International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 3, June 2017, pp. 1385~1397
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i3.pp1385-1397

Approximation Measures for Conditional Functional Dependencies Using Stripped Conditional Partitions

Anh Duy Tran[1], Somjit Arch-int[2], Ngamnij Arch-int[3]
[1,2,3] Department of Computer Science, Faculty of Science, Khon Kaen University

Article Info

Article history:

Received Apr 11, 2017
Revised May 5, 2017
Accepted May 24, 2017

Keyword:

Approximate conditional functional dependencies
Measures
Stripped conditional partitions
Incomplete knowledge granularity

ABSTRACT

Conditional functional dependencies (CFDs) have been used to improve the quality of data, including detecting and repairing data inconsistencies. Approximation measures have significant importance for data dependencies in data mining. To adapt to exceptions in real data, the measures are used to relax the strictness of CFDs for more generalized dependencies, called approximate conditional functional dependencies (ACFDs). This paper analyzes the weaknesses of dependency degree, confidence and conviction measures for general CFDs (constant and variable CFDs). A new measure for general CFDs based on incomplete knowledge granularity is proposed to measure the approximation of these dependencies as well as the distribution of data tuples into the conditional equivalence classes. Finally, the effectiveness of stripped conditional partitions and this new measure are evaluated on synthetic and real data sets. These results are important to the study of theory of approximation dependencies and improvement of discovery algorithms of CFDs and ACFDs.

Corresponding Author:

Anh Duy Tran,
Department of Computer Science, Faculty of Science,
Khon Kaen University,
123 Moo 16 Mittapap Road., Nai-Muang, Muang District, Khon Kaen 40002,Thailand
Email: duyanh208@gmail.com

1. INTRODUCTION

High data quality has a very important role for many organizations in making correct decisions. However, in real-world applications, the data often contains inconsistencies, inaccuracies and errors because of integration of data from various sources. A recent report showed that billions of dollars in losses annually for US business is due to poor data quality [1]. Although functional dependencies (FDs) are significant constraints and knowledge in relational database design and data mining [2-6], they are not robust enough to address data quality problem. Therefore, CFDs have been extended from FDs to solve this problem [7-8]. FDs only hold on a set of tuples satisfying the conditions characterized by CFDs. For example, let *cust* be a relation specifying customers with the attributes: CC (country code), ZIP (zip code), STR (street), AC (area code), CT (city) as introduced in [7][9]. Let's consider two CFDs: $\varphi_1 = ([CC, ZIP] \rightarrow STR, (44, - \| -))$ and $\varphi_2 = ([CC, AC] \rightarrow CT, (01, 212 \| NYC))$. CFD φ_1 only holds on the relation *cust* when the customer's country code is 44. CFD φ_2 shows that if all customers in the US (CC=01) have an area code of 212, then their city must be NYC. These constraints cannot be discovered from the databases using the concept of FD.

The main application of CFDs is data cleaning [7][10][11] in which CFD discovery is an important stage. The measures have been used to discover the interesting rules, reduce search space and relax strictness of CFDs with exceptions in data [9][12-13]. Chiang et al [12] introduced the various measures to evaluate the data quality rules, including Support, χ2-Test, Confidence, Interest and Conviction. Based on the subsumed classes in the partitions, these measures captured interesting CFDs such that there exist the conditional attributes on the left hand side (LHS) of these CFDs. The non-subsumed classes are used to formalize the approximate constant CFDs for identifying dirty data values. Therefore, it seems difficult to approximate the general CFDs based on these measures.

The interesting rules in the discovery problem of constant CFDs [14] were also evaluated based on the *χ2-Test*. Moreover the conviction is an effective measure for association rules because it tackles the weaknesses of the confidence and interest measures [15]. As shown in [12], the conviction is the best measure for providing the interesting CFDs and identifying the dirty data values. Therefore, the conviction measure will be selected for analysis in this study.

Recently, Nakayama et al [13] presented the formalization of ACFDs with the confidence measure based on the maximum number of tuples in a relation satisfying the conditional dependency. This measure is extended from the error measure g_3[16], which has been used widely in the study, discovery and application of approximate functional dependencies (AFDs) and comparable dependencies (CDs) [17-21]. Nakayama et al focused on extending three discovery algorithms for ACFDs (approxCFDMiner, approxCTANE and approxFastCFD) from CFD discovery algorithms [9].

Unfortunately the effectiveness of stripped conditional partitions and evaluation of this measure for ACFDs were not considered. Therefore we introduce the conditional indiscernibility relation, conditional equivalence class, conditional partition, stripped conditional partition and dependency degree γ as an extension from the concepts of Pawlak rough set [22-23] to confront this problem. Rough set theory is an effective approach for analyzing uncertain and incomplete data in many areas of data mining, knowledge discovery and attribute reduction [22-28]. In addition, information (knowledge) granularity can be used to measure uncertainty of information [29-33].

This study also infers that the measurement of ACFDs allows us to know the distribution degree of objects in the conditional equivalence classes. For example, we can represent how much degree patients corresponding to any symptom are distributed into disease groups. However the above measures cannot express this distribution. We therefore introduce the incomplete knowledge granularity of conditional partition induced by itemsets based on the knowledge granularity of the partition [33] to propose a new measure that not only measures the approximation degree of dependencies CFDs, but also the distribution of data tuples into the conditional equivalence classes. This measure can give us a more general view of ACFDs with expectation for extending ACFDs to other application domains such as classification and sociological investigation. Finally the computations of measures using the stripped conditional partitions allow the discovery time of CFDs and ACFDs to be improved effectively.

From this promising analysis, the paper focuses on solving the following issues:

- Computing the measures based on the conditional partitions and stripped conditional partitions.
- Evaluating the effectiveness of the stripped conditional partition for discovery algorithm of ACFDs (approxCTANE) based on the confidence measure.
- Evaluating the limitations of the measures for ACFDs, including the dependency degree, confidence and conviction.
- Proposing a new measure for CFDs and evaluating the utility of this measure.

The rest of the paper is organized as follows: Section 2 presents primary concepts of partition, dependency degree and conditional functional dependencies. In section 3, we compute the measures and propose a new measure for CFDs based on conditional partition and stripped conditional partition as well as analyze among the measures. Section 4 introduces the discovery problem of ACFDs and product of two stripped conditional partitions. The evaluation of measures and discovery of ACFDs are conduced on the synthetic and real data sets in Section 5. Section 6 concludes the paper.

2. PRELIMINARIES

In this section, we introduce some concepts of relational database, indiscernibility relation, partitions, dependency degree and conditional functional dependencies [7-9][22-23][34].

Let r(R) be a relation on a set of attributes $R = \{A_1, A_2, ..., A_m\}$, where $dom(A_k)$ is a domain of $A_k \in R$. Then an indiscernibility relation I_X is defined by

$$I_X = \{(t_i, t_j) \in r^2 \mid \forall A_k \in X, t_i[A_k] = t_j[A_k]\}$$

The relation I_X partitions the set of tuples r into equivalence classes. For $t_i \in r$, an equivalence class $I_X^o(t_i)$ of t_i on a set of attributes X is defined by

$$I_X^o(t_i) = \{t_j \in r \mid \forall A_k \in X, t_i[A_k] = t_j[A_k]\}$$

Then a set of equivalence classes $I_X^o = \{I_X^o(t_i) \mid t_i \in r\}$ is called a partition of *r* on *X*. The partition I_X^o is finer than I_Y^o if and only if for any equivalence class $w \in I_X^o$, there exists the equivalence class *q* in I_Y^o such that $w \subseteq q$. To reduce computational time with the partitions, Huhtala et al. [17] have proposed the stripped partition of I_X^o, denoted $\hat{I}_X^o$ as follows

$$\hat{I}_X^o = \{w \in I_X^o \mid |w| > 1\}$$

A set of tuples $O \subseteq r$ can be approximated with respect to $X \subseteq R$ by defining the lower $X_*(O)$ approximation, where $X_*(O) = \{t_i \in r \mid I_X^o(t_i) \subseteq O\}$. Then for two sets of attributes *X* and *Y*, $POS(X \to Y, r) = \cup_{O \in I_Y^o} X_*(O)$ is a positive region of dependency $X \to Y$.

The coefficient $\gamma(X \to Y, r) = \dfrac{POS(X \to Y, r)}{|r|}$ defines a dependency degree of $X \to Y$. If $\gamma(X \to Y, r) = 1$, then $X \to Y$ is a functional dependency.

Functional Dependencies (FDs) express the relationship between two sets of attributes on the relation *r(R)*. In more generalized form, CFDs specify the constraints between two sets of attributes fixed by particular values. A Conditional Functional Dependency (CFD) φ, denoted $(X \to Y, t_p)$ [9], is a Functional Dependency with respect to pattern tuple t_p such that for each $A_k \in XY$, $t_p[A_k]$ = *'a' or* $t_p[A_k]$ = *'-'*, where the constant $a \in dom(A_k)$ and value of unnamed variable *'-'* drawn from $dom(A_k)$. A tuple t_i matches the pattern tuple t_p on the set of attributes *X*, denoted $t_i[X] \leq t_p[X]$ if and only if for each $A_k \in X$, $(t_i[A_k] = t_p[A_k])$ or $(t_i[A_k] = 'a', t_p[A_k] = '-')$, $a \in dom(A_k)$. We write $t_i[X] << t_p[X]$ if $t_i[X] \leq t_p[X]$ but $t_p[X] \not\leq t_i[X]$.

The CFD $\varphi = (X \to Y, t_p)$ holds on *r* (or *r* satisfies φ, denoted $r \models \varphi$) if, for any $t_i, t_j \in r$ such that $t_i[X] = t_j[X] \leq t_p[X]$, then $t_i[Y] = t_j[Y] \leq t_p[Y]$.

As mentioned in [9][12] we can discover the CFDs of form $\varphi = (X \to A_k, t_p)$, where the single attribute A_k is in *R* and not in *X*. Let $\varphi = (X \to A_k, t_p)$ be a CFD and $X \neq \varnothing$. According to [9], $X^c \subseteq X$ is a set of attributes such that $t_p[X^c]$ is a constant pattern tuple. The remaining attributes $X^v = X - X^c$ is corresponding to variable pattern tuple such that $t_p[X^v] = (-, ..., -)$. Therefore, the CFD φ can be expressed by the form $\varphi = ([X^c, X^v] \to A_k, t_p)$.

3. APPROXIMATE MEASURES FOR CFDs

In this section, we present the conditional partition, stripped conditional partition, and approximate measures for dependencies. We first consider some measures for CFDs.

Definition 3.1. [9][12][35] Let *r(R)* be a relation, the support of CFD $\varphi = (X \to A_k, t_p)$ on *r*, denoted *sup(*φ*, r)*, is defined by

$$\sup(\varphi, r) = \frac{|r_{t_p}|}{|r|} \tag{1}$$

where $|r_{t_p}| = |r_{t_p[XA_k]}|$ is the number tuples in r matching t_p on set of attributes XA_k.

From the conviction measure in [12], Definition 3.2 defines this measure for general CFDs.

Definition 3.2. Let *r(R)* be a relation, the conviction measure of CFD $\varphi = (X \to A_k, t_p)$ on *r*, denoted *Conv(*φ*, r)*, is defined by

$$Conv(\varphi, r) = \frac{P((X, t_p[X])).P(\neg(A_k, t_p[A_k]))}{P((X, t_p[X]), \neg(A_k, t_p[A_k]))}$$

where probability $P(X,t_p[X])$ is equal to $sup(X, t_p[X])$ and if $(X, t_p[X])$ and $(A_k, t_p[A_k])$ are independent, then *Conv(*φ*, r)* is equal to 1.

Definition 3.3. [13] Let *r(R)* be a relation, the confidence of CFD $\varphi = (X \to A_k, t_p)$ on *r*, denoted *conf (*φ*, r)*, is defined by

$$conf(\varphi, r) = \frac{\max\{|r'| \mid r' \subseteq r, r' \models (X \to A_k, t_p)\}}{|r|} \tag{2}$$

Partitions of itemsets are introduced in [9] to check correctness of CFDs in the discovery problem. Let *r(R)* be a relation on *R* and *(Z, t*$t_p[Z]$*)* be an itemset, where $Z \subseteq R$. There exists a conditional indiscernibility relation $I_{(Z,t_p[Z])}$ on *r*, defined by

$$I_{(Z,t_p[Z])} = \{(t_i, t_j) \in r^2 \mid t_i[Z] = t_j[Z] \leq t_p[Z]\}$$

A conditional equivalence class of t_i with respect to itemset *(Z,* $t_p[Z]$*)*, denoted $I^o_{(Z,t_p[Z])}(t_i)$, is defined by $I^o_{(Z,t_p[Z])}(t_i) = \{t_j \in r \mid t_i[Z] = t_j[Z] \leq t_p[Z]\}$.Therefore there exists a conditional partition of r with respect to *(Z,* $t_p[Z]$*)*, defined by

$$I^o_{(Z,t_p[Z])} = \{I^o_{(Z,t_p[Z])}(t_i) \mid t_i \in r, \mid I^o_{(Z,t_p[Z])}(t_i) \mid \neq 0\} \tag{3}$$

Table 1. A data table

	A1	A2	A3
t_1	0	1	0
t_2	0	1	0
t_3	1	2	0
t_4	1	2	0
t_5	0	1	0
t_6	0	2	1
t_7	1	2	2
t_8	1	3	2
t_9	1	3	2
t_{10}	2	1	3
t_{11}	2	2	1

We have that $\cup_{t_i \in r} I^o_{(Z,t_p[Z])}(t_i)$, where $I^o_{(Z,t_p[Z])}(t_i) \neq \varnothing$, is not always equal to *r*. Therefore the conditional partition $I^o_{(Z,t_p[Z])}$ is a semi-partition.

For example, let *r(R)* be a relation as Table 1. Then

$I^o_{(A_1,A_{2,-,-})} = \{\{t_1,t_2,t_5\},\{t_6\},\{t_3,t_4,t_7\},\{t_8,t_9\},\{t_{10}\},\{t_{11}\}\}$; $I^o_{(A_1,A_{2,-,2})} = \{\{t_6\},\{t_3,t_4,t_7\},\{t_{11}\}\}$

$\cup_{w \in I^o_{(A_1,A_2,-,2)}} w = \{t_3,t_4,t_6,t_7,t_{11}\} \neq r$

Definition 3.4. Let $O \subseteq r$ be a set of tuples and let *(Z,* $t_p[Z]$*)* be an itemset such that $Z \subseteq R$, the lower $(Z,t_p[Z])_*(O)$ approximation of *O* is defined by

$$(Z,t_p[Z])_*(O) = \{t_i \in r \mid I^o_{(Z,t_p[Z])} \neq \varnothing \, and \, I^o_{(Z,t_p[Z])}(t_i) \subseteq O\} \tag{4}$$

Definition 3.5. Let *r(R)* be a relation, the dependency degree of CFD $\varphi = (X \rightarrow A_k, t_p)$ on *r*, denoted $\gamma(\varphi, r)$, is defined by

$$\gamma(\varphi, r) = 1 - \frac{|r_{t_p[X]}| - |POS(\varphi, r_{t_p[X]})|}{|r|} \tag{5}$$

where: $POS(\varphi, r_{t_p[X]}) = \cup_{q \in I^o_{(A_k,t_p[A_k])}} (X, t_p[X])_*(q)$ and $r_{t_p[X]} = \cup_{w \in I^o_{(X,t_p[X])}} w$

Proposition 3.1. Let *r(R)* be a relation, the support of CFD $\varphi = (X \rightarrow A_k, t_p)$ on *r* is computed by

$$\sup(\varphi, r) = \frac{|r_{t_p}|}{|r|} = \frac{|\cup_{w \in I^o_{(XA_k,t_p[XA_k])}} w|}{|r|} \tag{6}$$

Proof. we have $|r_{t_p}| = |r_{t_p[XA_k]}| = |\cup_{w \in I^o_{(XA_k,t_p[XA_k])}} w|$. Proposition 3.1 therefore can be inferred from Definition 3.1 □

Proposition 3.2. Let *r(R)* be a relation, the conviction measure of CFD $\varphi = (X \rightarrow A_k, t_p)$ on *r* is computed by

$$Conv(\varphi, r) = \frac{1}{|r|} . \frac{|r_{t_p[X]}| .(|r| - |r_{t_p[A_k]}|)}{|r_{t_p[X]}| - |r_{t_p[XA_k]}|} \tag{7}$$

where *Z=X* , *Z=A*$_k$*, or Z=XA*$_k$, $r_{t_p[Z]} = \cup_{w \in I^o_{(Z, t_p[Z])}} w$ and if *(X, t*$_p$*[X])* and *(A*$_k$*, t*$_p$*[A*$_k$*])* are independent, then *Conv(*φ*, r)* is equal to 1.

Proof. We have $P(\neg(A_k, t_p[A_k]))$ =1- $\sup((A_k, t_p[A_k])) = 1 - \frac{|r_{t_p[A_k]}|}{|r|}$ and

$$P((X, t_p[X]), \neg(A_k, t_p[A_k])) = \sup((X\ , t_p[X]))\ \text{-}\ \sup((XA_k, t_p[XA_k])) = \frac{1}{|r|} .(|r_{t_p[X]}| - |r_{t_p[XA_k]}|)$$

Therefore Proposition 3.2 can be proven from Definition 3.2. □

Proposition 3.3. Let *r(R)* be a relation and φ = *(X* → *A*$_k$*, t*$_p$*)*. Then the measure *Conf* is computed by

$$Conf(\varphi, r) = 1 - \frac{|r_{t_p[X]}| - \sum_{w \in I^o_{(X, t_p[X])}} \max\{|q| \mid q \in I^o_{(XA_k, t_p[XA_k])}, q \subseteq w\}}{|r|} \tag{8}$$

where

$$|r_{t_p[X]}| = |\cup_{w \in I^o_{(X, t_p[X])}} w| = \begin{cases} |w| \mid w \in I^o_{(X^c, t_p[X^c])} & if\ \ X^c \neq \emptyset \\ |r| & otherwise \end{cases} \tag{9}$$

Proof. From Equation 2, we obtain

$$Conf(\varphi, r) = \frac{\max\{|s| \mid s \subseteq r_{t_p[X]}, s \models \varphi\} + |r - r_{t_p[X]}|}{|r|}$$

$$= 1 - \frac{|r_{t_p[X]}| - \max\{|s| \mid s \subseteq r_{t_p[X]}, s \models \varphi\}}{|r|}$$

On the other hand, in a similar way to the computation of g$_3$ in [17], we have $\max\{|s| \mid s \subseteq r_{t_p[X]}, s \models \varphi\} = \sum_{w \in I^o_{(X, t_p[X])}} \max\{|q| \mid q \in I^o_{(XA_k, t_p[XA_k])}, q \subseteq w\}$. Proposition 3.3 is therefore proven. □

Example 3.1. Let's compute the measures S*up*, γ , *Conf*, and *Conv* for the dependencies φ_1 and φ_2 from Table 1:

φ_1 = *(A*$_1$ → *A*$_2$*, 0 || 1)* φ_2 = *(A*$_1$ → *A*$_2$*, - || -)*

Through φ_1 , φ_2 and Table 1, we can infer the conditional partitions as follows:

$I^o_{(A_1,0)} = \{\{t_1, t_2, t_5, t_6\}\}$ $\qquad$ $I^o_{(A_1,-)} = \{\{t_1, t_2, t_5, t_6\}, \{t_3, t_4, t_7, t_8, t_9\}, \{t_{10}, t_{11}\}\}$

$I^o_{(A_2,1)} = \{\{t_1, t_2, t_5, t_{10}\}\}$ $\qquad$ $I^o_{(A_2,-)} = \{\{t_1, t_2, t_5, t_{10}\}, \{t_3, t_4, t_6, t_7, t_{11}\}, \{t_8, t_9\}\}$

$I^o_{(A_1, A_2, 0, 1)} = \{\{t_1, t_2, t_5\ \}\}$ $\qquad$ $I^o_{(A_1, A_2, -, -)} = \{\{t_1, t_2, t_5\}, \{t_6\}, \{t_3, t_4, t_7\}, \{t_8, t_9\}, \{t_{10}\}, \{t_{11}\}\}$

From these conditional partitions and Equations 5-8, we have:

$Sup(\varphi_1, r) = 3/11$ $\qquad$ $Sup(\varphi_2, r) = 11/11$

$\gamma(\varphi_1, r) = 1 - \frac{4-0}{11} = 7/11$ $\qquad$ $\gamma(\varphi_2, r) = 1 - \frac{11-0}{11} = 0$

$Conf(\varphi_1, r) = 1 - \frac{4-3}{11} = 10/11$ $\qquad$ $conf(\varphi_2, r) = 1 - \frac{11-(3+3+1)}{11} = 7/11$

$Conv(\varphi_1, r) = \frac{1}{11} . \frac{4.(11-4)}{4-3} = 28/11$ $\qquad$ $Conv(\varphi_2, r) = 1$

Because $P((A_1A_2,-,-)) = P((A_1,-)).P((A_2,-)) = 1$, we infer that $(A_1,-)$ and $(A_2,-)$ are independent. Therefore, $Conv(\varphi_2, r) = 1$.

The following proposition expresses an interesting relationship between the confidence and dependency degree.

Proposition 3.4. Let *r(R)* be a relation and $\varphi = (X \rightarrow A_k, t_p)$. Then

$$Conf\,(\varphi, r) = \gamma(\varphi, r) + \frac{\sum_{w \in I^o_{(X,t_p[X])}} \max\{|\,q\,||\,q \in I^o_{(XA_k,t_p[XA_k])}, q \subset w\}}{|\,r\,|}$$

Proof. From (8), we obtain the following connection

$$Conf\,(\varphi, r) = 1 - \frac{|\,r_{t_p[X]}\,| - \sum_{w \in I^o_{(X,t_p[X])}} \max\{|\,q\,|\,q \in I^o_{(XA_k,t_p[XA_k])}, q \subseteq w\}}{|\,r\,|}$$

$$= 1 - \frac{|\,r_{t_p[X]}\,| - |\cup_{q \in I^o_{(A_k,t_p[A_k])}} (X, t_p[X])_*(q)|}{|\,r\,|}$$

$$= \gamma(\varphi, r) + \frac{\sum_{w \in I^o_{(X,t_p[X])}} \max\{|\,q\,|\,q \in I^o_{(XA_k,t_p[XA_k])}, q \subset w\}}{|\,r\,|} \qquad \square$$

We next present an example to analyze three measures, including dependency degree, confidence and conviction for general CFDs.

Example 3.2. Let's compute the measures for the following dependencies from Table 1

$\varphi_1 = (A_1 \rightarrow A_2, 0 \,\|\|\, 1)$	$\varphi_2 = (A_1 \rightarrow A_2, - \,\|\|\, -)$	$\varphi_3 = (A_1A_2 \rightarrow A_3, -,2 \,\|\|\, -)$
$\varphi_4 = (A_1 \rightarrow A_3, 1 \,\|\|\, -)$	$\varphi_5 = (A_1A_2 \rightarrow A_3, -,- \,\|\|\, 0)$	$\varphi_6 = (A_2A_3 \rightarrow A_1, 2,- \,\|\|\, -)$

From Equations 5, 7 and 8, we have

$\gamma(\varphi_1, r) = 7/11$	$\gamma(\varphi_2, r) = 0$	$\gamma(\varphi_3, r) = 8/11$
$Conf(\varphi_1, r) = 10/11$	$Conf(\varphi_2, r) = 7/11$	$Conf(\varphi_3, r) = 10/11$
$Conv(\varphi_1, r) = 28/11$	$Conv(\varphi_2, r) = 1$	$Conv(\varphi_3, r) = 1$
$\gamma(\varphi_4, r) = 6/11$	$\gamma(\varphi_5, r) = 3/11$	$\gamma(\varphi_6, r) = 9/11$
$Conf(\varphi_4, r) = 9/11$	$Conf(\varphi_5, r) = 5/11$	$Conf(\varphi_6, r) = 10/11$
$Conv(\varphi_4, r) = 1$	$Conv(\varphi_5, r) = 1$	$Conv(\varphi_6, r) = 1$

Remark 3.1. From Example 3.2, let's evaluate the measures based on theory analysis**:**

1. The measure *Conv* cannot define how much dependency (or violation) is there in $\varphi = ([X^c, X^v] \rightarrow A_k, t_p)$ with the following CFD forms:
   - $t_p[A_k]$ = '-', $X^c = \varnothing$: φ_2
   - $t_p[A_k]$ = '-', $X^c \neq \varnothing$, $X^v \neq \varnothing$: φ_3 *and* φ_6
   - $t_p[A_k]$ = '-', $X^v = \varnothing$: φ_4
   - $t_p[A_k] \neq$ '-', $X^c = \varnothing$: φ_5

   To unravel this, because $(X, t_p[X])$ and $(A_k, t_p[A_k])$ are independent in these forms, we infer that their Conv measures are always equal to 1 even when any values in data tuples are changed. We can see this limitation in $\varphi_2, \varphi_3, \varphi_4, \varphi_5,$ and φ_6.
2. Dependency degree γ is too strict for measuring the approximation of CFDs.
   Indeed, with dependency φ_2 from Example 3.2, we observe that just four tuples violating φ_2 make dependency degree of φ_2 to be 0.
3. As shown in Example 3.2 for dependencies from φ_1 to φ_6, the measure *Conf* can overcome the drawbacks of γ and *Conv* for general CFDs. However, they cannot measure the distribution of data tupes in the conditional equivalence classes.
   Indeed, if values in tupes t_3 and t_9 corresponding to attributes A_3 and A_2 are changed from 0 to 1 and 3 to 5, then *Conf* (even γ and *Conv*) of φ_2 and φ_4 never change.

Therefore, we propose the following lemma and definitions for a new measure.

From the knowledge granularity of a partition [33], we introduce the incomplete knowledge granularity of conditional partition induced by itemset *(Z, t_p[Z]).*

Definition 3.6. Let $I^{o}_{(Z,t_p[Z])} = \{w_1, \ldots, w_l\}$ be a conditional partition with the incomplete probability distribution $P_{I^{o}_{(Z,t_p[Z])}} = \left(P(w_1), \ldots, P(w_l)\right)$ on the set of tupes *r*, $P(w_i) = |w_i| / |r|$ and $\sum_{w_i \in I^{o}_{(Z,t_p[Z])}} P(w_i) \le 1$. Then incomplete knowledge granularity of $I^{o}_{(Z,t_p[Z])}$ is defined by

$$IE(I^{o}_{(Z,t_p[Z])}) = \sum_{i=1}^{l} |w_i| P(w_i) = \sum_{i=1}^{l} |w_i| \frac{|w_i|}{|r|} \tag{10}$$

Lemma 3.1. Let *r(R)* be a relation *φ = (X → A_k, t_p)* holds on r if and only if $IE(I^{o}_{(X,t_p[X])}) = IE(I^{o}_{(XA_k,t_p[XA_k])})$

Proof. We demonstrate that $IE(I^{o}_{(X,t_p[X])}) = IE(I^{o}_{(XA_k,t_p[XA_k])})$ if and only if there is the same distribution of the tuples *r* into classes $w \in I^{o}_{(X,t_p[X])}$ and $q \in I^{o}_{(XA_k,t_p[XA_k])}$ respectively, i.e, for any $w \in I^{o}_{(X,t_p[X])}$, there exists $q \in I^{o}_{(XA_k,t_p[XA_k])}$ such that $w = q$, i.e, *φ* holds on *r*. Lemma 3.1 is therefore proven. □

From Lemma 3.1, a distribution error degree of φ can be expressed by

$$\lambda_E(\varphi, r) = \frac{IE(I^{o}_{(X,t_p[X])}) - IE(I^{o}_{(XA_k,t_p[XA_k])})}{IE(I^{o}_{(X,t_p[X])})} \tag{11}$$

Then we have a new measure for CFD as follows

Definition 3.7. Let *r(R)* be a relation and *φ = (X → A_k, t_p).* A new measure, called distribution dependency degree, is defined by:

$$\lambda_D(\varphi, r) = \begin{cases} 1 - \lambda_E(\varphi, r) = \dfrac{IE(I^{o}_{(XA_k,t_p[XA_k])})}{IE(I^{o}_{(X,t_p[X])})} & if\ |I^{o}_{(X,t_p[X])}| \neq 0 \\ 0 & otherwise \end{cases} \tag{12}$$

Example 3.3. From Example 3.2, we have *λ_D(φ_2, r) = 5/9* and *λ_D(φ_4, r) = 13/25*. If we change data values as in Remark 3.1 (3.), then *λ_D(φ_2, r) = 23/45* and *λ_D(φ_4, r) = 11/25.*

We see that the bigger measure *λ_D,* the bigger dependency probability of *φ*. If *λ_D(φ, r) = 1*, then *φ* becomes a CFD. Therefore *λ_D* can be used to measure the approximation of ACFDs. Moreover, *λ_D* can measure the distribution degree of data tuples based on the conditional dependencies as indicated in following example.

Table 2. A data table of patients

	Name	Symptom	Disease
t_1	N_1	1	1
t_2	N_2	1	2
t_3	N_3	1	3
t_4	N_4	2	2
t_5	N_5	2	2
t_6	N_6	2	2
t_7	N_7	1	4
t_8	N_8	1	5
t_9	N_9	3	1
t_{10}	N_{10}	2	3
t_{11}	N_{11}	2	4
t_{12}	N_{12}	3	1
t_{13}	N_{13}	3	1
t_{14}	N_{14}	2	1
t_{15}	N_{15}	3	1

Example 3.4. Let *r(R)* be a relation as Table 2

With φ_{11} = *(Symptom → Disease,1 || -)*, φ_{12} = *(Symptom → Disease, 2 || -)*, and φ_{13} = *(Symptom → Disease, 3 || -)* in Table 2, we have $\lambda_D(\varphi_{11}, r) = 1/5$, $\lambda_D(\varphi_{12}, r) = 1/3$, and $\lambda_D(\varphi_{13}, r)=1$. If values in tupes t_{10} and t_{11} corresponding to Disease attribute are changed from 3 to 1 and 4 to 2, then $\lambda_D(\varphi_{12}, r) = 5/9$.

We observe that the nearer the measure $\lambda_D(\varphi_{1i}, r)$ is to 1, the bigger the centralized distribution of the patients in $r_{\varphi_{1i}}$ into one or more diseases is.

We now propose the computation for measures based on the stripped conditional partition to reduce the computational time in discovering ACFDs.

Definition 3.8. Let *r(R)* be a relation and let *(Z, t_p[Z])* be an itemset such that $Z \subseteq R$. A stripped conditional partition of $I^o_{(Z,t_p[Z])}$, denoted $\hat{I}^o_{(Z,t_p[Z])}$ is defined by

$$\hat{I}^o_{(Z,t_p[Z])} = \{w \in I^o_{(Z,t_p[Z])} \mid |w| > 1\} \tag{13}$$

Based on the stripped conditional partition, for any itemset *(Z, $t_p[Z]$)*, where $Z \neq \varnothing$ and $Z \subseteq R$, if $|r_{t_p[Z]}| > 1$, then we have the following definitions, propositions, and lemmas.

Definition 3.9. Let *r(R)* be a relation. Then, the support of φ is computed according Equation 1, where $|r_{t_p}|$ is defined by

$$|r_{t_p}| = |r_{t_p[XA_k]}| = \begin{cases} |w|\,|\,w \in \hat{I}^o_{(Z^c,t_p[Z^c])}, Z = XA_k & \textit{if } Z^c \neq \emptyset \\ |r| & \textit{otherwise} \end{cases} \tag{14}$$

Proposition 3.5. The incomplete knowledge granularity of $I^o_{(Z,t_p[Z])}$ is defined by

$$IE(I^o_{(Z,t_p[Z])}) = \frac{1}{|r|}\left(\sum_{w \in \hat{I}^o_{(Z,t_p[Z])}} |w|(|w|-1) + |r_{t_p[Z]}|\right) \tag{15}$$

Proof. From Equation 3 and 13, we infer that $|r_{t_p[Z]}| - \sum_{w \in \hat{I}^o_{(Z,t_p[Z])}} |w|$ is the number of single conditional equivalence classes removed from the conditional partition $I^o_{(Z,t_p[Z])}$. Hence,

$$\sum_{w \in I^o_{(Z,t_p[Z])}} |w|^2 = \sum_{w \in \hat{I}^o_{(Z,t_p[Z])}} |w|^2 + |r_{t_p[Z]}| - \sum_{w \in \hat{I}^o_{(Z,t_p[Z])}} |w| = \sum_{w \in \hat{I}^o_{(Z,t_p[Z])}} |w|(|w|-1) + |r_{t_p[Z]}| \tag{16}$$

Proposition 3.5 is therefore proven. □

Now through the stripped conditional partitions, the new measure λ_D can be computed using Equations 12, 14, and 15.

The computation of *Conf* is based on Proposition 3.6 as follows

Proposition 3.6. Let *r(R)* be a relation and $\varphi = (X \rightarrow A_k, t_p)$ where $X \subseteq R$, $A_k \in R$. Then, the confidence measure *Conf* of φ *is* defined by

$$Conf(\varphi, r) = \begin{cases} 1 - \dfrac{\sum_{w \in \hat{I}^o_{(X,t_p[X])}} f(w)}{|r|} & \textit{if } t_p[A_k] = \text{'}-\text{'} \\ 1 - \dfrac{|r_{t_p[X]}| - |r_{t_p[XA_k]}|}{|r|} & \textit{otherwise} \end{cases} \tag{17}$$

where

$$f(w) = \begin{cases} |w| - \max\{|q| \mid q \in \hat{I}^o_{(XA_k,t_p[XA_k])}\} & \textit{if there exists } q \in \hat{I}^o_{(XA_k,t_p[XA_k])} \textit{ such that } q \subseteq w \\ |w| - 1 & \textit{otherwise} \end{cases}$$

$|r_{t_p[X]}|$and $|r_{t_p[XA_k]}|$ are computed based on Equation 14.

Proof. If $t_p[A_k]$ = '-', then for any equivalence class $w \in I^o_{(X,t_p[X])}$, there exist the equivalence classes $q \in I^o_{(XA_k,t_p[XA_k])}$ such that $w = \cup_{q \in I^o_{(XA_k,t_p[XA_k])}} q$. On the other hand, for any $w \in I^o_{(X,t_p[X])}$, if cardinality of w is equal to 1 and $t_i \in w$ then t_i always satisfies $\varphi = (X \rightarrow A_k, t_p)$. Hence, for any $w \in \hat{I}^o_{(X,t_p[X])}$, if there exists $q \in \hat{I}^o_{(XA_k,t_p[XA_k])}$ such that $q \subseteq w$, then the number of tuples in w violating CFD φ is $f(w) = |w| - \max\{|q| \mid q \in \hat{I}^o_{(XA_k,t_p[XA_k])}, q \subseteq w\}$. Otherwise, q was removed from the conditional partition $I^o_{(XA_k,t_p[XA_k])}$. Thus, $f(w) = |w| - 1$.

If $t_p[A_k] \neq$ '-', then for any $w \in I^o_{(X,t_p[X])}$, if there exists $q \in I^o_{(XA_k,t_p[XA_k])}$ such that $q \subseteq w$, then q is unique. Hence,

$$Conf(\varphi, r) = 1 - \frac{\sum_{w \in I^o_{(X,t_p[X])}} |w| - \sum_{q \in I^o_{(XA_k,t_p[XA_k])}} |q|}{|r|} = 1 - \frac{|r_{t_p[X]}| - |r_{t_p[XA_k]}|}{|r|}$$

Proposition 3.6 is therefore proven. □

4. THE DISCOVERY PROBLEM OF ACFDs

Let *SupThr* be a support threshold, a CFD $\varphi = (X \rightarrow A_k, t_p)$ is frequent if $sup(\varphi, r) \geq SupThr$ [9]

As introduced in [13], with a confidence threshold *ConfThr*, the ACFD $\varphi = (X \rightarrow A_k, t_p)$ holds on the relation r (or r approximately satisfies φ, denoted $r \models_{Conf} \varphi$) if and only if $Conf(\varphi, r) \geq ConfThr$. Then $\varphi = (X \rightarrow A_k, t_p)$ is minimal if 1) $A_k \notin X$, 2) for any proper subset $Y \subset X$, $r \not\models_{Conf} (Y \rightarrow A_k, t_p[Y] \| t_p[A_k])$, and 3) for any pattern tuple s_p where $t_p << s_p$, $r \not\models_{Conf} (X \rightarrow A_k, s_p[X] \| -)$.

Problem 1. Let r(R) be a relation. The discovery problem of ACFDs is to discover the frequent and minimal ACFDs on r.

The right - hand - side (RHS) candidate set of (X, s_p), denoted $C^+(X, s_p)$, is used to check minimality of dependencies and prune the search space of discovery algorithms of CFDs and ACFDs (CTANE and appoxCTANE) based on attribute-set/pattern lattice.

To discover the frequent and minimal ACFDs on r, the algorithm appoxCTANE [13] starts from a set $L_1 = \{(A_k, -) \mid A_k \in R\} \cup \{(A_k, a) \mid \sup(A_k, a) \geq SupThr, A_k \in R, a \in dom(A_k)\}$ and generates L_2 from L_1, L_3 from L_2, …in which, L_l is the set of itemsets (X,s_p) such that the cardinality of X is equal to l. For each itemset $(X,s_p) \in L_l$, the RHS candidate set $C^+(X, s_p)$ is computed by intersection of RHS candidate sets $C^+(X \backslash A_k, s_p[X \backslash A_k])$ for any $A_k \in X$. Then the dependency $\varphi = (X \backslash A_k \rightarrow A_k, s_p[X \backslash A_k] \| s_p[A_k])$ is minimal if and only if $(A_k, s_p[A_k]) \in C^+(X, s_p)$. From that, to mine the frequent and minimal ACFDs on r, appoxCTANE checks the dependencies $\varphi = (X \backslash A_k \rightarrow A_k, s_p[X \backslash A_k] \| s_p[A_k])$ such that $(A_k, s_p[A_k]) \in C^+(X, s_p)$, $(X, sp) \in L_l$, and $A_k \in X$. Let u_p be any tuple pattern such that $u_p[A_k] = s_p[A_k]$ or - , and $u_p[X \backslash A_k] \leq s_p[X \backslash A_k]$. If $Conf(\varphi, r) \geq ConfThr$, then output φ, and remove $(A_k, *)$ from $C^+(X, u_p)$ for every $(X, u_p) \in L_l$ where '$*$' denotes all values for A_k, including the variable . If $Conf(\varphi, r) = 0$, *then* remove $(A_p, *)$ from $C^+(X, u_p)$ for every $A_p \in R \backslash X$, and $(X, u_p) \in L_l$. Next, the algorithm removes itemsets (X, s_p) from L_l such that $C^+(X, s_p)$ is empty. This process is performed for the next levels L_{l+1}, L_{l+2},… until there exists q such that L_q is equal to empty.

Readers can refer $C^+(X, s_p)$ and the algorithms CTANE and appoxCTANE in the papers [9][13].

Now, let $r(R)$ be a relation and let (X, t_p) and (Y, s_p) be two itemsets such that $|r_{t_p}| > 0$ and $|r_{s_p}| > 0$. Then, the itemset (Z, u_p) is generated from (X, t_p) and (Y, s_p) such that $Z = XY$ and

$$u_p = t_p \times s_p = \begin{cases} (t_p, s_p[Y - X \cap Y]) & if\ |X \cap Y| \neq 0\ and\ t_p[X \cap Y] = s_p[X \cap Y] \\ (t_p, s_p) & if\ |X \cap Y| = 0 \end{cases}$$

According to the product of two partitions in [9] and [17], the following lemmas hold based on the conditional partitions and stripped conditional partitions.

Lemma 4.1. The conditional partition $I^o_{(Z,u_p)}$ is computed by

$$I^o_{(Z,u_p)} = I^o_{(X,t_p)} . I^o_{(Y,s_p)}$$

where

$$I^{o}_{(X,t_p)}.I^{o}_{(Y,s_p)}=\{w\cap q \mid w\in I^{o}_{(X,t_p)}, q\in I^{o}_{(Y,s_p)}, |w\cap q|\neq 0\}$$

Lemma 4.2. Assume that $|r_{t_p}|>1$, $|r_{s_p}|>1$ and $|r_{u_p}|>1$. Then the stripped conditional partition $\hat{I}^{o}_{(Z,u_p)}$ is computed by

$$\hat{I}^{o}_{(Z,u_p)}=\hat{I}^{o}_{(X,t_p)}.\hat{I}^{o}_{(Y,s_p)}$$

where

$$\hat{I}^{o}_{(X,t_p)}.\hat{I}^{o}_{(Y,s_p)}=\{w\cap q \mid w\in \hat{I}^{o}_{(X,t_p)}, q\in \hat{I}^{o}_{(Y,s_p)}, |w\cap q|>1\}$$

Lemmas 4.1 and 4.2 are used to compute the products of two conditional partitions and stripped conditional partitions for the itemsets in the levels L_2, L_3, …of the attribute-set/pattern lattice.

The results in Section 3 and 4 allow us effectively improve the computation time for the CFD and ACFD discovery algorithms. These results are also used to evaluate the limitations of the measures (dependency degree, conviction and confidence) in general CFDs and the utility of the proposed measure (λ_D).

5. RESULTS AND DISCUSSION

This section evaluates the effectiveness of the stripped conditional partition for the discovery algorithm of ACFDs based on the confidence measure and the utility of new measure (λ_D).

Based on the algorithm appoxCTANE [13], the algorithms that discover ACFDs using the conditional partitions (CPs) and stripped conditional partitions (SCPs) are called CP-appoxCTANE and SCP-appoxCTANE, respectively.

CP-appoxCTANE algorithm mines the ACFDs through the product of CPs (Lemma 4.1) and the computations of the support and confidence of φ using CPs (Definition 3.1 and Proposition 3.3). While the discovery of SCP-appoxCTANE is based on the product of SCPs (Lemma 4.2) and the computations of the support and confidence of φ using SCPs (Definition 3.1, Definition 3.9 and Proposition 3.6).

Table 3. A description of data sets

	Dataset	# of attributes	# of tuples
1	Synthetic datasets	6-12	500000
2	Blood Transfusion	5	748
3	Nursery	9	12960
4	Chess	7	28056
5	Car Evaluation	7	1728

With the synthetic and real data sets, as shown in Table 3, the experiments are conduced under the CP-appoxCTANE and SCP-appoxCTANE algorithms. These algorithms are implemented in R on a computer with a 3.5 GHz Intel Core i7 processor and 8GB memory.

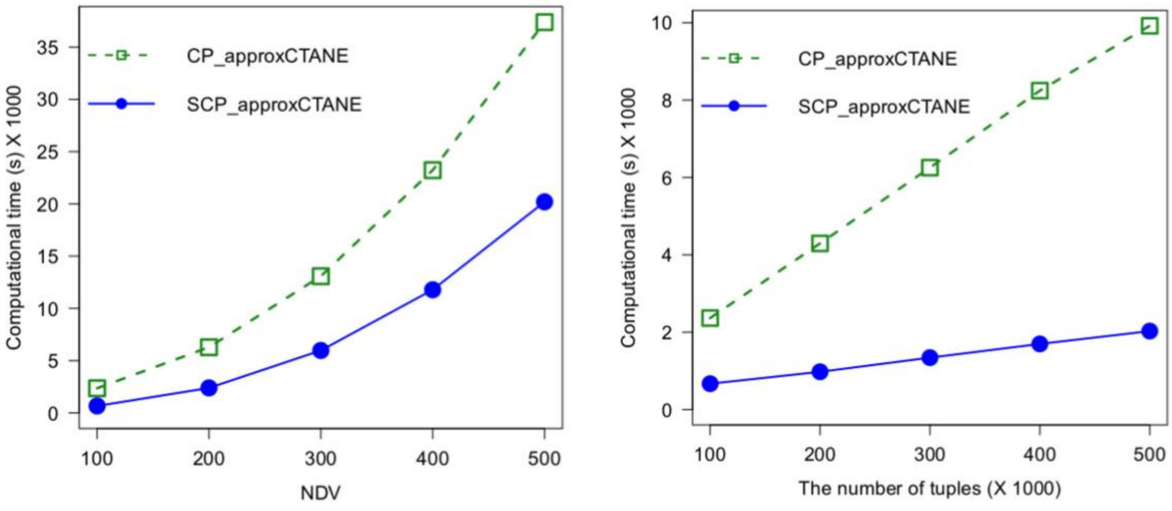
The synthetic datasets sets are generated randomly by varying the number of distinct values of attributes (*NDV*), the number of tuples (|*r*|), the number of attributes (*arity*), and the support threshold (*SupThr*). Note that attributes per data set have the same NDV.

To evaluate the effectiveness of stripped conditional partitions for discovery algorithms, the experiments are carried out as follows:

- We fix *ConfThr*, |*r*|, *arity* and *SupThr* equal to 0.8, 500000, 6, and 100 respectively. *NDV* is varied from 100 to 500.
- Fixing *ConfThr*, *NVD*, *arity* and *SupThr* equal to 0.8, 100, 6 and 0.001 respectively, we vary |*r*| from 100k to 500k.
- With *SupThr* = 0.005, *ConfThr* = 0.8 and *NDV* = 20, *arity* is varied from 6 to 12.
- With |*r*| = 500k, *arity* = 6, *NDV* = 100, we vary *SupThr* from 0.001 to 0.1.

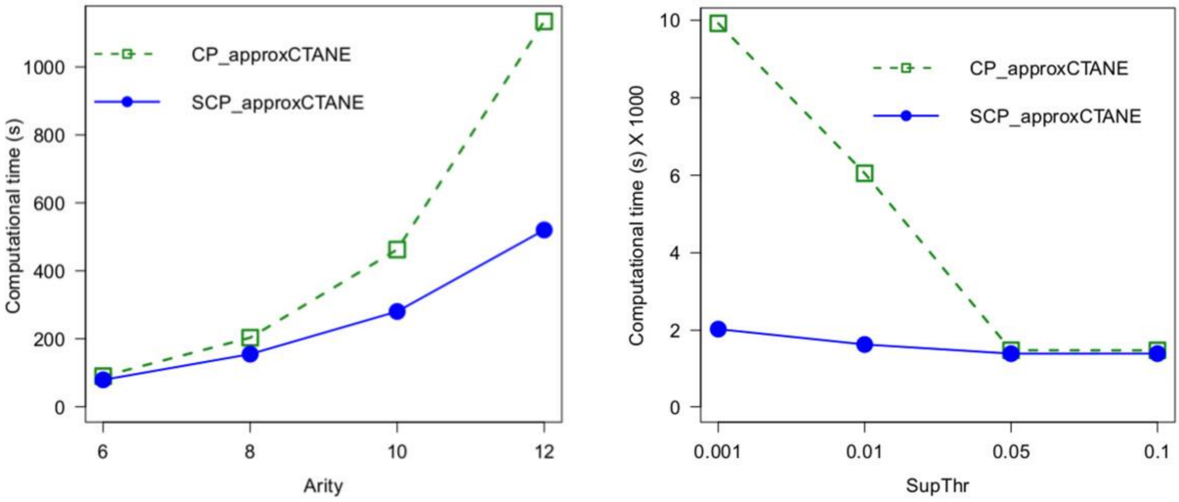
As shown in Figures 1 and 2, SCP-appoxCTANE outperforms CP-appoxCTANE with increasing the number of distinct values, arity as well as the number of tuples and decreasing the support thresholds.

Similarly, we can apply Lemma 4.2 and Lemma 3.1 (the incomplete knowledge granularity of SCPs for CFD φ) and Proposition 3.5 to reduce the discovery time of CFDs for CTANE algorithm in the paper [9].



(A) (B)

Figure 1: Comparison of discovery algorithms of ACFDs based on CP and SCP (varying NDV and |r|)

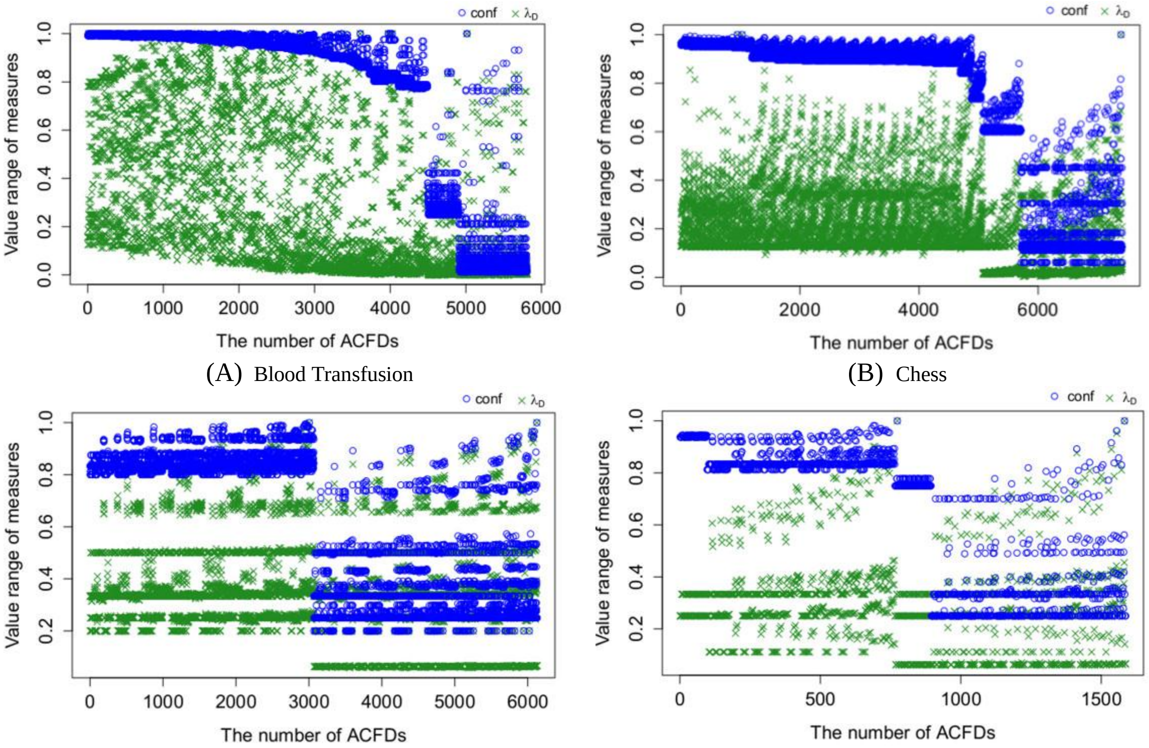


(A) (B)

Figure 2: Comparison of discovery algorithms of ACFDs based on CP and SCP (varying Arity and Support threshold)

With data sets (Blood Transfusion, Chess, Nursery and Car Evaluation) downloaded from the Repository [36], to ensure the evaluation of the measures *Conf* and λ_D on the same set of ACFDs using the algorithm SCP-appoxCTANE, if the measure values of CFDs are greater than 0 in this algorithm, we omitted pruned criteria of RHS candidate set $C^+(X,u_p)$. With the support thresholds 0.05 for Chess and Nursery and 0.01 for Blood Transfusion and Car Evaluation, the discovered ACFDs $\varphi = (X\backslash A_k \rightarrow A_k, s_p[X\backslash A_k]|| s_p[A_k])$ in each dataset are sorted in ascending by the supports of $(X\backslash A_k, s_p[X\backslash A_k])$. We observe that in Figure 3, the dependencies $\varphi = (X\backslash A_k \rightarrow A_k, s_p[X\backslash A_k]|| s_p[A_k])$ with lower supports of $(X\backslash A_k, s_p[X\backslash A_k])$ tended to have bigger the values of $Conf(\varphi, r)$. The dependencies based on the measure *Conf* focus much more on the group with high confidence, the remaining CFDs are in groups with medium and low confidence. Therefore *Conf* is too lenient when measuring the approximation of CFDs. Whereas the measure λ_D of dependencies spread from low to high values.

From the theoretical and experimental results, we demonstrate that the stripped conditional partition is efficient for discovering of CFDs and ACFDs. Moreover the measure λ_D not only measures centralized distribution degree of the objects into one or more groups, but also measures the approximation of CFDs effectively.



(A) Blood Transfusion

(B) Chess

(C) Nursery

(D) Car Evaluation

Figure 3: An evaluation between *conf* and λ_D

6. CONCLUSION

This paper has introduced the stripped conditional partitions and incomplete knowledge granularity for the computations of measures to effectively improve the discovery time for the general CFDs and ACFDs. From the analysis of the weaknesses of the measures (dependency degree, conviction and confidence), we propose a new measure for the general CFDs. This measure can help us to have a more general view of ACFDs with expectation for extending them to other application domains.

REFERENCES

[1] W. W. Eckerson. Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data. *Technical report, The Data Warehousing Institute*. 2002.

[2] R. Gummadi, A. Khulbe, A. Kalavagattu, S. Salvi, S. Kambhampati. SMARTINT: using mined attribute dependencies to integrate fragmented web databases. *Journal of Intelligent Information Systems*. 2012; 38(3): 575-599.

[3] R.S. King, J.J. Legendre. Discovery of functional and approximate functional dependencies in relational databases. *Journal of Applied Mathematics and DecisionS ciences.* 2003;7(1):49-59.

[4] U. Nambiar, S. Kambhampati. *Mining approximate functional dependencies sand concept similarities to answer imprecise queries.* Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS. 2004; 73-78.

[5] N. Novelli, R. Cicchetti. Functional and embedded dependency inference: a data mining point of view. *Information Systems*, 2001; 26(7): 477-506.

[6] H. Yao, H.J. Hamilton. Mining functional dependencies from data. *Data Mining and Knowledge Discovery*. 2008; 16(2): 197-219.

[7] P. Bohannon, W. Fan, F. Geerts, X. Jia, A. Kementsietsidis. *Conditional functional dependencies for data cleaning*. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, IEEE. 2007; 746-755.

[8] W.Fan, F.Geerts, X. Jia , A. Kementsietsidis. Conditional functional dependencies for capturing data inconsistencies. *ACM Transactions on Database Systems (TODS)*. 2008; 33(2): 6-48,

[9] W. Fan, F. Geerts, J. Li, M. Xiong. Discovering Conditional Functional Dependencies. *IEEE Transactions on Knowledge and Data Engineering*. 2011; 23(5): 683-698.

[10] G. Cong, W. Fan, F. Geerts, X. Jia, S. Ma. *Improving data quality: Consistency and accuracy.* In Proceedings of the 33rd international conference on Very large data bases. 2007; 315-326.

[11] W. Fan, S. Ma, N. Tang, W. Yu. Interaction between record matching and data repairing. *Journal of Data and Information Quality (JDIQ)*, 2014; 4(4): 16-38.

[12] F. Chiang, RJ. Miller. *Discovering data quality rules.* Proceedings of the VLDB Endowment. 2008; 1: 1166-1177.

[13] H. Nakayama, A. Hoshino, C. Ito, K. Kanno. *Formalization and Discovery of Approximate Conditional Functional Dependencies.*in: Database and Expert Systems Applications, Springer. 2013; 118-128.

[14] J. Li, J. Liu, H. I., et al. Effective Pruning for the Discovery of Conditional Functional Dependencies. *The Computer Journal*. 2013; 56(3): 378-392.

[15] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. *Dynamic itemset counting and implication rules for market basket data.* ACM SIGMOD Record. 1997; 26(2): 255-264.

[16] J. Kivinen, H. Mannila. *Approximate dependency inference from relations*. In: Database TheoryICDT92, Springer. 1992; 646: 86-98.

[17] Y. Huhtala, J. Krkkinen, P. Porkka, H. Toivonen. TANE: An efficient algorithm for discovering functional and approximate dependencies. *The computer journal*. 1999; 42(2):100-111.

[18] C. Giannella, E. Robertson. On approximation measures for functional dependencies. Information Systems, 2004 29(6); 483-507.

[19] R.S. King, J.J. Legendre. Discovery of functional and approximate functional dependencies in relational databases. *Journal of Applied Mathematics and Decision Sciences*. 2003; 7(1): 49-59.

[20] U. Nambiar, S. Kambhampati. *Mining approximate functional dependencies and concept similarities to answer imprecise queries*. In: Proceedings of the 7th International Workshop on the Web and Databases, ACM. 2004; 7378.

[21] S.Song, L. Chen, P.S. Yu. Comparable dependencies over heterogeneous data. *The VLDB Journal—The International Journal on Very Large Data Bases*. 2013 ; 22(2): 253-274.

[22] Z. Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 1982; 11(5): 341-356.

[23] Z. Pawlak. Rough sets: *Theoretical aspects of reasoning about data*. Springer. 1991; 9.

[24] B. Li, T.W. Chow, P. Tang. Analyzing rough set based attribute reductions by extension rule. *Neurocomputing* 123. 2014: 185–196.

[25] M. Li, C. Shang, S. Feng, J. Fan. Quick attribute reduction in inconsistent decision tables. *Information Sciences* 254. 2014: 155–180.

[26] J.W. Grzymała-Busse, *Characteristic relations for incomplete data: A generalization of the indiscernibility relation*. in: S. Tsumoto, R. Słowinski, J. Komorowski, J.W. Grzymała-Busse (Eds.), Rough Sets ´740 and Current Trends in Computing, Springer Berlin Heidelberg. 2004: 244–253.

[27] M. Kryszkiewicz. Rough set approach to incomplete information systems. *Information science* 112. 1998: 39–49.

[28] W. Xu, M. Zhang, B. Sun, M. Lin, R. Cheng. Rules mining based on rough set of compatible relation. *Indonesian Journal of Electrical Engineering and Computer Science* 12 . 2014: 6346–6353.

[29] L.A. Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy sets and systems* 90 .1997: 111–127.

[30] Y. Yao. Information granulation and rough set approximation. *International Journal of Intelligent Systems* 16. 2001: 87–104.

[31] Y. Yao, et al. *Granular computing: basic issues and possible solutions.* in: Proceedings of the 5th joint conference on information sciences 1. 2000: 186–189.

[32] R. Sasamal, R.K. Shial, Performance analysis of granular computing model on the basis of s/w engineering and data mining, *IAES International Journal of Artificial Intelligence* 1. 2012: 182.

[33] Q.R. Feng, D.Q. Miao, J. Zhou, Y. Cheng. A novel measure of knowledge granularity in rough sets. *International Journal of Granular Computing, Rough Sets and Intelligent Systems,* 2010; 1(3): 233-251.

[34] D. Maier. The theory of relational databases. Computer science press Rockville. 1983; 11.

[35] L. Golab, H. Karloff, F. Korn, D. Srivastava, and B. Yu. *On Generating Near-Optimal Tableaux for Conditional Functional Dependencies.* Proc. VLDB Endowment, 2008; 1(1): 376-390.

[36] M. Lichman. UCI machine learning repository. School of Information and Computer Sciences, University of California, Irvine. http: //archive.ics.uci.edu/ml. 2013