集約バスケットからの相関関係マイニング

沼尾 雅之　松澤 裕史　松尾 総一郎

電気通信大学大学院情報工学専攻 〒182-8585 東京都調布市調布ヶ丘１－５－１

E-mail: numao@cs.uec.ac.jp, {matuzawa, msou}@nm.cs.uec.co.jp

あらまし 頻出パターン抽出による相関関係分析は，購買パターン分析以外にも，様々な用途への応用が期待されているが，製造業のラインなどでは，取得可能なデータが個別製品単位では得られず，バスケット解析が適用できない場合が多い．本稿では，こうした複数のバスケットが集約されたデータセットから相関関係分析をして，故障パターンを抽出する方法を提案する．具体的には，集約バスケットから仮想的なバスケットを復元する方法および，実例としてプロセス業における故障パターン抽出について説明する

キーワード 相関関係分析，バスケット解析，頻出パターン抽出，データ生成

Association Mining from Aggregated Basket

Masayuki NUMAO　Hirofumi MATSUZAWA　and　Souichirou MATSUO

Graduate School of Computer Science, University of Electro-Communications,

1-5-1 Chofugaoka, Chofu, Tokyo 182-8585 Japan

E-mail: numao@cs.uec.ac.jp, {matuzawa, msou}@nm.cs.uec.ac.jp

Abstract Association rule mining is originally developed for the market basket analysis but expected to be used for a wide range of data analysis such as fault analysis in manufacturing. However, in real plant operation, only batch-level fault data is available. Thus, without unit-level data, the straight forward application of basket analysis is difficult. In this paper, we propose a way to extract a frequent fault pattern from the aggregated dataset. We show how to generate a virtual basket from aggregated basket. The fault pattern analysis in a process industry is also shown as a real example.

Keyword Association Mining, Basket Analysis, Frequent Pattern, Data Generation

1. はじめに

頻出パターン抽出による相関関係分析[1]は，POS データからの購買パターン分析として利用され，データベースマーケティングの重要なツールとして利用されている[3]．また，それ以外にも様々な用途への応用が期待されているが，例えば製造業のラインなどでは，取得可能なデータが個別製品単位ではなく，バッチ単位でしか得られない場合があり，バスケット単位でのデータが必要なバスケット解析ができない場合が多い．本稿では，こうした複数のバスケットが集約されたデータセットから,仮想的なバスケットを復元して，それをもとにして相関関係分析をする方法を示す．

集約バスケットデータとは，複数のバスケットが集約されたデータであり，具体的には，その中にあるバスケットの個数と，アイテムそれぞれの個数情報が得られることになる．バスケット解析をするためには，そこから仮想的なバスケットを復元することになるが，単純には，アイテム毎の平均出現確率を求め，それをもとにバスケットにアイテムを振り分けていく方法が考えられる．しかし，この方法は，アイテムの出現が独立であることを前提にしており，そこからアイテムの従属性である相関関係を導き出すのがふさわしいかの疑問がある．また，技術的にも，アイテムの振り分けられたバスケットの総和が集約バスケットのアイテム個数と不一致になる場合があるなども問題がある．本稿では，こうした集約バスケットデータから，いかにして仮想バスケットを抽出するかについての，理論的な考察および技術的な方法について議論する．

近年，製造業においては，品質管理のために原材料から製品までのデータを個品単位で管理するトレーサビリティが盛んになっているが，実際には物理的およびコスト等の制約によって個品単位にはデータが得られていないのが現実である．さらに製鉄やガラス製造などのプロセス業においては，最終製品の単位と，原

材料の単位は大きく異なっているため，最終製品単位のデータは原材料時点では得られない場合が多い．例えば，原材料についてのデータ収集法としては，一定の時間間隔で測定することしかないが，それは複数の最終製品に対応したデータになってしまうわけである．従ってプロセス業においては集約されたデータしか得られないのは不可避であると考えられ，こうした集約データからでも故障パターンなどの相関分析ができるような手法の開発が求められている．本稿では，集約バスケットからの相関関係抽出の実例として，プロセス業における故障パターン抽出についても説明する．

2. 集約バスケットからの相関分析

まず通常の単純バスケットからの相関分析を以下のように定義する．

- 定義１　頻出パターン抽出

アイテム集合

$$I = \{a_1, \ldots, a_n\}$$

トランザクション集合

$$T = \{t_1, \ldots, t_N\}, t_i \subset I$$

あるアイテム集合 A の**支持度**は以下のように定義される．

$$\mathrm{supp}^T(A) = \frac{|\{t \in T \mid A \subseteq t\}|}{N} \qquad \text{（式 2.1）}$$

また，相関ルール X⇒Y の**確信度**は以下のように定義される．ただし，X⊂A，Y⊂A としたときには，X∩Y=ϕ，X∪Y=A となる．

$$\mathrm{conf}^T(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)} \qquad \text{（式 2.2）}$$

集約バスケットは，複数のバスケットの集約なので，まず，いくつのバスケットが集約されたかという集約度，および，その中のアイテムとその個数データが必要となる．

- 定義２ 集約バスケット

アイテムカウント集合

アイテム(id)が何個ずつ(cnt)あるかを表現するものであり，以下のように表現される．

$$IC = \{\langle id, cnt\rangle \mid id \in I\}$$

集約バスケット

集約バスケットの ID である GID と，その中にいくつのバスケットが入っているかの Size，および，アイテムカウント集合の 3 つ組から構成され，以下のように表現される．

$$at_i = \langle GID_i, Size_i, \{\langle id_{i,1}, cnt_{i,1}\rangle, \ldots, \langle id_{i,L_i}, cnt_{i,L_i}\rangle\}\rangle \in AT$$

集約バスケット集合

集約バスケットの集合であり，これが，データマイニングの入力データセットとなる．

$$AT = \{at_1, .., at_N\}$$

- 定義３ 集約バスケットからの仮想バスケット復元

集約バスケット $at_i \in AT$ に対して，以下のような条件を満たす仮想バスケット列を生成することをバスケット復元と呼ぶ．

仮想バスケット列

$$VT_i = \{vt_{i,1}, .., vt_{i,Size_i}\}, vt_{i,j} \subset I$$

集約バスケット at_i に現れる全てのアイテム $id_{i,j}$, $j=1,..,L_i$ 対して以下の式が満たされる．

$$cnt_{i,j} \geq \left|\{vt \in VT_i \mid id_{i,j} \in vt\}\right| \qquad \text{(式 2.3)}$$

条件式 2.3 で等号が成り立つ時は，集約バスケットのあるアイテムの数と，そのアイテムを持つ仮想バスケット個数が，一致すべきであるという条件になる．これは，生成された仮想バスケットでは，各アイテムの個数がすべて 1 とされることに対応している．これは，定義３から，仮想バスケット列は，アイテム集合 I の部分集合としてしか表現できないためであり，仮想バスケット列を，定義１のトランザクション集合に対応させているためである．

しかし，実際には，ビールが 2 本買い物かごにあっても不思議はないように，同一アイテムを複数個 1 つのバスケットに割り当てて，トランザクション集合としては結果的に 1 つにつぶすようなことも考慮すべきである．また，集約バスケットでは，あるアイテムの個数が,バスケット数を超えていることも許されるが，その場合には，生成された仮想バスケット全てに，そのアイテムを 1 つ以上割り振る必要があり，このことからも不等号の必要性がわかる．

3. 仮想バスケットの復元

本節では，実際に第 2 節の条件式 2.3 を満たすような仮想バスケットの復元法について説明する．

3.1. テーブル定義

集約バスケット集合を表現するためのテーブルを定義する．正規化された 2 種類のテーブル SizeTable と ItemCount を Table1,Table2 のように定義する．

GID	Size
A	3
B	5

Table.1 SizeTable

GID	ID	CNT
A	engine	1
A	noise	2
B	engine	3
B	noise	3
B	brake	2

Table.2 ItemCount

3.2. 仮想バスケットテーブル

仮想バスケットを作成するための作業用テーブル VTWork を Table.3 のように定義する．これは，Size から{1,..,Size}を値とする新しいカラム ISize をつくり，GID と ISize を主キーとするテーブルとなる. ID と CNT は ItemCount テーブルの値がそのまま入る. RNUM は，ISize の順番をランダムに並び変えたものである. このテーブルは, MS SQL を使うと Fig.1 のようにして生成することができる.

GID	ISize	ID	CNT	RNUM
A	1	engine	1	1
A	2	engine	1	2
A	3	engine	1	3
A	1	noise	2	2
A	2	noise	2	1
A	3	noise	2	3
B	1	brake	2	1
B	2	brake	2	4
B	3	brake	2	3
B	4	brake	2	5
B	5	brake	2	2
B	1	engine	3	3
B	2	engine	3	5
B	3	engine	3	1
B	4	engine	3	4
B	5	engine	3	2
B	1	noise	3	3
B	2	noise	3	1
B	3	noise	3	4
B	4	noise	3	2
B	5	noise	3	5

Table.3 VTWork

```
WITH Temp ([GID],[ISize],[ID],[CNT]) AS
  (SELECT [GID],[Size],[ID],[CNT]
    FROM [ItemCount]
    WHERE ISize > 0
  UNION ALL
  SELECT [GID],[ISize]-1,[ID],[CNT]
    FROM Temp
    WHERE ISize > 1
  )
SELECT [GID],[ISize],[ID],[CNT],
        ROW_NUMBER() OVER
    (PARTITION BY [GID],[ID]
    ORDER BY NEWID()) AS ‘RNUM’
 FROM Temp
```

Fig.1 仮想テーブル生成 SQL

最終的な仮想バスケットテーブルは，RNUM≤CNT という条件でアイテムを選びだせば得ることができる. これを Tabel.4 に示す．この方法が第 2 節の仮想バスケットの復元条件式 2.3 を満たしていることは明らかである.

GID	ISize	ID
A	1	engine
A	1	noise
A	3	noise
B	3	brake
B	4	brake
B	1	engine
B	3	engine
B	4	engine
B	2	noise
B	3	noise
B	4	noise

Table.4 VTable

4. 実験と評価

本節では，バスケット復元の精度について実験，評価する．最初に通常のバスケット解析用データセットから，相関関係マイニングツールによって，相関ルールを抽出する．次に，同じバスケットデータを，集約の単位 Size をパラメータとして，Size={2，4，8}のように替えながらまとめていって集約バスケットデータを作成する．そして，ここから，第 3 節の復元方法によって，仮想バスケットを生成し，そこから同じマイ

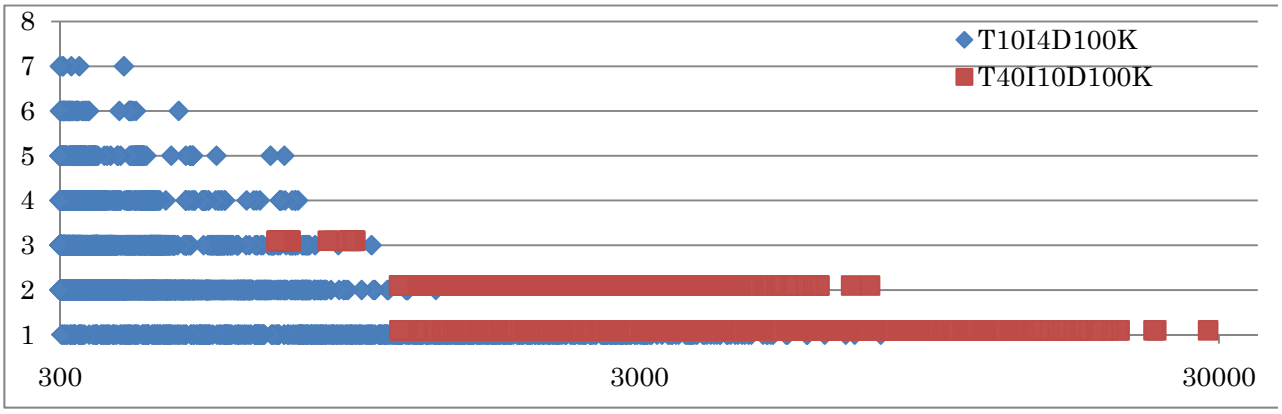

Fig.2　評価用データセットにおける頻出パターン集合の分布

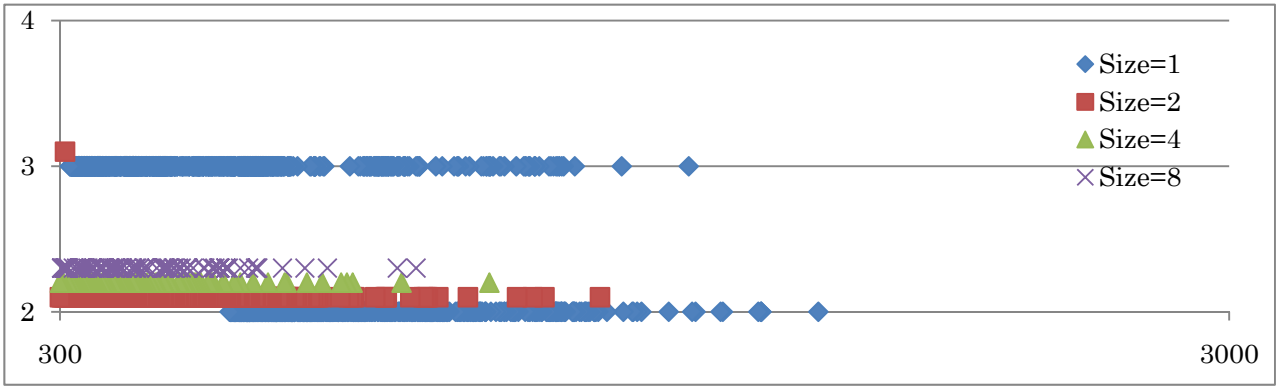

Fig3. 仮想バスケットからの頻出パターン集合の分布 (T10I4D100K)

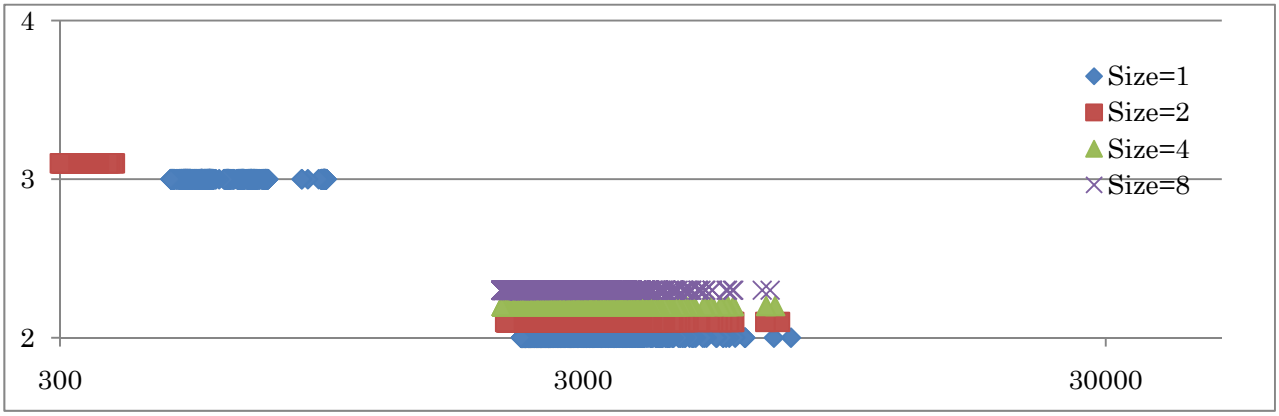

Fig.4 仮想バスケットからの頻出パターン集合の分布（T40I10D100K）

ニングツールによって相関ルールを抽出し，元のバスケットデータから得られた結果と比較する．

4.1. 実験用データセット

QUEST プロジェクト[1]で作成された，バスケットデータ生成プログラムを用いて，生成されたデータセットを使用した[6].

T10I4D100K：トランザクション数 10 万，アイテム数 1000，平均トランザクション長 10

T40I10D100K：トランザクション数 10 万，アイテム数 1000，平均トランザクション長 40

ここから，集約バスケット集合のデータセットを生成する．集約の単位を Size とし，例えば Size=2 の集約バスケットは，元のトランザクション集合 T の要素を順に 2 個ずつまとめて集約バスケットを作っていく．

アイテムカウント集合の和を＋とし，通常の集合の要素 x はアイテムカウント集合の要素<x,1>だと見なすと，集約バスケット集合の要素 at は，以下のように表現される．

$$at_i = < i, Size, t_{i*Size} + .. + t_{(i+1)*Size-1} >$$

ここから，元のデータセットに対して，それぞれ Size=1,2,4,8 の 4 種類の集約バスケット集合を生成する．ただし，Size=1 の集約バスケット集合は，元のデータセットと同等となる．

4.2. 仮想バスケットからの頻出パターン集合の抽出

元データセットによる頻出パターン集合をプロットした結果を Fig.2 に示す．マイニングには，MicroSoft SQL Server 2008 の Analysis Service[7]を用いている．

横軸は頻度の対数，縦軸はアイテム数である．2 種類のデータセットはかなり異なった分布をしていることが分かる[4]．**T40I10D100K** では，アイテム数 4 以上の頻度が意図的に削減されているように見える．一方，**T10I4D100K** では，アイテム数に応じてほぼ指数的に頻度が減少している．

また，平均トランザクション長の違いが頻度分布に大きく影響していることもわかる．特に **T40I10D100K** では，アイテムの種類が 1000 個しかない中で，その内の 40 個をそれぞれのトランザクションに分配しなければならないことから，パターンがかなり人工的に生成されていることが，この頻度分布にも現れていると考えられる．

Fig3,Fig4 には，上記 2 つのデータセットから，Size=1,2,4,8 の集約バスケットを生成し，そこから仮想バスケットを復元し，それに対して，頻出パターン解析をした結果を示す．ここで，アイテム数 1 の頻出パターンの個数は Size に関わらず Fig1,Fig2 と同じであるために省略してある．

T40I10D100K では，Size=2,4,8 の集約バスケットからも，Size= 1 とほぼ同じ頻度で頻出パターンが抽出されていることがわかる．例えば，最頻出のアイテム数 2 のパターンは (529,368) であるが，これは，Size=1,2,4,8 となったときには，それぞれ 7500,7145,6996,6884 という頻度で検出されている．また，頻度順位 100 番目のアイテム数 2 のパターンは (12,419) であるが，これはそれぞれ，頻度 3685,3198,2910,2837 で抽出されている．

一方，**T10I4D100K** では，集約サイズが 2,4,8 と指数的に増えていることに応じて，指数的に頻度が減少していることが読み取れる．例えば，最頻出のアイテム数 2 のパターンは (346,217) であるが，これは，Size=1,2,4,8 となったときには，それぞれ 1336,779,467,339 という頻度で抽出されている．また，頻度順位 100 番目のアイテム数 2 のパターンは (12,722) であるが，これはそれぞれ，頻度 740,468,338,258 で抽出されている．これは，元々同一のバスケットにあった 2 つのアイテムが，そのアイテムを含まない別のバスケット k-1 個と集約されて，結果として Size=k の集約バスケットとなった場合，仮想バスケットとして復元された時に同一バスケットに入る確率が，1/k となることから傾向としては自然であると考えられる．

次に，仮想バスケットから，元のバスケットと同じ頻出パターン集合が，どの程度抽出できるかを評価するために，仮想バスケットで得られた頻出パターン集合の上位 X 個の中に，元の頻出集合の上位 X 個のうち，いくつ(Y 個)が含まれているかを評価した．その結果をグラフ Fig.5,Fig.6 に示す．それぞれ，実験用データセット **T10I4D100K**，**T40I10D100K** に対応していて，パターン長は 2 である．横軸は X，縦軸は Y を示す．また，それぞれの線は，バスケットの集約度 1，2，4，8 に対応している．

例えば，Fig.5 においては，上位 100 番目の中に，集約度 2，4，8 の仮想バスケットからは，それぞれ，76 個，26 個，11 個の正しい頻出集合が復元されたことを示している．また，集約度 2 においては，正しい頻出集合を常に 7 割ぐらい含んでいることもわかる．ここから，仮想バスケットにおいては，頻出集合の中には，正しくないものが含まれており，その数は，集約度に従って増加することが分かる．また，その率は，集約

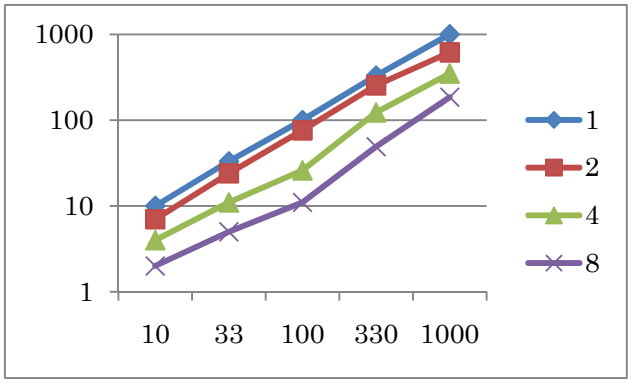

Fig.5　仮想バスケットからの復元率(T10I4D100K)

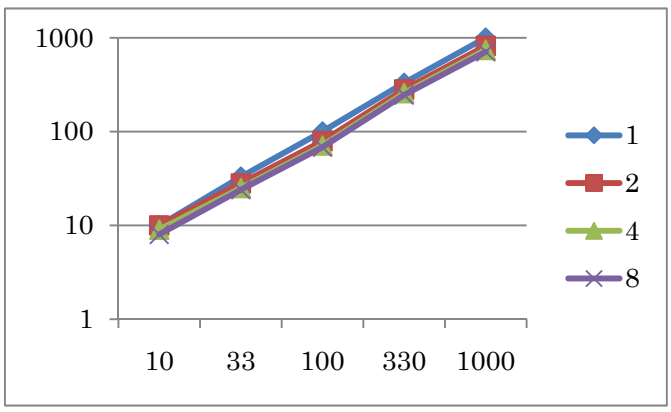

Fig.6　仮想バスケットからの復元率(T40I10D100K)

度が同じであれば一定であるともわかる．次節では，この正しくない頻出集合をいかにして排除するかについて考察する．

4.3. 検出限界に対する考察

4.2 節より，もしアイテムの生起が独立だとすると，アイテム数 n のパターンの現れる確率は式 4.1 で与えられる．これはアイテム間に全く相関がなくても出てくる値なので，支持度の検出限界と考えられる．一方，

Original Basket Items	Aggregated Basket Items	Virtual Basket Items
A,B,C	A,B,C,D	A,B
D		C,D
A,B,C	A,B,C,D	A,C
D		B,D
C	C,D	C
D		D
C	C,D	C,D
D		
C	C,D	D
D		C

Table.5　仮想バスケットの復元例

Size=k の集約バスケットから復元された仮想バスケットにおける支持度は，元の支持度の$(1/k)^{n-1}$となることから，集約サイズが大きくなるに従って，前述の下限に近付いていくことになる．しかし，これは逆に，仮想バスケットから得られた支持度が，その検出下限より大きければ，アイテム間に何らかの相関があったと判断することができるということを意味している．たとえば，Table.5 の表で，左の欄は元のバスケットのアイテム，中欄は，Size=2 の集約バスケットのアイテム，右欄は復元された仮想バスケットのアイテムを示しているものとする．ここで，それぞれの場合についての相関ルール A⇒B と，C⇒D の支持度と確信度を計算してみると以下のようになる．

(a) 元のバスケット
supp({A,B})=0.2, conf(A⇒B)=1.0
supp({C,D})=0.0, conf(C⇒D)=0.0

(b) 仮想バスケット
$supp^V$({A,B})=0.1, $conf^V$(A⇒B)=0.5
$supp^V$({C,D})=0.2, $conf^V$(C⇒D)=0.5

まず，相関関係 A⇒B においては，支持度，確信度ともに仮想バスケットでは半分になっている．これは，バスケットを 2 つずつ集約したことから明らかな結果となっている．一方，相関関係 C⇒D においては，支持度，確信度共に元々は 0 であったものが，相関があるものとしてとらえられている．しかしながら，C,D の単一アイテムとしての出現数は，それぞれ 5 ずつであることから，出現確率は 0.5 となり，C，D が同時に出現する確率は 0.25 ということになる．先ほどの仮想バスケットでの支持度 Supp({C,D})=0.2 は，この値より小さいために検出下限を下回っていると考えられる．ところで，A,B の単一アイテムとしての出現数は，それぞれ 2 ずつであるので，A,B が同時に出現する確率は 0.04 であり，仮想バスケットでの支持度 Supp({A,B})=0.1,はこの値を上回っているので，有意な相関があると判断できる．

単一アイテムの出現頻度は，仮想バスケットにおいても元のバスケットと同じ値になる．ここから，仮想バスケットにおいて抽出された頻出パターンの検出下限を求めることができ，それよりも大きかった時だけ，有意な相関パターンとして抽出することができる．

つまり，有意性の判定は以下の不等式が成り立つことが条件となる．

$$supp^V(X) > \prod_{x \in X} supp^V(\{x\}) \qquad \text{（式 4.2）}$$

5. プロセス業における故障パターン抽出

ガラス瓶製造プロセスにおける欠品データから，実際に欠品種類の相関関係を分析した結果を示す．

5.1. 欠点データ

欠点データは時間単位で得られ，第 3 節の SizeTable に対応するものとして，一定時間毎の欠品数データ，そして ItemCount に対応するものとして，その時間での欠点の種類とその個数データが得られる．これをまとめたものが，Table.6 の欠点データテーブルである．元の SizeTable のキーに対応するものは，データを収集時間である[収集日付]と[収集時刻]，データが収集された場所である[セクション]，[キャビティ]，[型番]であるが，この欠点データテーブルのキーは，これらに，欠点を示す[欠点名]が加わっている．これがアイテムに対応している．また，[欠点個数] は，アイテムの個数に，そして[欠品数]は，集約のサイズにそれぞれ対応するものとなっている．例えば，テーブルの最初の 4 行は，2009 年 9 月 29 日 9：00 に，場所 4B16 において，計 4 個の製品に，「肩びり」，「ねじ下」，「底びり」，「胴びり」の欠点がそれぞれ，4 個，2 個，2 個，1 個現れたということを示している．ここで注意したいの

収集日付	収集時刻	セクション	キャビティ	型番	欠点名	欠点個数	欠品数
20090929	90000	4	B	16	肩びり	4	4
20090929	90000	4	B	16	ねじ下	2	4
20090929	90000	4	B	16	底びり	2	4
20090929	90000	4	B	16	胴びり	1	4
20090929	90000	4	C	132	全高	3	3
20090929	90000	6	A	2	肩びり	2	2
20090929	90000	7	A	21	底びり	4	4

Table.6　欠点データテーブル

収集日付	収集時刻	セクション	キャビティ	型番	識別順	欠点名
20090929	90000	4	B	16	4	肩びり
20090929	90000	4	B	16	3	肩びり
20090929	90000	4	B	16	1	肩びり
20090929	90000	4	B	16	2	肩びり
20090929	90000	4	B	16	4	ねじ下
20090929	90000	4	B	16	2	ねじ下
20090929	90000	4	B	16	3	底びり
20090929	90000	4	B	16	2	底びり
20090929	90000	4	B	16	1	胴びり
20090929	90000	4	C	132	3	全高
20090929	90000	4	C	132	1	全高
20090929	90000	4	C	132	2	全高
20090929	90000	6	A	2	1	肩びり
20090929	90000	6	A	2	2	肩びり

Table.7 仮想バスケットテーブル

は，今回の分析は，欠点の出た製品の中での，欠点名の相関関係であるので，正常な製品は母集団から除外されていることである．

データサイズとしては，SizeTable に対応している欠品数データテーブルの行数が 39,200 行であり，これが集約バスケットの総数に対応している．そして，その中の欠品数の総和は 333,000 行となっているが，これが，復元すべき仮想バスケットの総数に対応する．ここから，集約のサイズの平均は 8.5 であることが分かる．

5.2. 仮想バスケットの生成と相関関係の抽出

この欠点データテーブルから得られた，製品単位で起きる欠点を示す仮想バスケットテーブルを Table.7 に示す．ここでのキーは[収集日付]，[収集時刻]，[セクション]，[キャビティ]，[型番]，[識別順]となり，これが 1 つのバスケットに対応する．そして，「欠点名」が，そのバスケットに入るアイテムに対応している．ここから，例えば，(20090929,90000,1,4,B,16,2) の製品には，肩びり，ねじ下，底びりという 3 つの欠点が同時に起こっていることがわかる．製品に対応する仮想バスケットの総数は 266,000 であるのに対して，欠点の総数は 333,000 で，製品 1 個あたりの平均欠点数は 1.25 であった．マイニングの結果得られた頻出パターンの一部を Table.8 に示す．

単一アイテムの出現頻度から，[胴へこみ]という欠点が，約 23%の製品に現れていることが分かる．このために，頻出頻度上位のアイテム数 2 の頻出パターンの要素としても入っている．ここで，式 4.2 の条件を用いると，[胴へこみ]を含む頻出パターンは除外され，結果として，{[ねじ下],[首びり]} というパターンが得られ，ここから[ねじ下]⇒[首びり]という相関ルー

ル支持度=0.0078，確信度=0.154 という形で得られる．ここから，本来の支持度，確信度は，それぞれ，バスケット集約度 8.5 を乗算して，支持度=0.0663，確信度=1 になることと予想される．

頻度	パターン長	パターン
62199	1	胴へこみ
34774	1	首びり
32819	1	ねじ上
26842	1	ネジ山
22233	1	天咬出
21235	1	天出不良
20006	1	胴泡
18662	1	天泡
13429	1	ねじ下
12888	1	底薄
…		
2493	2	ネジ山，胴へこみ
2318	2	ねじ上，胴へこみ
2225	2	首びり，胴へこみ
2150	2	ねじ上，首びり
2092	2	胴泡，胴へこみ
2081	2	ねじ下，首びり
1559	2	天泡　胴へこみ
1557	2	ネジ山，ねじ上
1542	2	ねじ下，ねじ上
…		

Table.8 欠点の頻出パターン

6. おわりに

本稿では，複数のバスケットが集約されたデータとして，バスケットの個数とアイテムそれぞれの個数情報からなる集約バスケットを定義し，そこから，いかにして仮想バスケットを抽出し，アイテム間の相関関係を抽出するかについての，理論的な考察および技術的な方法について議論した．

バスケット解析においては，個々のバスケット中のアイテムの集合が，頻出パターン抽出のために必要なデータの全てであり，バスケット中にある個々のアイテムの個数は無視される．例えば，ワインとビールが１つずつ入れられた場合も，それぞれ 10 本ずつ入れられた場合も同じバスケットとして扱われるが，後者はより強い相関関係を示すものとして扱われるべきだと思われる．また，別の例としては，懐中電灯１つと乾電池 2 本の組み合わせのようなアイテム*個数間に相関関係がある場合もある．従来，バスケット分析において，アイテムの個数まで扱っているものはない．時系列相関分析[2]では，同一アイテムが時系列的に複数回購買されていることを扱えるが，ある時点でのバスケットの単位では，1 アイテムは 1 個としてしか扱われないので，本質的な問題の解決にはなっていない．本稿で提案した手法は，こうしたアイテム*個数間の相関関係抽出方法への道筋を開くものになると思われる．

本稿では，さらにバスケットという単位自体があいまいであることも指摘した．バスケットは元来，POS での購買単位であることから，データ収集が簡単であることから，データ処理の単位として用いられている．しかし，購買の単位としてみた場合でも，週に 1 回まとめ買いする人もいれば，1 日に数回買い物をする人もいるわけであり，また，休日に，家族連れで全員のものを一度に買っている人が，平日には，自分のものだけを少数買っている場合もあるわけであり，全てを１つのバスケットとして単純に処理すべきかは疑問が残る．そうして考えてみると，今までのバスケット自体が，複数のより小さなバスケットの集約になっていると考えた方が自然である．本稿での提案は，マイクロバスケット分析に有効なものとなると考えられる．今回の集約バスケットでは，その中のバスケットの個数が与えられていたが，マイクロバスケット分析では，それも与えられていないことになるから，こうしたことは，今後の課題となる．

本稿では，支持度，確信度が，バスケットの集約の度合いに対してどのように変化するかについても分析した．その結果，集約度が増加するに従って，有意な支持度，確信度を抽出することが難しくなることもわかった．この優位性の判定基準を，さらに向上させることも今後の課題と考える．

謝辞　本研究の一部は，科学研究費補助金基盤研究(C)(#21500132)による．

参　考　文　献

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. VLDB '94, 1994.

[2] Agrawal, R. and Srikant, R, Mining sequential patterns. In Eleventh International Conference on Data Engineering(ICDE95), IEEE, 1995.

[3] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data, SIGMOD '97, 1997.

[4] C. Cooper and M. Zito, Realistic Synthetic Data for Testing Association Rule Mining Algorithms for Market Basket Databases, PKDD 2007, 2007.

[5] Z. Zheng, R. Kohavi, and L. Mason. Real world performance of association rule algorithms. KDD '01, 2001.

[6] Frequent Itemset Mining Dataset Repository, http://fimi.ua.ac.be/data/

[7] Microsoft Association Algorithm Technical Reference, http://technet.microsoft.com/en-us/library/cc280428.aspx