2006－DBS－140（II）（80）
2006／7／14

一般化されたノイズ入りデータに対する相関ルールマイニング

成田 和世 †　　北川 博之 †

† 筑波大学大学院システム情報工学研究科
〒305-8573 茨城県つくば市天王台 1-1-1

あらまし

情報技術の発達に伴い，ディジタル情報は増加，多様化の一途を辿っている．それに伴い，巨大なデータから隠れた特徴や規則を体系的に発見するデータマイニングの技術は，ますます重要となっている．しかし，実世界に存在するデータは欠損値や誤った値などのノイズを含むものも多い．このような**ノイズ入りデータ**からマイニングされる情報は不正確なものとなってしまう．先行研究 [1] で，我々はあるトランザクションに本来出現するはずのアイテムがそのトランザクションから消失するノイズと，本来出現しないはずのアイテムがそのトランザクションに出現するノイズの，2 種類のノイズを想定して，この 2 種類のノイズのみを含むデータベースから，ノイズのない真の状態のデータベースにおける頻出アイテム集合を推定する手法を提案した．しかし，実世界上のデータには，[1] が想定した 2 種類のノイズ以外のノイズを持つものも多く存在する．本稿では，[1] で提案した頻出アイテム集合の推定法を，より一般的なノイズ入りデータに対しても使用できるように一般化する．

Association Rule Mining for a Generalized Noisy Data Model

Kazuyo NARITA†　　Hiroyuki KITAGAWA†

† Graduate School of Systems and Information Engineering，University of Tsukuba
Tennodai，Tsukuba，Ibaraki，305-8573 Japan

abstract

As we face huge amounts of varied information, data mining, which helps us discover hidden features or rules from voluminous data systematically, has become more important. However, many data in real world are dirty, including noises such as missing values or irrelevant values. The information mined from such *noisy data* becomes incorrect. In our previous work[1], we assumed a noisy data model which involves two kinds of noise: one is that an item which should be in a transaction erronously disappears, and another that an item which should not be in a transaction erronously appears. We proposed the method to estimate *frequent itemsets*[2] on the noiseless data, by probabilistic calculation using the noisy one. However, the real world data may include more complex patterns of noises. In this paper, we present a more generalized noisy data model, and discuss association rule mining under the model.

1　はじめに

情報技術の発達に伴い，ディジタル情報は増加，多様化の一途を辿っている．それに伴い，巨大なデータから隠れた特徴や規則を体系的に発見するデータマイニングの技術は，ますます重要となっている．[2] から始まった相関ルールマイニングの研究分野では，Agrawal らが提案したアプリオリアルゴリズム [3] に始まり，より効率的なマイニングアルゴリズムの提案 [5, 4, 6]，数値型属性を含むデータに対するマイニング [7]，頻出極大アイテム集合 [8] や頻出飽和アイテム集合 [9] の概念についてなど，様々な研究が

広く行われてきた．近年では，ファジー理論の概念を取り入れたマイニングアプローチ [10, 11] やマイニングを行うときに個人情報に対するプライバシーを守るために，故意にデータを摂動し，そこからマイニングを行う研究 [12, 13, 14] など，データが持つあいまいさや信頼性を意識した研究も盛んに行われている．

さらに，実世界にあるデータは欠損値や誤った値などのノイズを含むものも多く，このような**ノイズ入りデータ**からマイニングされる情報は不正確なものとなってしまうことから，データにノイズが含まれていることを考慮に入れたデータマイニングの研究も行われてきている [16, 15]．

我々は特に，確率的にノイズが発生するノイズ入りデータに着目し，そこからノイズのない本来のデータにおける頻出アイテム集合を確率的な推定によってマイニングする手法を研究してきた．先行研究 [1] では，トランザクションに本来出現するはずのアイテムが確率 p によってそのトランザクションから消失するノイズと，本来出現しないはずのアイテムが確率 q によってそのトランザクションに出現するノイズのみを含むノイズ入りデータをマイニングの対象として，そこからノイズのない真の状態のデータベースにおける頻出アイテム集合を推定する手法を提案した．しかし，実世界上のデータには，[1] が想定した 2 種類のノイズ以外のノイズを持つものも多く存在する．

本稿では，[1] で提案した頻出アイテム集合の推定法を，より一般的なノイズ入りデータに対しても使用できるように一般化する．そのために，まず，一般に，ノイズ入りデータがノイズのないデータベースからどのように生成されるのかを検討し，ノイズ入りデータの一般化モデルを示す．そして，[1] の推定法のアプローチを一般化することによって，任意のノイズ入りデータのうちの，どのようなデータに対して本推定手法が有効であるのかを明らかにした上で，我々が提案する推定法の一般形を示す．

以降，本稿の構成は次のようになる．2 で本稿で使用する記法を列挙する．3 で，先行研究 [1] で我々がどのようなノイズ入りデータを想定し，そこからノイズのない真の状態のデータベースにおけるアイテム集合のサポートをどのように推定したかを概説する．4 で確率的なノイズ入りデータのモデルを一般化し，その後，[1] のアプローチに基づく推定法を一般化する．5 で，一般化された推定手法を用いて，ノイズ入りデータをどのようにマイニングするかについて，簡単なノイズ入りデータの例を用いて議論する．6 で本稿に関連する研究に言及する．7 でまとめる．

表 1　トランザクションデータベース TDB

TID	トランザクション
101	a, c, d, e, f
102	a, b, c, e
103	b, d, f
104	a, b, c, f
105	a, c, f

2　定義

この章では，本稿で用いる表現や記法を定義する．

あるトランザクションにアイテムが出現していることを 1，出現していないことを 0 として表現すれば，トランザクションデータベースは全ての要素がブール値の行列として捉えることが出来る．例えば表 1 のトランザクションデータベース TDB は，次のような行列 M_{TDB} で表せる．

$$M_{TDB} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

行列の各行が表 1 の各トランザクションのアイテムの出現を表している．例えば TID001 のトランザクションにアイテム c が出現していることは，行列 M_{TDB} の 1 行 3 列目の要素が示している．

本稿では便宜上，トランザクションデータベースをこのような行列表現として扱う．ノイズのない真の状態のデータベース行列 DB に対して，ノイズ入りデータベース行列を DB' と表す．DB を DB' に対する**真のデータベース**と呼ぶ．行列の大きさはどちらも $N \times m$ である．すなわち，どちらのデータベースも全トランザクション数は N で，出現する全

アイテムの集合を I とすると，$|I| = m$ である．行列 DB，DB' の k 行目はデータベースの k 番目のトランザクションを意味し，それぞれベクトル t_k，t'_k で表す．また，アイテム $a \in I$ に対して，$t_k[a]$，$t'_k[a]$ はベクトル t_k，t'_k において，トランザクションにアイテム a が出現しているかどうかを表す要素の値である．例えば M_{TDB} が DB であるとき，$t_1[c] = 1$ である．

アイテム集合 $X = \{x_1, ..., x_n\} \subset I$ の DB 上での出現回数は，$|\{t_k|t_k[x_1] = 1 \wedge t_k[x_2] = 1 \wedge \ldots \wedge t_k[x_n] = 1\}|$ を意味し，$cnt(X)$ で表す．サポートは $sup(A) = cnt(A)/N$ である．表 1 では，集合 $\{a, b, c\}$ を含むトランザクションは TID が 102 と 104 のものであり，$cnt(\{a, b, c\}) = 2$，$sup(\{a, b, c\}) = 2/5$ である．

3 先行研究

[1] で，我々はデータベース上の各アイテムに独立に働く 2 種類のノイズを想定し，それらのノイズのみを含むノイズ入りデータから真のデータベース上の頻出アイテム集合を発見するために，真のデータベース上のアイテム集合のサポートを，ノイズ入りデータベースから推定する手法を提案した．

[1] で想定したノイズ入りデータは次のような動きで生成される．

定義 1 2 種類のノイズが入った確率的ノイズ入りデータ

任意のアイテム $a \in I$ は，各トランザクション $t_i \in DB$ とその対応トランザクション $t'_i \in DB'$ に対して，

1. $a \in t_i$ ならば，確率 p で $a \in t'_i$ となる．
2. $a \notin t_i$ ならば，確率 q で $a \notin t'_i$ となる．

ただし $p \neq 0.5$ かつ $q \neq 0.5$ である．

今，1-アイテム集合 $X = \{x_1\}(X \subset I)$ に対して，$|\{t_k|t_k[x_1] = 1\}|$，$|\{t_k|t_k[x_1] = 0\}|$ をそれぞれ c_1，c_2 とし，$|\{t'_k|t'_k[x_1] = 1\}|$，$|\{t'_k|t'_k[x_1] = 0\}|$ をそれぞれ c'_1，c'_2 とすると，ベクトル $[c_1, c_2]^T$ と $[c'_1, c'_2]^T$ の間には確率的に次の関係が成り立つことが分かる．

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} p & 1\text{–}q \\ q & 1\text{–}p \end{bmatrix}^{-1} \cdot \begin{bmatrix} c'_1 \\ c'_2 \end{bmatrix}$$

上式から得られる c_1 は真のデータベース上の $cnt(X)$ の推定値で，$sup_{est}(X) = c_1/N$ を我々は**推定サポート**と呼ぶ．

また，2-アイテム集合 $X = \{x_1, x_2\}(X \subset I)$ に対しては，$|\{t_k|t_k[x_1] = 1 \wedge t_k[x_2] = 1\}|$，$|\{t_k|t_k[x_1] = 1 \wedge t_k[x_2] = 0\}|$，$|\{t_k|t_k[x_1] = 0 \wedge t_k[x_2] = 1\}|$，$|\{t_k|t_k[x_1] = 0 \wedge t_k[x_2] = 0\}|$ をそれぞれ c_1，c_2，c_3，c_4 とし，$|\{t'_k|t'_k[x_1] = 1 \wedge t'_k[x_2] = 1\}|$，$|\{t'_k|t'_k[x_1] = 1 \wedge t'_k[x_2] = 0\}|$，$|\{t'_k|t'_k[x_1] = 0 \wedge t'_k[x_2] = 1\}|$，$|\{t'_k|t'_k[x_1] = 0 \wedge t'_k[x_2] = 0\}|$ をそれぞれ c'_1, c'_2, c'_3, c'_4 とすると，ベクトル $[c_1, c_2, c_3, c_4]^T$ と $[c'_1, c'_2, c'_3, c'_4]^T$ の間には次の関係が成り立つ．

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = M^{-1} \cdot \begin{bmatrix} c'_1 \\ c'_2 \\ c'_3 \\ c'_4 \end{bmatrix}$$

ここで，M は確率 p, q から作られる**確率行列**である．

$$M = \begin{bmatrix} p^2 & p(1\text{–}q) & (1\text{–}q)p & (1\text{–}q)^2 \\ p(1\text{–}p) & pq & (1\text{–}q)(1\text{–}p) & (1\text{–}q)q \\ (1\text{–}p)p & (1\text{–}p)(1\text{–}q) & qp & q(1\text{–}q) \\ (1\text{–}p)^2 & (1\text{–}p)q & q(1\text{–}p) & q^2 \end{bmatrix}$$

同様に，n-アイテム集合 $X = \{x_1, ..., x_n\}(X \subset I)$ に対して，

$$\begin{aligned} c_1 &= |\{t_k|t_k[x_1] = 1 \wedge t_k[x_2] = 1 \wedge, \ldots, \wedge t_k[x_n] = 1\}| \\ c_2 &= |\{t_k|t_k[x_1] = 0 \wedge t_k[x_2] = 1 \wedge, \ldots, \wedge t_k[x_n] = 1\}| \\ &\vdots \\ c_n &= |\{t_k|t_k[x_1] = 0 \wedge t_k[x_2] = 0 \wedge, \ldots, \wedge t_k[x_n] = 0\}| \end{aligned}$$

また，

$$\begin{aligned} c'_1 &= |\{t'_k|t'_k[x_1] = 1 \wedge t'_k[x_2] = 1 \wedge, \ldots, \wedge t'_k[x_n] = 1\}| \\ c'_2 &= |\{t'_k|t'_k[x_1] = 0 \wedge t'_k[x_2] = 1 \wedge, \ldots, \wedge t'_k[x_n] = 1\}| \\ &\vdots \\ c'_n &= |\{t'_k|t'_k[x_1] = 0 \wedge t'_k[x_2] = 0 \wedge, \ldots, \wedge t'_k[x_n] = 0\}| \end{aligned}$$

とすると，ベクトル $C = [c_1, c_2, \ldots, c_n]^T$ と $C' = [c'_1, c'_2, \ldots, c'_n]^T$ の間には，対応する $n \times n$ の確率行列 M を用いて次の関係が成り立つ．

$$C = M^{-1} \cdot C'$$

これによって，n-アイテム集合 X の推定サポート $sup_{est}(X)$ も得られることが分かる．

[1] では，定義 1 のノイズ入りデータベースに対して，事前に何らかの方法で得られた確率 p, q が与えられたとき，推定される真のデータベースにおける頻出アイテム集合を全てマイニングする幅優先のアルゴリズムを提案している．そのアルゴリズムで，あるアイテム集合が真のデータベース上で頻出アイテム集合であるかどうかを判定する条件は，与えられた最小サポート min_sup を用いて，以下のようになる．

$$sup_{esp}(X) \geq min\_sup \Rightarrow$$

X は頻出アイテム集合である．

4 一般化されたノイズ入りデータマイニング

本章ではまず確率的ノイズ入りデータの一般化モデルを示し，次に，[1] の推定法をより一般的なノイズ入りデータに対しても使用できるように一般化する．また，一般化した推定法が一般化モデルで表される任意のノイズ入りデータのうち，どこまでのデータに対して使用できるのかを明確にする．

4.1 ノイズ入りデータの一般的表現

ノイズには様々な種類が考えられる．アイテムレベルで見れば，[1] で扱ったノイズのように各アイテムに独立に働くノイズもあれば，トランザクションにあるアイテムが出現することで，同じトランザクションから本来出現するはずの別のアイテムが消えてしまうケースも考えられる．また，あるトランザクションで引き起こされたノイズが，その後のトランザクションに発生するノイズに影響を与えるケースも考えられる．しかし，データベースレベルで見れば，ノイズ入りデータベースのモデルは容易に一般化出来る．一般に，ノイズ入りデータベース DB' は，真のデータベース DB と任意の変換オペレータ F によって，以下のように定義できる．

定義 2 ノイズ入りデータの一般化モデル
真のデータベース DB に対して，任意の確率的なノイズが入ったノイズ入りデータベース DB' を返す任意の変換オペレータ F が与えられたとき，ノイズ入りデータベースの一般化モデルは，

$$DB' = F(DB) \tag{1}$$

となる．

4.2 より一般的なノイズ入りデータからのサポートの推定

ここでは [1] で提案したアイテム集合のサポートの推定法を，より一般的なノイズ入りデータに対しても使用できるように一般化した上で，一般化した推定法が定義 2 の一般化モデルで表される任意のノイズ入りデータのうち，どこまでのデータに対して使用できるのかを明確にする．

3 で概説した先行研究 [1] では，ノイズは各アイテムに独立に働いていた．つまり，あるアイテムの出現が，他のアイテムに働くノイズに及ぼす影響はなかった．しかし，一般には，各アイテムに働くノイズは互いに影響しあっているかも知れず，ノイズが各アイテムに独立に働くとは限らない．そのような場合は 3 で述べた方法では推定サポートが求められない．[1] ではあるアイテム集合 $X = \{x_1, ..., x_n\}$ の推定サポートを求めるには，データ中の各トランザクションがアイテム $x_1, ..., x_n$ をどのように含んでいるかを調べれば良かった．しかし，ノイズが各アイテムに独立に働かない場合は，X に含まれない任意のアイテム y の出現が X の要素アイテムが受けるノイズに影響を及ぼしている可能性があるので，各トランザクションが y をどのように含んでいるかも調べる必要がある．

もし，あるトランザクション t_k と，異なるトランザクション $t_h (h \neq k)$ が，$\forall a \in I, t_k[a] = t_h[a]$ ならば，t_k と t_h はベクトルとして等しい．一般に，全要素がブール値である大きさ m のベクトルは 2^m 個ある．このようなベクトル $[1, 1, ..., 1]$, $[0, 1, ..., 1]$, ..., $[0, 0, ..., 0]$ をそれぞれ $v_1, v_2, \ldots, v_{2^m}$ とし，DB に存在するベクトル v_k に等しいトランザクションの数を $C(v_k) = |\{t_i | t_i = v_k\}|$, DB' に存在するベクトル v_k に等しいトランザクションの数を $C'(v_k) = |\{t'_i | t'_i = v_k\}|$ とすると，ベクトル $[C(v_1), \ldots, C(v_{2^m})]^T$ と $[C'(v_1), \ldots, C'(v_{2^m})]^T$ との間には式 2 が成り立つ．

$$\begin{bmatrix} C(v_1) \\ C(v_2) \\ \vdots \\ C(v_{2^m}) \end{bmatrix} = \begin{bmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,2^m} \\ r_{2,1} & r_{2,2} & \dots & r_{2,2^m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{2^m,1} & r_{2^m,2} & \dots & r_{2^m,2^m} \end{bmatrix}^T \cdot \begin{bmatrix} C'(v_1) \\ C'(v_2) \\ \vdots \\ C'(v_{2^m}) \end{bmatrix} \tag{2}$$

ここで，$r_{i,j}$ は DB 上で v_j であるトランザクション t_k が，ノイズの影響で v_i となる確率で，変換オペレータ F によって定まる．式 2 によって得られる $C(v_k)$ は，明らかに DB' から推定された，DB 上でベクトル v_k に等しいトランザクションの数である．従って，任意のアイテム集合 $X = \{x_1, \dots, x_n\}$ の推定サポート $sup_{esp}(X)$ は以下のように求められる．

$$sup_{esp}(X) = \frac{\sum_{\forall i(1 \le i \le n), v_k[x_i]=1 \text{ となる } v_k} C(v_k)}{N}$$

以上のことから，式 2 を解けば，任意のアイテム集合の推定サポートが得られることが分かる．

式 2 は，同一ベクトルのトランザクションの個数に対して確率的な推定を行っている．同一ベクトルの異なるトランザクションを区別しないので，各トランザクションで同一アイテムに働くノイズが異なる任意のノイズ入りデータからのサポートの推定には対応できない．よって，[1] の推定手法を一般化した式 2 は，同一アイテムに働くノイズが全てのトランザクションにおいて同じである任意のノイズ入りデータにのみ適用できる．トランザクションごとに同一アイテムに働くノイズが異なるようなノイズ入りデータに対してサポートを推定する手法は本稿では議論しない．

データベース上の全てのアイテムの数 m に対して，$2^m \times 2^m$ の行列の計算を行うのは非常に高コストである．コストを下げるためには，どのような変換オペレータ F に基づいてノイズ入りデータが生成されたかによって，可能な限り行列サイズの縮退を行う必要がある．

5 マイニングの例

ここでは，一般化された推定手法を用いて，ノイズ入りデータをどのようにマイニングするかについて，簡単なノイズ入りデータの例を用いて議論する．

[1] で想定したノイズは，各アイテムに独立に働く上，どのアイテムに働くノイズも，同じ二つの確率によってのみ決まった．しかし実世界上のデータに働くノイズはより複雑なものが多い．ここでは，表 1 のトランザクションデータベース TDB を，次のような変換オペレータにより生成されたノイズ入りデータベースとして，具体的にサポートの推定を行うところを見せる．

変換オペレータ F の例

1. アイテム a，b は確率 p_1 で出現の状態が入れ替わる．
2. アイテム c は，真のデータベース上のあるトランザクションに出現しているとき，確率 p_2 でそのトランザクションから消失し，トランザクションに出現していないとき，確率 q_2 でそのトランザクションに出現する．
3. アイテム d は，真のデータベース上のあるトランザクションに出現しているとき，確率 p_3 でそのトランザクションから消失し，トランザクションに出現していないとき，確率 q_3 でそのトランザクションに出現する．
4. アイテム e，f はノイズの影響を受けない．

1. は，$t_k[a] = 1 \wedge t_k[b] = 1$ のとき確率 p_1 で $t'_k[a] = 1 \wedge t'_k[b] = 1$，$t_k[a] = 1 \wedge t_k[b] = 0$ のとき確率 p_1 で $t'_k[a] = 0 \wedge t'_k[b] = 1$，$t_k[a] = 0 \wedge t_k[b] = 1$ のとき確率 p_1 で $t'_k[a] = 1 \wedge t'_k[b] = 0$，$t_k[a] = 0 \wedge t_k[b] = 0$ のとき確率 p_1 で $t'_k[a] = 0 \wedge t'_k[b] = 0$ となることを意味する．このとき，確率 p_1，p_2，p_3，q_2，q_3 によって，式 2 の確率行列の要素確率 $r_{i,j}$ は容易に得られる．

TDB 上のユニークなアイテムの総数は 6 であるので，式 2 の確率行列の大きさは 64×64 となる．しかし，あるアイテム集合に着目して，その推定サポートを求める場合は，そのアイテム集合の要素アイテムと，要素アイテムに働くノイズに

$$\begin{bmatrix} |\{t_k|t_k[a]=1\wedge t_k[b]=1\}| \\ |\{t_k|t_k[a]=1\wedge t_k[b]=0\}| \\ |\{t_k|t_k[a]=0\wedge t_k[b]=1\}| \\ |\{t_k|t_k[a]=0\wedge t_k[b]=0\}| \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-p_1 & p_1 & 0 \\ 0 & p_1 & 1-p_1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}^T \cdot \begin{bmatrix} |\{t'_k|t'_k[a]=1\wedge t'_k[b]=1\}| \\ |\{t'_k|t'_k[a]=1\wedge t'_k[b]=0\}| \\ |\{t'_k|t'_k[a]=0\wedge t'_k[b]=1\}| \\ |\{t'_k|t'_k[a]=0\wedge t'_k[b]=0\}| \end{bmatrix} \quad (3)$$

$$\begin{bmatrix} |\{t_k|t_k[c]=1\}| \\ |\{t_k|t_k[c]=0\}| \end{bmatrix} = \begin{bmatrix} 1-p_2 & q_2 \\ p_2 & 1-q_2 \end{bmatrix}^T \cdot \begin{bmatrix} |\{t'_k|t'_k[c]=1\}| \\ |\{t'_k|t'_k[c]=0\}| \end{bmatrix} \quad (4)$$

影響を与える非要素アイテムが各トランザクションにどのように出現しているかを調べれば十分であり，確率行列は縮退できる．例えば 1-アイテム集合 $\{a\},\{b\},...,\{f\}$ の推定サポートを計算する場合は次のようになる．上記変換オペレータにより，$sup_{est}(\{e\}) = 2$，$sup_{est}(\{f\}) = 4$ は自明である．a に働くノイズは b の出現に影響され，b に働くノイズは a の出現に影響されるので，それぞれ推定サポートを求める際はトランザクションに a と b がどのように出現するかを調べる必要がある．よって，$sup_{est}(\{a\})$ および $sup_{est}(\{b\})$ は式 3 を解くことによって得られる．$sup_{est}(\{c\})$ を求める際は式 4 を解く．$sup_{est}(\{d\})$ を求めるときは確率 p_3, q_3 を用いて式 4 と同様に解けばよい．

ここで，式 3 を解けば，2-アイテム集合 $\{a,b\}$ の推定サポートまで得られることが分かる．また，$\{a,c\}$ の推定サポートは，トランザクションに a と c および b がどのように出現するかを調べれば計算可能であるが，この計算で $\{a,c\}$ の推定サポートと同時に $\{a,b,c\}$ の推定サポートが得られる．このように，変換オペレータによっては，あるアイテム集合に着目して推定サポートの計算を行えば式 2 の確率行列が縮退可能であったり，一回の計算で異なるサイズのアイテム集合の推定サポートが同時に得られることがある．一般に，確率行列の最適な縮退方法や，効率的な処理を行うためのマイニングアルゴリズムの最適な設計は，与えられた変換オペレータによって異なる．

6 関連研究

先行研究 [1] の手法は，プライバシーを考慮に入れたデータマイニング手法 [13] を基にしている．[13] は個人情報が含まれたデータをプライバシーを守りつつマイニングすることを目的に，故意にデータを摂動し，そこから摂動前のデータにおける頻出アイテム集合を，摂動前のデータにおけるアイテム集合のサポートを推定することにより発見する手法を提案している．この摂動には一つのパラメタが使用されているが，それに対して [1] では二つの確率によって 2 種類のノイズが入ったデータからサポートの推定を行っている．本稿の推定法は，先行研究 [1] の手法を，より一般的なノイズ入りデータに対しても使用できるように一般化したものである．

[15] は行列で表現されたノイズ入りデータに対して，ある程度のエラーを許容するという考え方に基づいて頻出アイテム集合のマイニングを行っている．

[16] は，実世界情報の多くがダーティで，そのことがアイテム集合のサポートや相関ルールの確信度に少なからず影響を与え，マイニングされるほとんどの頻出アイテム集合や相関ルールが，しきい値に近い小さなサポートや確信度を持っていることに注目し，しきい値を与えて最小サポートよりある程度小さなサポートを持つアイテム集合をマイニングすることで，従来のマイニングでは見落としてしまうアイテム集合の中から興味深いパターンをマイニングしようと試みている．

[15] と [16] はどちらもデータにノイズが混入していることを考慮したデータマイニングの研究だが，本研究と異なり，確率的な推定計算を行わない．

7 おわりに

本稿では，先行研究 [1] で提案した頻出アイテム集合の推定法を，より一般的なノイズ入りデータに対しても使用できるように，推定手法の一般化を行った．また，ノイズ入りデータの一般化モデルを示し，[1] の推定法のアプローチを一般化することによって一般化モデルで表されるノイズ入りデータのうちの，どのようなデータに対して本推定手法が有効であるのかを議論した．

今後は，同一アイテムに働くノイズが各トランザクションで異なる任意のノイズ入りデータから真のデータベースにおける頻出アイテム集合を推定するにはどのようにすればよいかを検討する．

謝辞 本研究の一部は科学研究費補助金特定領域研究（#18049005）の助成による．

参考文献

[1] 成田 和世, 北川 博之,「ノイズ入りデータからの頻出アイテム集合の推定」，電子情報通信学会第 17 回データ工学ワークショップ (DEWS2006), 2006.

[2] R. Agrawal, T. Imielinski and A. Swami, “Mining Association Rules between Sets of Items in Large Databases, ” Proc. ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, D.C., USA, May, 1993.

[3] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, “Fast discovery of association rules, “ Proc. 20th VLDB, pp. 487-499, Santiago de Chile, Chile, Sep. 1994.

[4] C. Borgelt, “Recursion Pruning for the Apriori Algorithm, “ Proc. the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, vol. 126, Brighton, UK, Nov. 2004.

[5] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation, “ Proc. ACM SIGMOD International Conference on Management of Data, pp. 1-12, Dallas, USA, May 2000.

[6] G. Grahane and J. Zhu, “Fast Algorithm for Frequent Itemset Mining Using FP-Trees, “ IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 10, pp. 1347-1362, Oct, 2005.

[7] R. Srikant, R. Agrawal, “Mining Quantitative Association Rules in Large Relational Tables”, Proc. the 1996 ACM SIGMOD International Conference on Management of Data, pp. 1-12, Montreal, Quebec, Canada, June, 1996.

[8] D. Burdick, M. Calimlim, J. Gehrke, “MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases”, Proc. the 17th International Conference on Data Engineering, pp. 443-452, Heidelberg, Germany, April, 2001.

[9] J. Wang, J. Han and J. Pei, “CLOSET+: searching for the best strategies for mining frequent closed itemsets, “ Proc. 9th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 236-245, Washington, D.C., USA, 2003.

[10] T. P. Hong, K. Y. Lin, S. L. Wang, “Fuzzy data mining for interesting generalized association rules”, Fuzzy Sets and Systems, vol. 138, issue 2, pp. 255-269, 2003.

[11] G. Chen, Q. Wei, “Fuzzy association rules and the extended mining algorithms”, Information Sciences, vol. 147, no. 1-4, pp. 201-228, 2002.

[12] A. V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules, “ Proc. 8th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 217-228, Edmonton, Canada, Jul. 2002.

[13] S. Rizvi and J. R. Haritsa, “Maintaining data privacy in association rule mining, ” Proc. 28th VLDB Conference, pp. 682-693, Hong Kong, China, Aug. 2002.

[14] V. S. Verykos, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin and Y. Theodoridis,

"State-of-the-art in Privacy Preserving Data Mining, " SIGMOD Rec., vol. 33, no. 1, pp. 50-57, 2004.

[15] J. Liu, S. Paulsen, W. Wand, A. Nobel, and J. Pris, "Mining approximate frequent itemsets from noisy data, " Proc. 5th IEEE International Conference on Data Mining, pp. 721-724, Houston, USA, Nov. 2005.

[16] J. Pei, A. K. H. Thung, and J. Han, "Fault-tolerant frequent pattern mining: Problems and challenges, " ACM SIGMOD Workshop on Research Issues in DMKD, Santa Barbara, USA, May 2001.