), \mathit{age}(X, Q_1), \mathit{marriedTo}(X, Y), \mathit{income}(Y, Q_2), \\ &\quad \langle Q_1, Q_2 \rangle \in \langle [l_1, u_1], [l_2, u_2] \rangle \rightarrow \mathit{bigSpender}(X). \end{aligned} \tag{2}$$

where l_i (u_i) is a value of the domain of the attribute Q_i ($i = 1, 2$), respectively.

Formally, let A be a conjunction such that A contains quantitative variables $Q_1, \ldots, Q_k$ ($k \geq 1$). Then, we call an expression $\boldsymbol{c}$ of the form "$\langle Q_1, \ldots, Q_k \rangle \in \langle I_1, \ldots, I_k \rangle$" an *interval constraint of* A, where I_i ($1 \leq i \leq k$) is an interval pattern for Q_i.

Let θ be a substitution for $\mathit{Var}(A)$ and $I_i = [e_i, f_i]$ for some e_i and f_i. Then, $\mathit{DB} \models (A, \boldsymbol{c})\theta$ iff $\mathit{DB} \models A\theta$ and $Q_i\theta \in [e_i, f_i]$ for $i = 1, \ldots, k$.

For simplicity, we write simply "A, I" instead of $A, \langle Q_1, \ldots, Q_k \rangle \in \langle I_1, \ldots, I_k \rangle$, where $I = \langle I_1, \ldots, I_k \rangle$ and we call I an *interval pattern* of A.

For a conjunction A which has no categorical variables except a key variable, we can define a *closed pattern* in the same way as [8]; let $S = \{\theta_1, \cdots, \theta_n\}$ ($n \geq 0$) be a set of substitutions for variables in A and let $Q_1, \ldots, Q_k$ be quantitative variables in A. Then, we define a mapping $\delta(\cdot)$ as follows: for a substitution $\theta \in S$ such that $\theta \supseteq \{Q_1/a_1, \ldots, Q_k/a_k\}$, $\delta(\theta) = \langle [a_1, a_1], \ldots, [a_k, a_k] \rangle$. Namely, the mapping δ maps θ into an k-dimensional interval pattern $\langle [a_1, a_1], \ldots, [a_k, a_k] \rangle$.

Definition 1 (closed pattern).

For a conjunction A such that A has no categorical variables except a key variable, let $S = \{\theta_1, \ldots, \theta_n\}$ ($n \geq 0$) be a set of substitutions for variables in A, and I an interval pattern of A. We consider the following two operators $(\cdot)^\square$:

$I^\square = \{\theta \mid \theta \in \mathit{answerset}(A, I; \mathit{DB})\}$,

$S^\square = \langle \delta(\theta_1) \sqcap \cdots \sqcap \delta(\theta_n) \rangle$.[1].

Let I and J be an interval pattern of A. Then, I and J are *equivalent* if $I^\square = J^\square$ and we write it by $I \equiv J$. We call I *closed* if there does not exist any other interval pattern J such that $I \equiv J$ and $I \sqsubseteq J$. □

2.2 Correlation Measures

Since the framework using the support/confidence only generates too many rules, we usually use another measure to find "interesting" ones among the generated rules. χ^2-value is such a measure to find correlated rules; it is defined as a normalized derivation of observation from expectation. Given a contingency table in Table 1, where given m and n are both assumed to be constants, χ^2 values are determined by x and y, and we thus denote it by $\chi^2(x, y)$. The following property of χ^2-value is shown by Morishita *et al.* [11].

Lemma 1 (Morishita *et al.*). *[11]* Let $r(I_0)$ be a rule with an interval pattern I_0, and let a (b, resp.) be the number of (positive) tuples that satisfy the antecedent of $r(I_0)$. Let $r(I)$ be a rule with an interval pattern I, and let p (q, resp.) be the number of (positive) tuples that satisfy the antecedent of $r(I)$ such that $0 \leq p \leq a$, $0 \leq q \leq b$, $q \leq p$ and $(p - q) \leq (a - b)$. Then, we have

$$\chi^2(p, q) \leq \max\{\chi^2(b, b), \chi^2(a - b, 0)\}. \tag{3}$$

□

[1] For $I_1 = \langle [a_i, b_i] \rangle_{i \in \{1,;k\}}$ and $I_2 = \langle [e_i, f_i] \rangle_{i \in \{1,;k\}}$, $\sqcap$ is the infimum operator defined by $I_1 \sqcap I_2 = \langle [\min(a_i, e_i), \max(b_i, f_i)] \rangle_{i \in \{1,\ldots,k\}}$, and $I_2 \sqsubseteq I_1 \iff [e_i, f_i] \subseteq [a_i, b_i], \forall i \in \{1, \ldots, k\}$.

The right-hand side of (3) gives an upper bound of $\chi^2(p,q)$, and we thus denote it by $ub(r(I))$.

Table 1. Contingency Table for Rule r=$A \rightarrow C$.

	C is true	C is false	Sum$_{row}$
A is true	$sup(r)$=y	$x-y$	$sup(A)$=x
A is false	$m-y$	$n-x-(m-y)$	$sup(\neg A)$= $n-x$
Sum$_{col}$	$sup(C)$=m	$sup(\neg C)$= $n-m$	n

QuantMiner [14], a GA-based algorithm, searches rules with high *fitness function* rules. The fitness function Fitness($\cdot$) is an evaluation measure for a rule, and it is based on the *Gain* measure proposed in [4]: $Gain(A \rightarrow C) = supp(A \wedge C) - \text{min\_conf} \cdot supp(A)$. The *Gain* value is a measure giving a trade-off between support and confidence. Using x and y in Table 1, we write $Gain(A \rightarrow B) = G(x,y)$, and $G(x,y)$ is also a convex function.

3 Mining Quantitative ARs with IPs from a MRDB

Algorithm 1 shows the outline of our algorithm for mining correlated ARs with interval patterns from a MRDB.

Given a MRDB DB, the user first specifies a *rule template* of the form: $A \rightarrow C$; it specifies conjunctions occurring in the left-hand side and the right-hand side, and the right-hand side contains a single target atom C. The user also specifies values of categorical variables occurring in A and C. In case the values of the categorical variables in the rule template are not given, its possible values will be computed in the algorithm so that each of the categorical variable is instantiated to some value in its domain.

Next, we compute the answersets of A and $A \wedge C$, and we make the *initial association rule* r_{init} of the form: $A, I^{\perp} \rightarrow C$, where $I^{\perp}$ is the *minimal* interval constraint of A. In the aforementioned rule (2), for example, $I^{\perp} = \langle [l_1, u_1], [l_2, u_2] \rangle$, where l_i (u_i) ($i = 1, 2$) is the minimum (maximum) value of the domain of the attribute Q_i, respectively.

If $r_{init}$is infrequent (i.e., its support $supp(A \wedge C)$ is less than *minsupp*), then exit. Otherwise, we compute the set $\mathcal{R}$ of strong rules with the best χ^2 value on DB, by calling a function MIC$^{+p(\chi^2)}(r_{init}, 0)$.

Algorithm 1: Correlated AR Mining from a MRDB

```
   input  : a MRDB $DB$, $minsupp$, $minconf$.
   output: a set $\mathcal{R}$ of rules $r_b$ with the best $\chi^2$ value on $DB$.
 1 A rule template of the form: $A \to C$ is specified by the user;  // initial
    step
 2 Compute answersets $answerset(A; DB)$ and $answerset(A \wedge C; DB)$;
    // mining step for categorical attributes
 3 Make an initial association rule $r_{init}$ of the form: $A, I^{\perp} \to C$;
 4 if $A \wedge C$ is infrequent then return;
 5 Initialize $\mathcal{R} \leftarrow \emptyset$; $\tau \leftarrow -\infty$, and compute correlated ARs by calling
    $\mathrm{MIC}^{+p(\chi^2)}(r_{init}, 0)$ ;
 6 return $\mathcal{R}$
 7 Function $\mathrm{MIC}^{+p(\chi^2)}$(a rule $r(I)$: $A, I \to C$, an integer $j$) : a set $\mathcal{R}$ of
    rules with the best $\chi^2$ value on Database $DB$ is
 8     $\mathcal{A} \leftarrow \{\mu_{i,\alpha} \mid \mu_{i,\alpha}$ is applicable to $I$ for some $i \geq j,\ \alpha \in \{l, r\}\}$;
 9     foreach $\mu_{i,\alpha} \in \mathcal{A}$ do
10         $I' \leftarrow \mu_{i,\alpha}(I)$;
11         if $sup(r(I')) < minsup$ or $ub(r(I')) < \tau$ then continue
12         $I_1 \leftarrow I'^{\square\square}$;
13         if $I_1$ fails the canonicity test then continue
14         $\tau_1 \leftarrow \chi^2$ value of $r(I_1)$;
15         if $\tau_1 > \tau$ and $conf(r(I')) \geq minconf$ then $\tau \leftarrow \tau_1$; $\mathcal{R} \leftarrow \{r(I_1)\}$
16         else if $\tau_1 = \tau$ and $conf(r(I')) \geq minconf$ then $\mathcal{R} \leftarrow \mathcal{R} \cup \{r(I_1)\}$;
17         call $\mathrm{MIC}^{+p(\chi^2)}(r(I_1), i)$;
18     end
19 end
```

The function $\mathrm{MIC}^{+p(\chi^2)}(r(I), j)$ is essentially the same as the MinIntChange algorithm by Kaytoue *et al.* [8]; the enumeration of closed IPs is done in the same way as the original MinIntChange. Namely, the algorithm generates its direct subsumers whose supports are strictly lower than its support. New interval patterns are generated by applying *minimal changes* to a given interval pattern (line 10). Since a closed interval pattern may be generated several times, we employ the *canonicity test* due to CloseByOne [10] (line 13).

Definition 2 (minimal change). [8]

Let I be an interval pattern of a conjunction A, where $I = \langle [a_1, b_1], \ldots, [a_k, b_k] \rangle$ for some $k \geq 1$.

A *right minimal change* $\mu_{i,r}(I)$ $(1 \leq i \leq k)$ is defined as I', where I' is I with its i-th interval replaced by $[a_i, v]$ such that $v = \max\{x \in V_i \mid x < b_i\}$ and V_i is the set of values which the quantitative variable Q_i will take in a given database. A *left minimal change* $\mu_{i,l}(I)$ is defined dually.

A right minimal change $\mu_{i,r}$ is *applicable* to I if the resulting interval $[a_i, v]$ does not collapse, i.e., $v - a_i > 0$. An *applicable* left minimal change $\mu_{i,l}(I)$ is defined dually. □

MIC$^{+p(\chi^2)}$ incorporates into the original MinIntChange a pruning mechanism (line 11) based on Lemma 1 and a mechanism storing rules with currently best χ^2 value (line 15–16). We then have the following properties:

Theorem 1 (Correctness of Algorithm 1). Let *minsup* be a given minimum support and *minconf* a given minimum confidence. Let DB be a given database. Then,

[**Soundness**] All output rules in $\mathcal{R}$ of Algorithm 1 give the best χ^2 value on DB, and satisfy both *minsup* and *minconf*.

[**Completeness**] Let $r(I)$ be an association rule of the form: $A, I \rightarrow C$ for some interval I. Then, if r gives the best χ^2 value on DB and satisfies *minsup* and *minconf*, then Algorithm 1 outputs a rule r_1 in $\mathcal{R}$ of the form: $A, I_1 \rightarrow C$ such that $I_1 = I^{\square\square}$.

Proof. Since the soundness is rather obvious, we omit its proof.

For the completeness, we have from the assumption and the completeness of MinIntChange that there exists a sequence s of minimal changes from the root to $I_1 = (I)^{\square\square}$ such that I_1 passes the canonicity test, i.e., it is not pruned in line 13. Furthermore, since the closure operator $I'^{\square\square}$ (line 12) does not change its support, the computation corresponding to s is not pranced in line 11, either. Therefore, we have that $r(I_1) \in \mathcal{R}$. □

We note that Algorithm 1 is generic in a sense that it works for another correlation measure, m, by replacing MIC$^{+p(\chi^2)}$ by MIC$^{+p(m)}$, provided that the correlation measure m has a property such as convexity so that it allows us to compute an upper bound $ub(r(I))$ (line 11). Such correlation measures include *information gain*, *gini index* and *Gain*, to mention a few.

Although the proposed pruning method makes the IP search space smaller, the search space becomes larger, when a given database has many quantitative attributes. To handle such cases, we utilize *Super CWC* [15], an off the shelf feature selection algorithm to reduce the number of attributes to use. We will show some experimental results in the following section.

4 Experimental Results

We show in Table 2 some datasets used in our experiments; one is from the UCI Machine Learning [2] and the others are from the CTU Prague Relational Learning Repository [12][3]. We also show in Table 3 rule templates for those datasets used to compute correlated association rules, where Q_i $(i = 1, 2)$ are quantitative variables, while the other variables are categorical ones.

Table 2. Example Databases: † from the UCI Machine Learning Repository and the others from the CTU Prague Repository [12]. #Relations: the number of tables in the database. #Instances: the number of rows in the target table. Size: size in MB.

Database	#Relations	#Instances	Size (MB)	Domain
Mutagenesis	3	188	0.9	Medicine
Financial	8	682	94.1	Finance
Mondial	33	454	3.3	Geography
Heart†	1	270	0.016	Medicine

Table 3. Rule Templates for the Datasets in Table 2.

Database	Rule Template
Mutagenesis	$mol\_Id(X), ind\_1(A, C), eLumo(X, Q_1), logP(X, Q_2) \to active(X)$
Financial	$loan(X), amount(X, Q_1), duration(X, 60), avg\_salary(X, Y, Q_2) \to status(X, C)$
Mondial	$country(X), continent(X, Europe), agri(X, Q_1), serv(X, Q_2) \to christian(X)$
heart	$id(X), sc(X, Q_1), max\_hra(X, Q_2), cp\_t(X, T) \to disease(X, C)$

We have implemented our proposed method by using Java 8 on a PC with an Intel Core i7 processor running at 2.30GHz, 8GB of main memory, working under Windows 7 (64 bit). We have performed the following experiments varying the thresholds *min_sup* at fixed $min\_conf = 0.6$.

Effects of the Pruning Method To see the effects of the pruning method, we present some results for the two datasets (mutagenesis and financial) in Figure 2. The figures (left) show the numbers of strong rules generated in computing correlated rules for the rule templates in Table 3, where those categorical attributes in each rule template take some values in their domains. The figures (right) show the corresponding execution times in milliseconds.

We have observed that the pruning method based on the branch-and-bound heuristics enables us to generate much less rules compared with the naive approach (i.e., without pruning). The execution time of both cases are also reduced accordingly.

Effects of the Number of Quantitative Attributes Next, to see the effects of the number of quantitative variables in mining correlated rules, we consider two more rule templates from the rule template, R_2, for the Mondial dataset in Table 3, by varying the number of quantitative variables from 2 to 4; namely, one is a rule template R_3 obtained by adding $industry(X, Q_3)$ to the antecedent of R_2, while the other rule template, R_4, is obtained similarly by adding $inflation(X, Q_4)$ to the antecedent of R_3.

[2] http://archive.ics.uci.edu/ml/datasets/statlog+(heart).

[3] https://relational.fit.cvut.cz/.

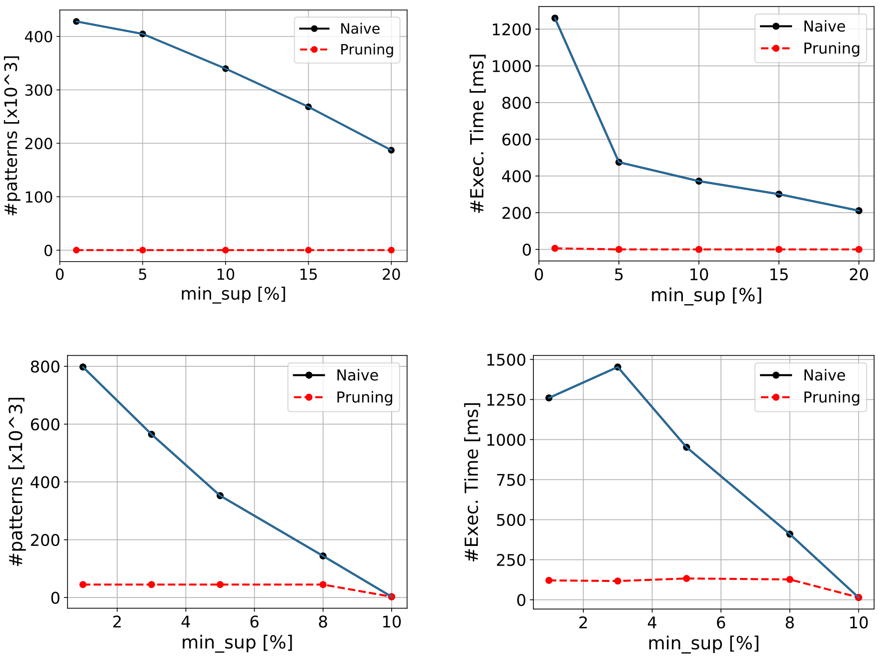

Fig. 2. #(Generated Strong Rules) and Execution Time for Computing Correlated Rules for the Mutagenesis (above) and the Financial (below) Dataset. Rule templates in Table 3 are used.

The number of generated strong rules and the execution time of computing the association rules are shown in Fig. 3. The number of different values in the domain of each quantitative variable Q_i are shown in Table 4. The numbers of possible interval patternss made from the values of Q_1, Q_2, Q_3 and Q_4 thus could become very large. However, the figure shows that the numbers of generated rules using the pruning method only moderately increase. The pruning method thus works well also in this case.

Table 4. Some Statistics of the Mondial Dataset in Table 2. #Values: # of different values in dom(Q_i) ($1 \leq i \leq 4$).

	Q_1 (agri.)	Q_2 (service)	Q_3 (industry)	Q_4 (inflation)
#Values	27	29	30	29

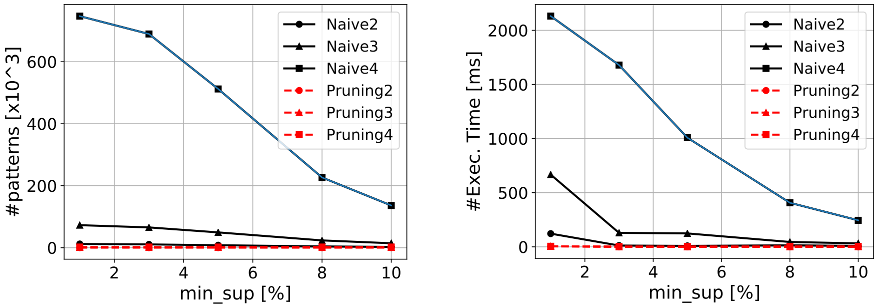

Fig. 3. #(Generated Strong Rules) and Execution Time for Computing Correlated Rules for the Mondial Dataset. *N-i* (resp., *P-i*): the naive (resp., pruning) method for rule R_i with i quantitative attributes ($i = 2, 3, 4$).

The heart dataset in Table 2 contains 13 attributes; among them, 6 attributes take numerical values. In this case, the MinIntChange algorithm generates a large number of CIPs. In fact, we could not obtain outputs of our algorithm within a reasonable time when applying it directly to the complete dataset. We alleviate this problem by first choosing some of the numerical attributes from the dataset by using *Super CWC* mentioned in Sect. 3, and then applying our algorithm to this reduced dataset. Figure 4 shows the numbers of strong rules generated and the execution time in computing correlated rules for an initial association rule.

Table 5 shows some rules with the best χ^2 values obtained by our algorithm as well as the other approaches, i.e., *QuantMiner* [14], an optimization-based approach, and CAIM [9], a data discretization approach. We notice that the rule obtained by our algorithm has the best χ^2 value in this case.

Table 5. Some Obtained Rules: An Example of the Heart Dataset. The Initial Association Rule: $id(X), sc(X, Q_1), max\_hra(X, Q_2), cp\_t(X, 4), \langle Q_1, Q_2\rangle \in \langle I_1, I_2\rangle \rightarrow disease(X, 2)$. $minsupp = 0.1$, $minconf = 0.6$.

	Interval Patterns $\langle I_1, I_2\rangle$	$(supp, conf)$	χ^2 value
Closed IP	$\langle [164.0, 409.0], [71.0, 177.0]\rangle$	(0.33, 0.75)	**81.7**
QuantMiner	$\langle [234.0, 326.0], [122.0, 147.0]\rangle$	(0.09, 0.85)	20.2
CAIM	$\langle [126.0, 407.0], [71.0, 195.0]\rangle$	(0.33, 0.70)	66.0

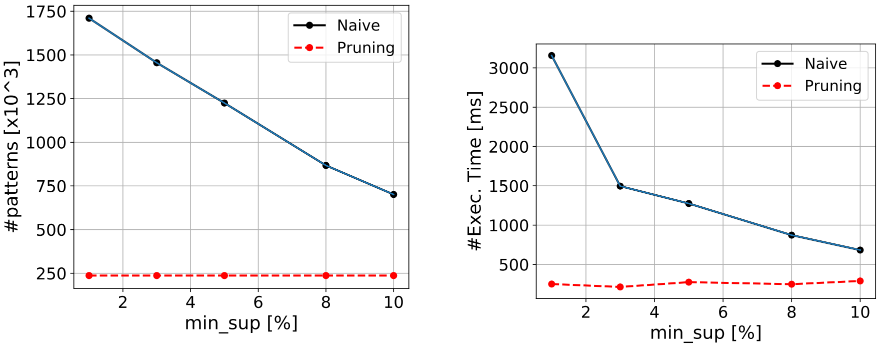

Fig. 4. #(Generated Strong Rules) and Execution Time for Computing Correlated Rules for the Heart Dataset with Initial Association Rule: $id(X)$, $sc(X, Q_1)$, $max\_hra(X, Q_2)$, $cp\_t(X, 4), \langle Q_1, Q_2 \rangle \in \langle I_1, I_2 \rangle \rightarrow disease(X, 2)$. $minsupp = 0.1$, $minconf = 0.6$.

5 Concluding Remarks

In this paper, we have considered the problem of mining relational association rules, especially focusing on the use of closed interval patterns (CIPs) for finding correlated rules with the best χ^2 values. Since the number of mined CIPs increases as the number of attributes and values in the domain of each attribute increases, we have examined the effectiveness of the original MinIntChange algorithm [8] with the pruning technique by Morishita *et al.* on the problem. We have also examined the effectiveness of the use of a feature selection algorithm, *Super CWC*, to reduce the search space of IPs.

Most of the work in the field of MRDM have handled numerical data by using data discretization. To the best of our knowledge, there has been no approach which uses closed interval patterns for mining correlated association rules in multi-relational data.

For future work, we will examine another correlation measure m in $\mathrm{MIC}^{+p(m)}$, since our algorithm is generic and will work for another measure. Since the search space for CIPs is still large in general and the computation of CIPs is costly, we will need some method for reducing the computational time and space to a manageable size.

Acknowledgment The authors would like to thank anonymous reviewers for their useful comments on the previous version of the paper. This work was partially supported by JSPS Grant-in-Aid for Scientific Research (C) 15K00305 and 18K11432.

References

1. De Raedt, L., Ramon, J.: Condensed representations for Inductive Logic Programming. In: Proc. KR'04, pp. 438–446 (2004)
2. Dehaspe, L.: Frequent pattern discovery in first-order logic. Ph.D. thesis, Dept. Computer Science, Katholieke Universiteit Leuven (1998)
3. Dzeroski, S.: Multi-Relational Data Mining: An Introduction. SIGKDD Explorations Newsletter 5(1), 1–16 (2003)
4. Fukuda, T., Morimoto, Y., Morishita, S., Tokuyama, T.: Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. In: Proc. of ACM SIGMOD '96. pp. 13–23 (1996)
5. Ganter, B., Kuznetsov, S.: Pattern Structures and Their Projections. In: ICCS-01, LNCS, vol. 2120, pp. 129–142. Springer (2001)
6. Guyet, T., Quiniou, R., Masson, V.: Mining relevant interval rules. In: Supplementary Proceedings of ICFCA. pp. 79–82 (2017)
7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2006)
8. Kaytoue, M., Kuznetsov, S.O., Napoli, A.: Revisiting numerical pattern mining with formal concept analysis. In: International Conference on Artificial Intelligence (IJCAI 2011). pp. 1342–1347 (2011)
9. Kurgan, L.A., Cios, K.J.: CAIM discretization algorithm. IEEE Transactions on Knowledge and Data Engineering 16(2), 145–153 (Feb 2004)
10. Kuznetsov, S.O., Obiedkov, S.A.: Comparing performance of algorithms for generating concept lattices. J. Exp. Theor. Artif. Intell. 14(2-3), 189–216 (2002)
11. Morishita, S., Nakaya, A.: Parallel branch-and-bound graph search for correlated association rules. In: Proc. of ACM SIGKDD Workshop on Large-Scale Parallel KDD Systems. pp. 265–276 (1999)
12. Motl, J., Schulte, O.: The CTU Prague Relational Learning Repository. ArXiv e-prints (Nov 2015)
13. Nagao, M., Seki, H.: On mining quantitative association rules from multi-relational data with FCA. In: Proc. IEEE IWCIA2016. pp. 81–86 (2016)
14. Salleb-Aouissi, A., Vrain, C., Nortet, C.: QuantMiner: A genetic algorithm for mining quantitative association rules. In: IJCAI. pp. 1035–1040 (2007)
15. Shin, K., Kuboyama, T., Hashimoto, T., Shepard, D.: Super-cwc and super-lcc: Super fast feature selection algorithms. In: proceedings of Big Data. pp. 61–67 (2015)
16. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Proc. of ACM SIGMOD '96. pp. 1–12 (1996)