多次元データ向けデータマイニング手法の開発

2R－2

牧　秀行　芦田仁史　鮎川江里香　前田　章

（株）日立製作所システム開発研究所

1　はじめに

企業戦略における意思決定のスピードアップのために、エンドユーザコンピューティングが重要視されている。これを効率良く機能させるためには、データ分析の専門家ではないユーザが分析作業を容易に進めるための支援システムが必要である。

そこで、データの中から特徴的なデータ分布を自動的に発見し、着目すべき視点をユーザに提示するデータマイニング手法を開発した。

2　多次元データ分析

ＲＤＢ（リレーショナルデータベース）の検索結果は図１に示すような、テーブルのイメージでとらえることができる。各行は１つの対象物に対応し、各列は対象物の属性に対応する。図１は商品の売上記録の例である。対象物は「１回の売上」であり、「購入年月日」、「商品種別」、「店舗所在地域」、「売上金額」などがその属性となる。

購入年月日	商品種別	店舗所在地域	売上金額
19960602	商品Ａ	地域１	18000
19960602	商品Ｂ	地域２	32600
19960603	商品Ｂ	地域３	52600
19960603	商品Ｃ	地域２	22000
⋮	⋮	⋮	⋮

図１：ＲＤＢの検索結果

通常、ＲＤＢの検索結果は対象物の一覧表だが、データ分析の際には、これらを集計して得られた結果を用いるのが普通である。集計結果は、例えば図２に示す表のようになる。この表はＲＤＢの検索結果であるテーブルを元に、商品売上高を地域別、年別に集計したものだが、各地域における売上高の推移、他地域との比較を知りたいときには、このような集計表が使われるであろう。

Development of Data Mining Method for Multidimensional Data

Hideyuki MAKI, Hitoshi ASHIDA, Erika AYUKAWA, Akira MAEDA

Systems Development Laboratory, Hitachi, Ltd.

	1994年	1995年	1996年	合計
地域１	340000	480000	530000	1350000
地域２	120000	140000	350000	610000
地域３	810000	690000	940000	2440000
合計	1270000	1310000	1820000	4400000

図２：集計表の例

図２は、ＲＤＢの検索結果であるテーブルの「地域」と「年」を軸とする２次元空間の格子点上に、「売上金額」の値の集計値を置いた形になっている。ここで、軸として用いられている項目（この場合は「地域」と「年」）を「カテゴリ項目」、集計の対象になっている項目（この場合は「売上金額」）を「対象項目」と呼ぶことにする。図２は２次元の表だが、一般に、ＲＤＢの検索結果であるテーブル中の１つの項目を対象項目とし、その他の複数の項目をカテゴリ項目とすると、カテゴリ項目を直交する軸とした多次元空間の格子点に対象項目の集計値を置いたデータ構造を作ることができ、図２の集計表は、この多次元データ構造の、２次元への写像、あるいは断面と見ることができる。また、カテゴリ項目同士が直交しない場合もある。例えば、「都道府県」という項目と「都市名」という項目を考えると、これらは直交せず、階層をなすと考える方が自然である。

このように、分析対象のデータを多次元空間とその空間上の点としてとらえ、低次元空間への写像、断面によってデータの集計値を得るという考え方をここでは「多次元データ分析」と呼ぶことにする。また、これら低次元の写像、断面を「ビュー」と呼ぶことにする。多次元データ構造から任意の低次元への写像、断面を得る手段があれば、低次元空間を構成するカテゴリ項目を取り替えることにより、様々な軸に沿ったビューを得ることができる。

3　ビューの探索

多次元データ分析では、様々なビューからデータを観察することができるが、カテゴリ項目数が多くなると、着目すべきビューを探し出すことが困難である。そこで、多次元データ構造の中から、有用なビューを

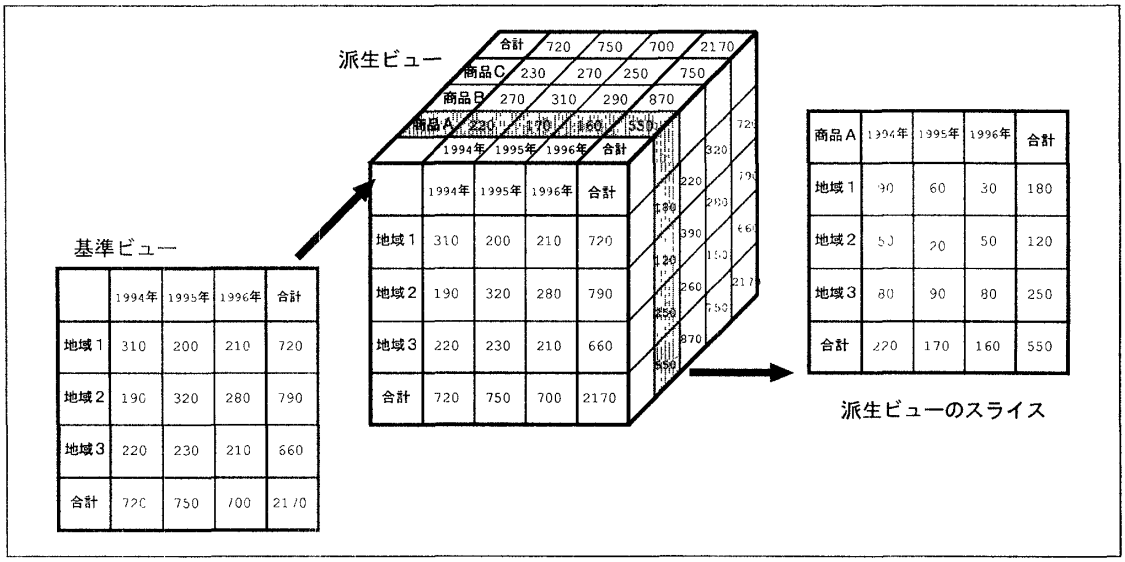

図 3: 派生ビューのスライス

自動的に発見するデータマイニング手法を開発した。その手順を説明する。

まず、ユーザはデータマイニングの基準となるビューを設定する。ユーザは何らかの興味を持ってデータを観察すると思われるので、最初の段階として、何らかのビューを自ら設定することは自然であると思われる。

データマイニング手法においては、ユーザが設定した基準ビューにカテゴリ項目を1つ加え、次元の数が基準ビューより1つ多いビューを生成する。これを派生ビューと呼ぶことにする。派生ビューは、すなわち、基準ビューよりも一段階詳細な集計表である。次に、この派生ビューにおいて、基準ビューに追加したカテゴリ項目についてのスライス（断面）を生成する（図3）。そして、以下に説明する評価基準にしたがって、このスライスの評価値を算出する。

4 ビューの評価

データマイニングの目的は、特徴的なデータ分布を示す切口を自動的に発見することであり、容易に推測できるようなデータ分布よりも、予想外のデータ分布を発見した方が、ユーザにとって役に立つはずである。そこで、「期待値からの乖離の大きいデータを持つビューは有用である」との方針にしたがって、ビューの有用さの評価基準を定義することにする。

派生ビューのスライスは、基準ビューと同じ次元を持つ。ここで、派生ビューのスライスにおける、格子点上の期待値を次の式によって求める。

$$E_i = V_i \cdot \frac{S_D}{S_B} \qquad (1)$$

ただし、E_i は派生ビューのスライスにおける格子点 i の期待値、V_i は基準ビューにおける格子点 i の実現値、S_D、S_B はそれぞれ、派生ビューのスライス、基準ビューにおける格子点上の値の総和である。そして、派生ビューのスライスの評価尺度を、次の式で定義する。

$$E = \sum_i \left(\frac{v_i - E_i}{E_i} \right)^2 \qquad (2)$$

ただし、E は派生ビューのスライスの評価値、v_i、E_i はそれぞれ、派生ビューのスライスにおける格子点 i の実現値、期待値である。

データマイニングでは、派生ビューのスライスを次々に探索し、上記評価値にしたがってそれらの順位づけを行い、上位のものをユーザに提示する。

5 おわりに

種々の項目を軸とした集計表を用いるデータ分析は、従来からごく一般的に行われてきた方法であるが、近年のデータ量の増加、意思決定のスピードアップの要求により、多次元データ分析支援システムは特にビジネス系データ分析で、今後さらに重要になると考えられる。多次元データ向けデータマイニング技術により、分析の効率、精度向上が期待できる。

参考文献

[1] 芦田、前田、高橋：
データマイニングにおける特徴的ルール生成方式、
情報処理学会第５０回全国大会 3-19