

医療統計表からの情報抽出を容易にする記法に関する一提案

高羽 実[†] 古橋 武[‡]

[†] 鈴鹿医療科学大学医用工学部 〒510-0293 鈴鹿市岸岡町 1001-1

[‡] 名古屋大学工学研究科 〒461-8603 名古屋市千種区不老町

E-mail: [†] takaba@suzuka-u.ac.jp, [‡] furuhashi@cse.nagoya-u.ac.jp

あらまし 医療統計表は、利用者が必要とする以上の分類項目で分類されており、そのフォーマットは複雑であるため、利用者が必要とする統計値を得ることは容易ではない。我々は、既存の医療統計表のデータを木構造により記述することで情報抽出を容易にする新しい記法を提案する。既存の医療統計表を本記法に自動で変換することは可能である。

キーワード 統計表, 記法, XML

1. まえがき

統計表は、統計値を特定の分類項目で集計した結果をまとめたものである。現在、厚生労働省が公開している医療統計表は、利用者のニーズを広く満たすために、利用者が必要とする以上の分類項目で分類されており、そのフォーマットは、分類項目が左側と上部に分けて記述され、それぞれに複数の分類項目が存在し、それらが階層になっていることや、単独で階層構造になっている分類項目があるなど複雑である。このため、利用者が必要とする統計値を得ることは容易ではない。

過去において統計表に関する様々な研究[1]～[5]がなされているが、既存の統計表からの情報抽出の容易性に関する議論は行われていない。表のデータ構造は、Wang 記法[6]により属性から値までのパスとして記述することができるため、本論文では、Wang 記法を基にして、統計表のデータを木構造により記述する新しい記法を提案する。

2. 統計表のフォーマット

2.1. 統計表を構成する要素

統計表は、分類項目を列又は行に割り当て、対応する欄に統計値を記入した表である。ここでは分類項目に利用される属性を“分類属性”、分類属性の名称を“分類属性名”、統計値の属性を“統計属性”、分類属性と統計属性が取り得る値のことを“属性値”と呼ぶ。図1の統計表を例にとると、“施設の種別”と“年次”が分類属性名であり、“昭和59年”と“62年”、“総数”、“病院”などが分類属性の属性値、“施設数”と“構成割合”、“人口10万対施設数”が統計属性の属性値である。なお、医療統計表では、分類属性名が表中に現れないため、図1では便宜的に分類属性名を補っている。

また、本論文ではヘッダーやデータの領域に次のような名称を使用する[6]。

・ Stub head : 左上の属性名の部分（図1の施設の種

別の部分）

・ Stub : 左側の属性値の列（図1の施設数と総数の列）

・ Boxhead : 上の属性名と属性値の行（図1の年次の部分）

・ Body : データの部分

2.2. 統計表特有のフォーマット

一般に統計表は、次のような統計表特有のフォーマットを持つ。

(1) 統計属性の表記

一般に統計属性が1つ存在する統計表の場合は、統計属性の属性値は表のヘッダーには現れない。例えば図2は、1つの統計属性が5つの分類属性で分類された統計表であるが、統計属性の属性値である“受療率”は表のどこにも現れない。複数の統計属性がある場合は、分類属性と同様、Stub または Boxhead にレイアウトされる。統計属性の属性値が複数ある場合は、それらの統計属性は同じ分類属性で分類される。例えば図1の統計表は、Stub に3つの統計属性の属性値（“施設数”、“構成割合”、“人口10万対施設数”）を持つが、統計属性値のそれぞれに“総数”や“病院”などの“施設の種別”の分類属性で分類されている。

(2) 階層構造を持つ分類属性

分類属性には、大分類や中分類、小分類のように階層構造を持つものがある。例えば図1の統計表は、分類属性“施設の種別”が3階層になっている。このような場合、下の階層の属性値は上の階層の属性値により異なる。例えばこの統計表の“施設の種別”の場合、“病院”の下階層には“精神科病院”や“伝染病院”などがあり、“一般診療所”の下には“有床”と“無床”がある。

	施設の種類	年次			
		昭和 59 年	62 年	平成 2 年	．．．
施設数	総数	131832	137275	143164	
	病院	9574	9841	10096	
	精神科病院	1015	1044	1049	
	伝染病院	12	13	10	
	結核療養所	31	19	15	
	一般病院	8516	8765	9022	
	地域医療支援病院（再掲）	・	・	・	
	．．．				
	一般診療所	78332	79134	80852	
	有床	26459	24975	23589	
	．．．				
構成割合	総数	100	100	100	
	病院	7.3	7.2	7.1	
	精神科病院	10.6	10.6	10.4	
	．．．				
人口 1 0 万対施設数	総数	109.6	112.3	115.8	
	．．．				

図 1 施設数・構成割合・人口 1 0 万対施設数，年次・施設の種類の別

		入院			外来			
		総数	病院	一般診療所	総数	病院	一般診療所	歯科診療所
全国								
総数	総数	1145	1089	56	5551	1461	3091	1000
	0 ～ 4 歳	360	349	11	6439	1237	4808	394
	5 ～ 1 4	111	110	1	3204	490	1900	814
	．．．							
男	総数	1080	1042	39	4815	1377	2557	881
	0 ～ 4 歳	392	376	16	6727	1359	5023	345
	．．．							
女	総数	1206	1134	72	6252	1541	3599	1113
	0 ～ 4 歳	327	321	6	6137	1108	4582	447
	．．．							
北海道								
総数	総数	1667	1589	77	5923	2014	2775	1134
	0 ～ 4 歳	409	409	—	6425	1582	4417	425
	．．．							
男	総数	1580	1517	63	5085	1852	2269	964
	0 ～ 4 歳	421	421	—	6895	1842	4846	206
	．．．							
．．．								

図 2 受療率（人口 1 0 万対），入院－外来・施設の種類の別×性・年齢階級×都道府県別の別

3. 提案記法

表のデータ構造は，属性から値までのパスとして記述することができる[6]．このため，本研究では，記述と抽出のしやすさを考慮して，統計表を木構造で記述する．

木構造からの抽出は，木構造を構成する各要素を属性値で絞り込むことができれば可能である．そのためには，任意の要素に抽出の指示を与えられるようにする必要がある．しかし医療統計表には，分類属性が明示されておらず，特に階層のある分類属性の場合は，その属性値がどの分類属性に属しているのかわからない．属性の種類は，Stub 側の属性は列の位置により，Boxhead 側の属性は行の位置により特定することがで

きる．そこで，同じ属性であることを示すため，各要素に要素名として属性値の位置情報から求めたレベル番号を付ける．階層のある分類属性の場合は，階層ごとに別のレベル番号を振る．Stub と Boxhead のレベル番号の順は，本研究では Stub 側から先に番号を振り，Boxhead 側にそのあとの続きの番号を振る．これにより，分類属性ごと，階層ごとに同じレベル番号が振られるため，抽出の指示をレベル番号により与えることができるようになる．なお，階層のある分類属性の場合，上のレベルには下のレベルの合計値を持つが，“総数”のような名称は付いていないため，上のレベルの合計値は，下のレベルがないものを抽出することになる．

本研究では、木構造への変換と抽出の機能の豊富さから、統計表を XML 形式で記述する記法とする。XML による記述方法として、要素名に e1 からの連番を付け、要素に属性値を id 属性として記述し、最後の要素のテキストノードに統計値を記述する。なお、統計表には数字で始まる属性値が存在するが、先頭に数字を含む値を属性値として使えないという XML の記述規則に対応するため、全ての属性値の先頭に“n”を加える。

図 1 の統計表の本記法による記述例を図 3 に示す。

```
<root>
  <e1 id="n 施設数">
    <e2 id="n 総数">
      <e5 id="n 昭和 59 年">131832</e5>
      <e5 id="n62 年">137275</e5>
      <e5 id="n 平成 2 年">143164</e5>
      . . .
    </e2>
    <e2 id="n 病院">
      <e5 id="n 昭和 59 年">9574</e5>
      <e5 id="n62 年">9841</e5>
      <e5 id="n 平成 2 年">10096</e5>
      . . .
    <e3 id="n 精神科病院">
      <e5 id="n 昭和 59 年">1015</e5>
      <e5 id="n62 年">1044</e5>
      <e5 id="n 平成 2 年">1049</e5>
      . . .
    </e3>
  </e1>
</root>
```

図 3 図 1 の統計表の本記法による記述

4. 提案記法からの抽出

4.1. 抽出のパターンと抽出の考え方

利用者が行うと想定される統計表からの抽出は、以下のようなパターンに分けられる[5]。

- (1) 分類属性の属性値を指定しての抽出
- (2) 統計属性の属性値を指定しての抽出
- (3) 選択した分類属性だけで分類しての抽出
- (4) 階層のある分類属性から特定のレベルの抽出

統計表を木構造にした場合のこれらの抽出の機能別の考え方を述べる。

分類属性の属性値を指定しての抽出

統計表は、各分類属性の属性値の直積に統計値を対応させたものであるため、統計表を木構造に変換すると、各要素が同じ属性値を持つ子要素を持つ形になる。例えば図 4 では、“全国”は、その下に“総数”、“男”、“女”を持ち、“北海道”や“青森”も“総数”、“男”、“女”を持つという構造になる。この抽出は、分類属性に対応した要素を属性値を指定して抽出する。

統計属性の属性値を指定しての抽出

分類属性と統計属性は、実際の統計表では明確には区別されていない。例えば図 1 の統計表では、統計属性として“施設数”、“構成割合”、“人口 10 万対施設数”があるが、これらは分類属性の属性値と見なしても特に問題ない。このため、この抽出は(1)と同等と見なすことができる。

選択した分類属性だけで分類しての抽出

この抽出は、選択していない分類属性の全ての属性値の値を合計する必要があるが、実際の統計表には“計”や“総数”のような属性値があるため、選択しない分類属性はこれらの属性値を抽出することにより行う。

階層のある分類属性から特定のレベルの抽出

階層のある分類属性の場合、下位レベルの属性値は上位レベルによって異なるため、上位レベルから下位レベルまで順を追って属性値を指定して抽出する。但し実際には、階層ではなく異なる分類属性の場合でも、上位の分類属性によって下位の分類属性の属性値が変わる場合があるため、各分類属性の直積になっているとは限らない。例えば図 2 の統計表は、Boxhead に分類属性“入院－外来”と“施設の種類”があるが、“入院－外来”の属性値が“入院”の場合は“施設の種類”の属性値“歯科診療所”は省略されている。

以上のことから、これらの抽出は全て、各属性の属性値を上位レベルから順に指定して絞り込んでいくという方法により対応可能である。

4.2. XPath による抽出

4.1. に示した抽出は、XPath により次のように記述することが可能である。

分類属性の属性値を指定しての抽出

分類属性に対応する各要素に属性値で絞り込む抽出条件を指定する。

例えば図 2 の統計表から北海道の男性で年齢階級が 0～4 歳の外来で施設の種類が病院の受療率を求める XPath は次のようになる。

```
root/e1[@id='n 北海道']/e2[@id='n 男性']/e3[@id='n 0 ～ 4 歳']/e4[@id='n 外来']/e5[@id='n 病院']
```

統計属性の属性値を指定しての抽出

統計属性に対応する要素に属性値で絞り込む抽出条件を指定する。

例えば図 1 の統計表から施設の種類が総数の平成 2 年、5 年、8 年の施設数を求める XPath は次のようになる。

```
root/e1[@id='n 施設数']/e2[@id='n 総数']/e5[@id='n 平成2年' or @id='n 5年' or @id='n 8年']
```

選択した分類属性だけで分類しての抽出

指定した分類属性以外の分類属性に対応する要素の属性値を“総数”や“計”で絞り込む。

例えば図 2 の統計表から全国以外の各都道府県の入院患者の受療率（都道府県が全国以外、性が総数、年齢階級が総数、入院－外来が入院、施設の種類が総数の受療率）を求める XPath は次のようになる。

```
root/e1[@id!='n 全国']/e2[@id='n 総数']/e3[@id='n 総数']/e4[@id='n 入院']/e5[@id='n 総数']
```

階層のある分類属性から特定のレベルの抽出

指定するレベルの上位レベルの要素の属性値を指定して絞り込み、指定するレベルの各属性値に対する抽出条件を or で結合する。指定するレベルに下のレベルがある属性値では、複数の属性値が返る。

例えば図 1 の統計表から施設の種類が老人病院の平成 2 年、5 年、8 年の施設数を求める XPath は次のようになる。

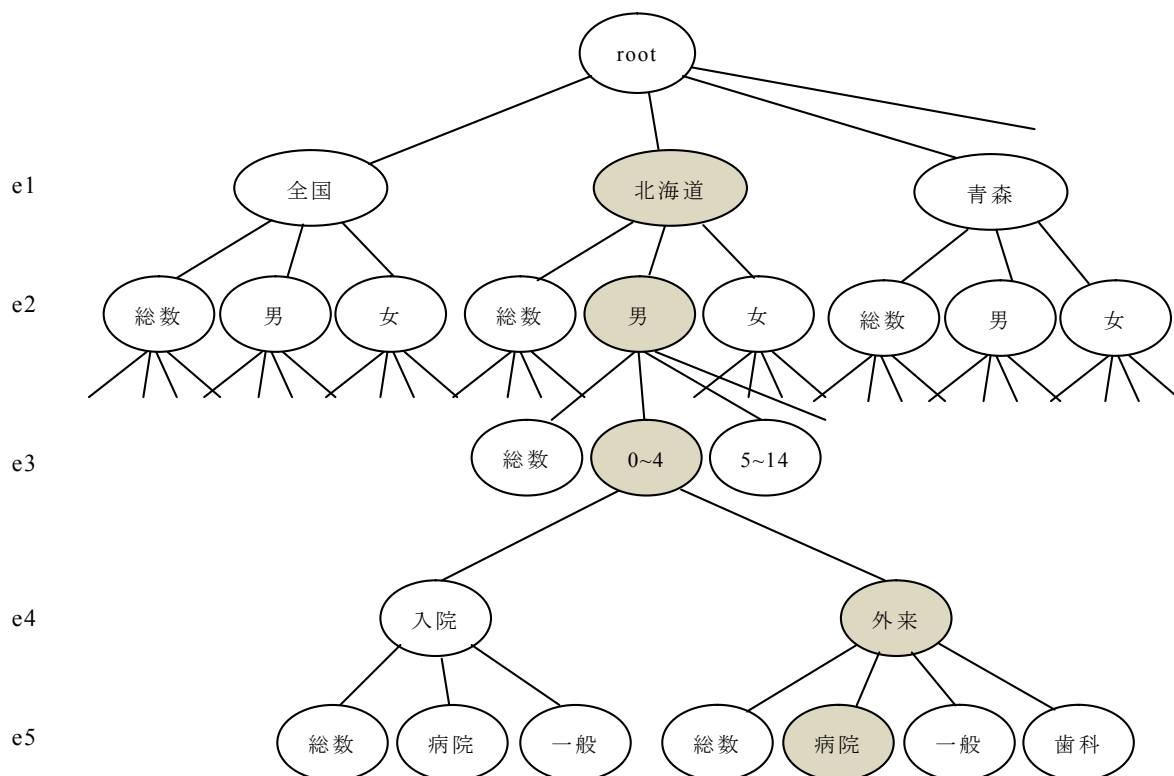


図 4 4.2.の抽出を木構造で表した例

root/e1[@id='n 施設数']/e2[@id='n 病院
']/e3[@id='n 一般病院']/e4[@id='n 老人病院
']/e5[@id='n 平成 2 年' or @id='n5 年' or
@id='n8 年']

このように、本記法で記述した統計表に対して XPath により抽出の記述を行うには、分類属性や統計属性に対応した要素の属性に対して抽出の条件を指定するだけであり、簡潔に表現することが可能である。

5. むすび

本研究では、医療統計表に関する新しい記法を提案した。本記法は、統計表の属性から値までのパスを XML で記述することで、統計表からの抽出に要求される機能である、分類属性や統計値の属性値を指定しての抽出や、指定した分類属性の属性だけで分類しての抽出、階層のある分類属性の上位レベルの抽出を容易に行うことを可能にしている。

今後の課題として、XPath による抽出の指定や XSLT による抽出結果の加工の記述を利用者が容易に行えるシステムを開発することが挙げられる。

文 献

- [1] M. Rafanelli, A. Shoshani, "STORM: A statistical Object Representation Model," 5th International Conference on Statistical and Scientific Database Management (SSDBM), pp.14-29, 1990.
- [2] S.P.Ghosh, "Statistical Relational Databases: Normal Forms," IEEE Trans. Knowledge and Data Eng., 3,

pp.55-64, 1991.

- [3] F.Ferri, P.Grinfoni, L.Meo-Evoli et al., "ADAMS: Aggregate data management system for epidemiologists and health-care managers," Computer Methods and Programs in Biomedicine, 40, pp.43-53, 1993.
- [4] H. J. Lenz, A. Shoshani, "Summarizability in OLAP and Statistical Data Bases," 9th International Conference on Statistical and Scientific Database Management (SSDBM), pp.132-143, 1997.
- [5] M. Okada, M. Takaba, S. Kaihara, M. Okada, "Formal Representation of Summary Tables for Health Care Statistical Database Management," Computers and Biomedical Research, 31, pp.426-450, 1998.
- [6] X. Wang, "Tabular Abstraction, Editing, and Formatting," Ph.D. thesis, University of Waterloo, 1996.