データベース・システム 74－2
（１９８９．１１．１７）

統計データベースのある最適設計手法について

小林康幸

島根大学

素データから作成される要約表の集まりが与えられた時、集約、射影による操作によってそれらの要約表をすべて導出できる要約表の集まりの中で、総レコード数が最小となるものを見つける問題を考える。与えられた要約表は、利用者が必要とする要約表であり、求める要約表は、計算機にデータベースとして格納する要約表（ベース表）である。すなわちこの問題は、ある意味での計算機での格納容量を最小にする、データベース設計の問題である。

ここでは、まずこの問題を数学的に定式化し、問題を解くためのいくつかの定理を述べた後、それらの定理を用いて作った、問題の解を求めるための方法を与える。

AN OPTIMAL DESIGN METHOD OF STATISTICAL DETABASES

Yasuyuki Kobayasi

Shimane University

1060, Nishikawatsu-cho, Matsue, 690, Japan

This paper treats a problem on statistical database designs; for a given set of summary tables, find a set of base tables to be organized as a database such that all given summary tables are derivable from them and that the total number of records is minimum among such sets of base tables.

After formalizing this problem and describing some theorems to solve the problem, we give a method to obtain a solution of the problem.

1．まえがき

統計データは、通常大容量であることおよび、個々のレコードの機密保護の点から、素データを計算機に格納せず、素データを集約したいくつかの要約表を格納することが多い。この場合、利用者が必要とする要約表の集まりが与えられた時、① どのような要約表の集まりを計算機に格納すれば、それらの要約表を選択、射影、結合、集約などの操作によって導出できるか、② ①で得られた要約表の集まりの中で、計算機での格納容量が最小となるものはどれか、という問題が生じる。

ここでは、これらの問題の定式化および、問題を解くための方法について述べる。

2．問題の定式化

工業統計や商業統計のように、通常統計表は、一つの調査によって得られた素データ（一次統計）から、集約によっていくつかの要約表（二次統計）が作成される。利用者は、与えられた要約表の集まりから、各自に必要な表を、選択、射影、結合、集約などの操作で導出することになる。

素データRを領域の集まり $\{D_i : i = 1,\ldots, N\}$ の直積 $D = \prod_{i=1}^{N} D_i$ の有限部分集合とする。また任意の集合Aに対し、A*をAの有限個の元の列のすべての集まりとし、A上の要約演算子gを、A*の元 a の任意の置換 a'に対し $g(a') = g(a)$ となる関数と定義する。さらにA上の要約演算子 g が次の条件をみたす時、gを結合的要約演算子と呼ぶ：

A*の任意の元 a, a'に対し、$g(g(a), g(a')) = g(a \cup a')$,

但し a ∪ a'は a および a'を一列に並べたAの列とする。合計、最大値などは結合的要約演算子であるが、中央値は結合的でない要約演算子である。

素データR $(\subset \prod_{i=1}^{N} D_i)$、$\{1,\ldots, N\}$ の空でない真部分集合 X および $\{1,\ldots, N\} - X$の元 j に対し、D_X から D_jへの関数$E_{j,X}$を

$$E_{j,X}(d) = (pr_j(r) : r \in R,\ pr_j(r) = d) \quad (d \in D_X)$$

と定義する（但し $\{1,\ldots, N\}$ の空でない任意の部分集合 X に対し、$D_X = \prod_{i \in X} D_i$ および D_X から D_i への射影をpr_iとする。）。

また、$\emptyset \neq Y = Y(G) \subset \{1,\ldots, N\} - X$ に対し、$G = \{g_j : j \in Y(G)\}$ をD_j上の要約演算子g_jの集合とする時、D_X から$D_{X \cup Y}$ への関数$S_{G,X}$，

$$S_{G,X}(d) = (d, (g_j(E_{j,X}(d)) : j \in Y)) \quad (d \in D_X)$$

をR上の集約関数と呼び、$X \subset X' \subset \{1,\ldots, N\}$ なる X'に対し、$D_{X'}$から $D_{X \cup Y}$ への関数 $S_{G,X} \circ pr_{X'}$ を同じ記号$S_{G,X}$で表すことにする（X'⊃Xの時、$D_{X'}$からD_Xへの射影pr_Xは、Xの任意の元iに対し、$pr_i \circ pr_{X'}$が$D_{X'}$からD_iへの射影pr_iと一致する関数とする。）。$S_{G,X}$ によるRの像 $S(G, X; R) = S_{G,X}(R)$ $(\subset D_{X \cup Y})$ を$S_{G,X}$ によるRの要約表と呼び、XおよびYはそれぞれ分類項目および集約項目の集合と呼ぶ。

例2.1. $D_1 = \{ABC, ACC, NSS, QQQ, WSE, BTT\}$,
$D_2 = \{KYOTO, TOKYO\}$ とし、D_3, D_4, D_5, D_6を非負整数の集合とすると、表2.1のR $(\subset \prod_{i=1}^{6} D_i)$ は素データである。また X = {2}，i = 5とすると、

$E_{i,X,R}$ (KYOTO) = (200, 310, 140, 100),
$E_{i,X,R}$ (TOKYO) = (580, 120, 230, 130),
$E_{i,X,R}$ (D_2) = {(200, 310, 140, 100), (580, 120, 230, 130)}
となる。さらに X= {2}, Y= {5, 6}, g_5およびg_6をSUM (合計) とすると、
$S_{G,X}(R)$= {(KYOTO, 750, 134), (TOKYO, 1060, 148)}
となる。

表2.1　素データR

ABC	KYOTO	1	1	200	27
ACC	KYOTO	1	2	310	31
NSS	KYOTO	2	1	140	56
PCC	KYOTO	2	2	100	20
QQQ	TOKYO	2	2	580	21
WSE	TOKYO	2	1	120	74
BTT	TOKYO	1	2	230	35
AXT	TOKYO	1	1	130	18

素データR ($\subset \prod_{i=1}^{N} D_i$) およびR上の集約関数の有限集合の集まりＦＳＡＦが与えられた時、ＦＳＡＦの元 $\mathbb{S}$および $\mathbb{S}'$が次の条件（Ｃ１）を満たす時、$\mathbb{S}$ は $\mathbb{S}'$から導出可能（$\mathbb{S}' \rightarrow \mathbb{S}$と書く）であると呼ぶ：

（Ｃ１）　$\mathbb{S}$の任意の元$S_{G,X}$ に対し、
　ｉ）$\cup \{G' : S_{G',X'} \in \mathbb{S}''\} = G$　　かつ
　ii）$\mathbb{S}''$の任意の元$S_{G',X'}$ に対し
　　$X' \supset X$　かつ$S(G', X; S(G', X'; R)) = S(G', X; R)$
を満たす $\mathbb{S}'$ の部分集合 $\mathbb{S}''$ が存在する（但し上の等式で、$S(G', X; R')$は、$R' = S(G', X'; R)$ ($\subset \prod_{j \in X' \cup Y} D_j$) を素データとした時の$S_{G',X}$ によるR'の要約表である。）。

例2.2.　例2.1においてX1= {2, 3}, X2= {2, 4}, G1= {g5}, G2= {g6}, $\mathbb{S}'$= {$S_{G1,X1}$, $S_{G2,X2}$}, X= {2}, G = {g5, g6}, $\mathbb{S}$ = {$S_{G,X}$} とすると $S_{G1,X1}(R)$, $S_{G2,X2}(R)$, $S_{G,X}(R)$は次のようになり、$\mathbb{S}' \rightarrow \mathbb{S}$を容易に示すことができる。

$S_{G1,X1}(R)$

KYOTO	1	510
KYOTO	2	240
TOKYO	1	360
TOKYO	2	700

$S_{G2,X2}(R)$

KYOTO	1	83
KYOTO	2	51
TOKYO	1	92
TOKYO	2	56

$S_{G,X}(R)$

KYOTO	750	134
TOKYO	1060	148

ＦＳＡＦの元$\mathbb{S}$に対し、次の式で$\mathbb{S}$のレコード数 NRec($\mathbb{S}$)を定義する：

$$NRec(\mathbb{S}) = \sum_{S_{G,X} \in \mathbb{S}} |S_{G,X}(R)|,$$

但し、集合Aに対し、|A|をAの元の数とする。

以上の定義を用いて、当初の目的であった"与えられた要約表をすべて導出できる要約表の集合の中で、計算機での格納容量の最小となるもの"の定義をする。

ＦＳＡＦの元$\mathbb{S}_0$が与えられた時、$\mathbb{S}(\in FSAF)$が次の条件を満たす時、$\mathbb{S}$は$\mathbb{S}_0$に対するレコード最小集合であると呼ぶ：

i) $\mathbb{S} \rightarrow \mathbb{S}_0$

ii) $NRec(\mathbb{S}) = min\{NRec(\mathbb{S}') : \mathbb{S}' \in FSAF \text{ かつ } \mathbb{S}' \rightarrow \mathbb{S}_0\}$

またこの時、$\mathbb{S}$は$\mathbb{S}_0$に対する統計データベース設計の問題に対する解あるいは単にＳＤＤ($\mathbb{S}_0$)の解と呼ぶ。

３．問題の解法

ここでは、まず次のような問題を考える。

空でない有限集合Ｕ、Ｕの被覆$\mathbb{B}$($\mathbb{B}$はＵの部分集合の集まりで、$<\mathbb{B}>$（集合族$\mathbb{A}$に対し、$<\mathbb{A}> = \bigcup_{A \in \mathbb{A}} A$とする。）、およびＵの全ての元ｘに対する非負整数ｆ(ｘ)が与えられた時、次の条件を満たす$2^U - \{\emptyset\}$の部分集合$\mathbb{A}$を$<U, \mathbb{B}, f>$に対する最小積和分解問題の解あるいは$<U, \mathbb{B}, f>$に対するＭＳＰＤと呼ぶ：

i) $\mathbb{A} \succ \mathbb{B}$、すなわち$\mathbb{B}$の任意の元Ｂに対し、$\mathbb{A}$の元Ａが存在し、$A \supset B$

ii) $\mathbb{A}' \succ \mathbb{B}$となる任意の$2^U - \{\emptyset\}$の部分集合$\mathbb{A}'$に対し、$SP_f(\mathbb{A}) \leqq SP_f(\mathbb{A}')$

但し$2^U - \{\emptyset\}$の任意の部分集合$\mathbb{A}'$に対し、$SP_f(\mathbb{A}') = \sum_{A \in \mathbb{A}'} \prod_{x \in A} f(x)$と定義する。

すると次のようなことが示せる。

定理１．$R(\subset \prod_{i=1}^{N} D_i)$を素データ$\mathbb{S}_0(\in FSAF)$をＲ上の有限個の集約関数の集まりとし、$\mathbb{B} = \mathbb{A}(\mathbb{S}_0) = \{X : S_{G,X} \in \mathbb{S}_0\}$およびＵ＝$<\mathbb{B}>$とおく。$pr_U(R) = \prod_{i \in U} pr_i(R)$および要約演算子の有限集合$G_0$が存在し、$\mathbb{S}_0$の任意の元$S_{G,X}$に対し、$G = G_0$であることを仮定する。

i) $\mathbb{S}$がＳＤＤ($\mathbb{S}_0$)の解なら、$\mathbb{A}(\mathbb{S}) \cap U = \{X \cap U : S_{G,X} \in \mathbb{S}\} - \{\emptyset\}$は$<U, \mathbb{B}, f>$のＭＳＰＤである(但し、$f(i) = |pr_i(R)| \quad (i \in U)$)。

ii) 逆に、$\mathbb{A}$が$<U, \mathbb{B}, f>$の解なら、任意の要約演算子の有限集合G_0に対し、$\mathbb{S}(\mathbb{A}) = \{S_{G_0,X} : X \in \mathbb{A}\}$はＳＤＤ($\mathbb{S}_0$)の解である。

例３.１．例２.１においてＹ＝{５，６}，Ｕ＝{２，３，４}，$f_5 = f_6 = SUM$，Ｇ＝{g_5, g_6}，$f(i) = |pr_i(R)| = 2 \ (2 \leqq i \leqq 4)$，$\mathbb{B} = \{(2, 3), (2, 4), (3, 4)\}$，$\mathbb{S}_0 = \{S_{G,B} : B \in \mathbb{B}\}$とすると、$pr_U(R) = \prod_{i \in U} pr_i(R)$であることは容易にわかる。従ってＳＤＤ($\mathbb{S}_0$)の解を求めることは、次のように表わされた$<U, \mathbb{B}, f>$に対するＭＳＰＤを求めることになる。

	2	3	4
B_1	1	1	
B_2	1		1
B_3		1	1
f	2	2	2

さらに $\{U\}$ が$<U,\mathbb{B},f>$に対するＭＳＰＤであること、従って $\{S_{G,U}\}$ がＳＤＤ($\mathbb{S}_0$) の解であることも容易にわかる。

従って、以降では統計データベース設計の問題を解くために、$<U,\mathbb{B},f>$に対するＭＳＰＤを求める方法を考える。

まず簡単な場合について考える。命題および定理の証明は、文献［2］を参照されたい。

命題１．$f(x)=0$となるＵの元ｘが存在するなら、任意のＭＳＰＤ $\mathbb{A}$に対し、$SP_f(\mathbb{A})=0$となる。

命題２．$B\subseteq B'$となる$\mathbb{B}$の元ＢおよびＢ'が存在するなら、$\mathbb{A}$が$<U,\mathbb{B},f>$に対するＭＳＰＤであるための必要十分条件は、$\mathbb{A}$が$<U,\mathbb{B}-\{B\},f>$に対するＭＳＰＤであることである。

命題３．$f^{-1}(0)=\emptyset$, $f^{-1}(1)\neq U$の時

$V=f^{-1}(1)$, $U'=U-V$, $f'=f|_{U'}$

とし、$2^U-\{\emptyset\}$ の部分集合$\mathbb{A}$に対し、$\mathbb{A}-V=\{A-V : A\in\mathbb{A}\}-\{\emptyset\}$ と定義すると、

i) $\mathbb{A}$が$<U,\mathbb{B},f>$に対するＭＳＰＤなら、$\mathbb{A}-V$は$<U',\mathbb{B}-V,f'>$に対するＭＳＰＤとなる。

ii) 逆に$\mathbb{A}'$が$<U',\mathbb{B}-V,f'>$に対するＭＳＰＤなら、$\{A'\cup V : A'\in\mathbb{A}'\}$ は$<U,\mathbb{B},f>$に対するＭＳＰＤとなる。

例３.２．図３.１に示される$<U,\mathbb{B},f>$に対するＭＳＰＤを求めるためには、命題３より図３.２に示される$<U',\mathbb{B}',f'>$に対するＭＳＰＤを求めれば良い。さらに命題２より、図３.３に示される$<U',\mathbb{B}'-\{B_4\},f'>$に対するＭＳＰＤを求めれば良いことがわかり、$\{B_1,B_2,B_3\}$がＭＳＰＤであることが簡単に求められる。

	u_1	u_2	u_3	u_4
B_1	1	1		1
B_2	1		1	
B_3		1	1	
B_4	1			1
B_5		1	1	1
f	3	2	2	1

図 3.1．$<U,\mathbb{B},f>$

	u_1	u_2	u_3
B_1	1	1	
B_2	1		1
B_3		1	1
B_4	1		
f	3	2	2

図 3.2．$<U',\mathbb{B}',f'>$

	u_1	u_2	u_3
B_1	1	1	
B_2	1		1
B_3		1	1
f	3	2	2

図 3.3. $<U', \mathbb{B}' - \{B_4\}, f'>$

命題1～3より、今後は次のことを仮定する：

（A1）$\mathbb{B}$の任意の元Bは$\mathbb{B}-\{B\}$の任意の元B'を含まない。

（A2）Uの任意の元xに対し$f(x) \geqq 2$。

上の仮定の下で、$<U, \mathbb{B}, f>$に対する特殊なタイプのMSPDを求めるために、いくつかの定義をする。$2^U - \{\emptyset\}$の部分集合$\mathbb{C}$が次の条件を満たす時、$\mathbb{C}$を分解不能族と呼ぶ：

（C2）$<\mathbb{C}>$の任意の元a, bは$\mathbb{C}$のある元Cに含まれる。

仮定（A1）を満たす$2^U - \{\emptyset\}$の部分集合$\mathbb{B}$、Uの部分集合Cおよび$<\mathbb{A}> \cap C = \emptyset$となる$\mathbb{B}$の部分集合$\mathbb{A}$に対し、$\mathbb{C}$が次の条件を満たす時、$\mathbb{C}$を$\mathbb{B}$における$\mathbb{A}$のC－極大分解不能族あるいはC－MIFと呼ぶ：

（C3）$\mathbb{A} \subset \mathbb{C} \subset \mathbb{B}$, $<\mathbb{C}> \cap C = \emptyset$, かつ$\mathbb{C}$は分解不能族である。

（C4）$\mathbb{A} \subset \mathbb{C}' \subset \mathbb{B}$, $<\mathbb{C}'> \cap C = \emptyset$となる任意の$\mathbb{C}'$に対し、$\mathbb{C}' \supset \mathbb{C}$かつ$\mathbb{C}'$が分解不能族ならば、$\mathbb{C}' = \mathbb{C}$となる。

$\mathbb{A} \subset \mathbb{B}$の時、$\mathbb{C}$が次の条件を満たす時、$\mathbb{C}$を$\mathbb{A}$における最大分解不能族と呼ぶ：

（C5）$\mathbb{C} \subset \mathbb{A}$かつ$\mathbb{C}$は分解不能族である。

（C6）$\mathbb{C}' \subset \mathbb{A}$かつ$<\mathbb{C}'> = <\mathbb{C}>$ならば、$\mathbb{C}' \subset \mathbb{C}$である。

例3.3において、$\{B_1, B_2, B_3\}$および$\{B_1, B_4, B_5\}$は、$\mathbb{B}$における$\{B_1\}$の$\emptyset$－MIFであり、従って分解不能族でもある。またu_3, u_4を含む$\mathbb{B}$の元が存在しないことから$\mathbb{B}$は分解不能族ではない。

例3.3.

	u_1	u_2	u_3	u_4
B_1	1	1		
B_2	1		1	
B_3		1	1	
B_4	1			1
B_5		1		1

定理2.（A1～2）の仮定の下で、条件（C7～9）を満たす$\mathbb{B}$の分割βに対する$\mathbb{A}_\beta = \{<\mathbb{C}> : \mathbb{C} \in \beta\}$の集合の中に、$<U, \mathbb{B}, f>$に対するMSPDが存在する。

（C7）βの任意の元𝕮は、分解不能族である。
（C8）βの任意の異なる元𝕮,𝕮'に対し、𝔸⊂𝕮かつ<𝔸>⊂<𝕮'>なら、<𝕮－𝔸>=<𝕮>かつ𝕮－𝔸は分解不能族である。
（C9）｜𝕮｜≧3となるβの元𝕮に対し、$SP_f(\{<\mathbb{C}>\}) < SP_f(\mathbb{C})$が満たされる。

C－MIFの定義における𝔹，𝔸および𝕮に対し、
$D^0(\mathbb{A},\mathbb{C}) = D^0(\mathbb{C}) = \{A \in \mathbb{B} : A \cap \mathbb{C} = \emptyset\}$
$D^{i+1}(\mathbb{A},\mathbb{C}) = \{A \in D^i(\mathbb{A},\mathbb{C}) : <\mathbb{A}>$の任意の元uに対し、$A \subset <D^i(\mathbb{A},\mathbb{C})_u>\}$
とすると、ある整数iが存在し、$D^i(\mathbb{A},\mathbb{C})=D^{i+1}(\mathbb{A},\mathbb{C})$となる。 この時の$D^i(\mathbb{A},\mathbb{C})$を$D^*(\mathbb{A},\mathbb{C})$と表わす。次の定理により$D^*(\mathbb{A},\mathbb{C})$を用いて𝔸のC－MIFを求めることができる（但し、$2^U - \{\emptyset\}$の部分集合𝕮，Uの元uおよびUの部分集合Bに対し、$\mathbb{C}_u = \{A \in \mathbb{C} : A \ni u\}$と定義する。）。

定理3．$\mathbb{A} \not\subset D^*(\mathbb{A},\mathbb{C})$、または$D^*(\mathbb{A},\mathbb{C})=\mathbb{A}$かつ𝔸が分解可能族なら、𝔸のC－MIFは存在しない。

定理4．$D^*(\mathbb{A},\mathbb{C}) \supset \mathbb{A}$かつ$D^*(\mathbb{A},\mathbb{C})$が分解不能族なら、$D^*(\mathbb{A},\mathbb{C})$は唯一の𝔸のC－MIFである。

定理5．$D^*(\mathbb{A},\mathbb{C}) \supset \mathbb{A}$かつ$D^*(\mathbb{A},\mathbb{C})$が分解可能族の時は、
1）$B \cap \mathbb{C} = \emptyset$となる$\mathbb{B} - \mathbb{A}$の元Bが存在して、𝕮が$\mathbb{A} \cup \{B\}$のC－MIFならば、𝕮は$\mathbb{C} \neq \mathbb{A}$となる𝔸のC－MIFである。
2）1）の仮定を満たす𝕮が存在しないなら、𝔸のC－MIFは、𝔸が分解不能族なら𝔸となり、𝔸が分解可能族なら存在しない。

定理6．定理3～5を用いることにより、𝔸のすべてのC－MIFを求めることができる。

次に𝔸の最大分解不能族を求めるための定理を述べる。

定理7．$\mathbb{A} \subset \mathbb{B}$，$\mathbb{A} \in \mathbb{A}$および$\mathbb{C} \subset <\mathbb{A}> - \mathbb{A}$とする時、$\mathbb{C}(\subset \mathbb{A})$に対して次の条件1）～3）を考えると、条件1）から条件2）が、また条件3）から条件1）が成り立つ。
1）𝕮は、$\mathbb{C} \ni \mathbb{A}$，$<\mathbb{C}> \cap \mathbb{C} = \emptyset$および$\mathbb{C} \neq \mathbb{A}$を満たす、𝔸における最大分解不能族である。
2）<𝔸>－𝔸の元uおよび𝕮'が存在して、𝕮'は$\mathbb{C}' \supset \mathbb{C}$となる、𝔸における{A}の$(\mathbb{C}' \cup \{u\})$－MIFである。
3）$\mathbb{C} \subseteq \mathbb{C}' \subset <\mathbb{A}> - \mathbb{A}$なる𝕮'が存在して、𝕮は𝔸における{A}のC'－MIFである。

定理8．定理7を用いると、𝔸における{A}のC'－MIFを求めることにより、定理8の条件1）を満たす、すべての𝕮を求めることができる。

これらの定理を用いて、次のような方法によって、<U,𝔹,f>に対するMSPDを求めることができる。

（Ⅰ）𝔸（⊂𝔹）のある元Aに対して、Aにおける｛A｝の∅－極大分解不能族ℂをすべて求める。
（Ⅱ）（Ⅰ）で求めたℂの集合を用いて、すべての𝔸の最大分解不能族𝔻を求める。
（Ⅲ）（Ⅱ）で求めた𝔻の集合を用いて、すべての𝔸の分解不能族を求める。
（Ⅳ）（Ⅰ）における𝔸およびAを変えることにより、すべての𝔹の分解不能族を求める。
（Ⅴ）（Ⅳ）で求めた𝔼の集合を用いて、分解不能族による𝔹の分割の集合ＩＰ(𝔹)を求める。
（Ⅵ）ＩＰ(𝔹)の各元に対し、条件（Ｃ８），（Ｃ９）をチェックし、それらの条件を満たす元βの集合Ｐ(𝔹,ｆ)を求める。
（Ⅶ）（Ⅵ）で求めたＰ(𝔹,ｆ)の元βに対しＳＰｆ($\mathbb{A}_\beta$)を計算することにより、＜Ｕ,𝔹,ｆ＞に対するＭＳＰＤを求める。

例３.４．図３.４は＜Ｕ,𝔹,ｆ＞が与えられた時の、ＭＳＰＤの候補として考えられる𝔹および｛Ｕ｝と、ＭＳＰＤである。$\mathbb{A}_0$のＳＰｆの値を比較したものである。

		U							MSPDの候補		
		u_1	u_2	u_3	u_4	u_5	u_6	u_7	𝔹	$\mathbb{A}_0$	{U}
𝔹	B_1	1	1	1	1				80		
	B_2	1	1	1		1			80		
	B_3	1	1		1	1			80		
	B_4	1	1			1	1		200	200	160
	B_5	1		1	1	1			80		
	B_6		1	1	1	1			16		1600
	B_7		1	1	1		1		40		
	B_8		1	1		1	1		40		80
	B_9		1		1	1	1		40		
	B_{10}			1	1	1	1		40		
	B_{11}		1	1	1			1	16		
	B_{12}		1	1		1		1	16		32
	B_{13}		1		1	1		1	16		
	B_{14}			1	1	1		1	16		
f		10	2	2	2	2	5	2			
SP_f									760	472	1600

図３.４．

上に示したＭＳＰＤを求める方法(Ⅰ)～(Ⅶ)に従って、$\mathbb{A}_0$を求める。
1）D*(｛B_1｝, ∅)＝｛B_i：i＝1,...,10｝（＝$\mathbb{C}_1$とおく）であり、$\mathbb{C}_1$は分解不能族だから、𝔹における｛B_1｝の∅－ＭＩＦは$\mathbb{C}_1$である。また＜$\mathbb{C}_1$＞＝｛u_i：i＝1,...,6｝より、𝔹における｛B_1｝の｛u_7｝－ＭＩＦでもある。
2）𝔹における｛B_1｝の｛u_5｝－ＭＩＦは｛B_1｝である。𝔹における｛B_1｝の｛u_6｝－ＭＩＦは、｛B_1,B_2,B_3,B_5,B_6｝（＝$\mathbb{C}_2$とおく）である。＜$\mathbb{C}_2$＞＝｛u_i：i＝1,...,5｝より𝔹における

$\{B_1\}$ の $\{u_6, u_7\}$ −ＭＩＦでもある。
3) 1),2) よりＢにおける $\{B_1\}$ の最大分解不能族は、$\{B_1\}$, $\mathbb{C}_1$, $\mathbb{C}_2$ である。
4) $\mathbb{B}-\{B_1\}$ における $\{B_2\}$ の ∅−ＭＩＦは、$\mathbb{C}_1-\{B_1\}$ である。$<\mathbb{C}_1-\{B_1\}>=\{u_i : i=1,\ldots,6\}$ より、$\mathbb{B}-\{B_1\}$ における $\{B_2\}$ の $\{u_7\}$ −ＭＩＦも、$\mathbb{C}_1-\{B_1\}$ である。
5) $\mathbb{B}-\{B_1\}$ における $\{B_2\}$ の $\{u_4\}$ −ＭＩＦおよび $\{u_4, u_7\}$ −ＭＩＦは、$\{B_2, B_4, B_8\}$ $(=\mathbb{C}_3$とおく) である。$\mathbb{B}-\{B_1\}$ における $\{B_2\}$ の $\{u_6\}$ −ＭＩＦおよび $\{u_6, u_7\}$ −ＭＩＦは、$\{B_2, B_3, B_5, B_6\}$ $(=\mathbb{D}$とおく) である。$\mathbb{B}-\{B_1\}$ における $\{B_2\}$ の $\{u_4, u_6\}$ −ＭＩＦは、$\{B_2\}$ である。
6) $<\mathbb{C}_1-\{B_1\}>=<\mathbb{C}_1>$および$<\mathbb{D}>=<\mathbb{C}_2>$だから、3),4),5) より、Ｂにおける $\{B_2\}$ の最大分解不能族は、$\{B_2\}$, $\mathbb{C}_1$, $\mathbb{C}_2$, $\mathbb{C}_3$ である。
7) 以後同様の方法で、Ｂにおける $\{B_i\}$ $(i=1,\ldots,14)$ のいづれかの最大分解不能族は、
$\{B_i\}$ $(i=1,\ldots,14)$, $\mathbb{C}_1=\{B_i : i=1,\ldots,10\}$, $\mathbb{C}_2=\{B_1, B_2, B_3, B_5, B_6\}$,
$\mathbb{C}_3=\{B_2, B_4, B_8\}$, $\mathbb{C}_4=\{B_3, B_4, B_9\}$, $\mathbb{C}_5=\{B_i : i=6,\ldots,10\}$,
$\mathbb{C}_6=\{B_i : i=6,11,\ldots,14\}$ である。
8) $\{\beta \in IP(\mathbb{B}) : \beta$は条件（Ｃ8）を満たす$\} = \{\alpha_i : i=1,\ldots,11\}$,
$\alpha_1=\{\{B_i\} : i=1,\ldots,14\}$, $\alpha_2=\{\mathbb{C}_1\} \cup \{\{B_i\} : i=11,\ldots,14\}$,
$\alpha_3=\{\mathbb{C}_1\} \cup \{\mathbb{C}_6-\{B_6\}\}$, $\alpha_4=\{\mathbb{C}_2\} \cup \{\{B_i\} : i=4,6,\ldots,14\}$,
$\alpha_5=\{\mathbb{C}_2, \mathbb{C}_5\} \cup \{\{B_i\} : i=4,11,\ldots,14\}$, $\alpha_6=\{\mathbb{C}_2, \mathbb{C}_5, \mathbb{C}_6-\{B_6\}\} \cup \{\{B_4\}\}$,
$\alpha_7=\{\mathbb{C}_3\} \cup \{\{B_i\} : i=1,3,5,6,7,9,\ldots,14\}$,
$\alpha_8=\{\mathbb{C}_4\} \cup \{\{B_i\} : i=1,2,5,\ldots 8,11,\ldots,14\}$,
$\alpha_9=\{\mathbb{C}_5\} \cup \{\{B_i\} : i=1,\ldots,5,11,\ldots,14\}$,
$\alpha_{10}=\{\mathbb{C}_5, \mathbb{C}_6-\{B_6\}\} \cup \{\{B_i\} : i=1,\ldots,5\}$,
$\alpha_{11}=\{\mathbb{C}_6\} \cup \{\{B_i\} : i=1,\ldots,5,7,\ldots,10\}$ となる。
ＩＰ(Ｂ)の元であるが、条件（Ｃ8）を満たさないβは多くある。例えば$\mathbb{C}_2$および$\mathbb{C}_3$を用いてＩＰ(Ｂ)の元βを作ろうとすると、
i) $\beta \ni \mathbb{C}_2, \mathbb{C}_3-\{B_2\}$ または
ii) $\beta \ni \mathbb{C}_2-\{B_2\}$, $\mathbb{C}_3$
のいづれかが成り立つ。i) の場合は$\mathbb{C}_3-\{B_2\}$ は分解可能となる。またii) の場合は、（Ｃ8）において$\mathbb{C}=\mathbb{C}_3$, $\mathbb{C}'=\mathbb{C}_2-\{B_2\}$, $\mathbb{A}=\{B_2\}$ とおくと、$\mathbb{A} \subset \mathbb{C}$かつ$<\mathbb{A}> \subset <\mathbb{C}'>$が成り立つが、$\mathbb{C}-\mathbb{A}$は分解可能になる。他の場合も同様にチェックできる。
9) $SP_f(<\mathbb{C}_1>)$ (＝８００) ＞ $SP_f(\mathbb{C}_1)$ (＝６９６),
$SP_f(<\mathbb{C}_2>)$ (＝１６０) ＜ $SP_f(\mathbb{C}_2)$ (＝３２０),
$SP_f(<\mathbb{C}_i>)$ (＝４００) ＞ $SP_f(\mathbb{C}_i)$ (＝３２０), $(i=3,4)$
$SP_f(<\mathbb{C}_5>)$ (＝８０) ＜ $SP_f(\mathbb{C}_5)$ (＝１７６),
$SP_f(<\mathbb{C}_6>)$ (＝３２) ＜ $SP_f(\mathbb{C}_6)$ (＝８０),
$SP_f(<\mathbb{C}_6-\{B_6\}>)$ (＝３２) ＜ $SP_f(\mathbb{C}_6-\{B_6\})$ (＝６４)
だから$P(\mathbb{B}, f) = \{\alpha_i : i=1,4,5,6,9,10,11\}$ となる。
10) $SP_f(\mathbb{A}_{\alpha_1})=760$, $SP_f(\mathbb{A}_{\alpha_4})=600$, $SP_f(\mathbb{A}_{\alpha_5})=504$,
$SP_f(\mathbb{A}_{\alpha_6})=456$, $SP_f(\mathbb{A}_{\alpha_9})=664$, $SP_f(\mathbb{A}_{\alpha_{10}})=616$,
$SP_f(\mathbb{A}_{\alpha_{11}})=712$より、$\mathbb{A}_{\alpha_6}$ $(=\mathbb{A}_0)$ がＭＳＰＤとなる。

4．まとめ

我々は、統計データベースの設計問題の定式化と解を求めるための手法を与えた。しかし、これで統計データベースの設計の問題が完全に解決したわけではない。次に残された問題について述べる。

1）$S_{G,X}$の定義でわかるように、ここでの操作は射影と集約のみを考えている。結合によってできる要約表は、元々素データRより射影と集約で表現できるため、ここでは取り扱う必要がないと考えた。今後選択を含めた問題の定式化をする必要がある。

2）定理1において、$\mathbb{S}_\theta$の任意の元$S_{G,X}$に対し$G=G_\theta$を仮定したが、一般的にはこのような仮定はできない。従ってこの仮定が成立しない場合の問題の解法を求める必要がある。

3）要約表の計算機での格納容量としてレコード数を考えた。これは分類項目の値は索引として各項目ごとに異なる値のみを記憶できること、および 2）における仮定があるため、格納容量はレコード数の定数倍と考えても不自然でないためであった。従って 2）の仮定をはずした場合には別の値を考える必要がある。

4）$|\mathbb{B}|$や$|U|$の値が大きくなった時、ＭＳＰＤを求める方法は効率的かといった心配がある。統計データベースでは$|U| \leqq 20$、$|\mathbb{B}| \leqq 200$を満たす場合がほとんどであるため、この範囲内では我々の求めた方法は、ＭＳＰＤを求めるために効果的であると考えている。今後この方法の改良および評価を行う必要がある。

［謝辞］　本稿作成において、多くの助言をいただいた、広島大学の菅原正博教授および池田秀人助教授に深く謝意を表します。

［参考文献］

［1］ Y. Kobayashi and H. Ikeda, Statistical data model and semantics, Proc. of World Conference on Information Processing/Communication (1989), 438-443.

［2］ Y. Kobayashi, A mathematical study on statistical database designs, to appear in Hiroshima Mathematical Journal.