2005－DBS－137（Ⅱ）（42）
2005／7／14

多重比率規則抽出のためのデータ分析手法

濱本 雅史 †　　北川 博之 †,††　　Christos Faloutsos †††

概要 各属性間における属性値の相関関係を表した比率は比率規則と呼ばれ、データの性質の分析や値の予測など幅広い応用がなされる。データ中には一般に複数の比率規則が含まれるが、これらを総じて多重比率規則と定義する。多重比率規則の各比率規則をデータから抽出する手法として、本稿では比率規則におけるアプリオリ特性を考慮する。これにより多数の属性で構成される比率規則もすべて２属性からなる比率規則から導出することができる。２属性の比率規則を抽出する手法として、バケットと呼ぶ小さな領域にデータを分割し、そのヒストグラムを利用する手法を提案する。この提案手法について人工データを用いた実験を行い、その有効性を確認する。

Data Analysis Method for Extracting Multiple Ratio Rules

Masafumi Hamamoto †　　Hiroyuki Kitagawa †,††　　Christos Faloutsos †††

Abstract Ratio Rules are correlation among attribute values, and are applicable to data analysis, value estimation, and so on. Many Ratio Rules are generally included in data; we call them Multiple Ratio Rules. To extract each Ratio Rule in Multiple Ratio Rules, we consider the Apriori property for Ratio Rules. This enables us to extract any Ratio Rules by using two dimentional Ratio Rules. To extract two dimentional Ratio Rules, we divide tuples into small areas called buckets and extract Ratio Rules from the histogram, the number of tuples in buckets. We examine our proposed method using synthetic data and validate its usefulness.

1　はじめに

大量のデータから重要な情報を抽出するため、多数のデータマイニング手法が近年研究されている。本研究では特に、定量的なデータに対し数値属性間の相関を分析することに焦点を当てる。具体例を挙げると、商品の購買履歴情報から顧客は“パン”と“バター”に５：２の割合で金を費やす、といった情報を発見する問題を考える。また異なる問題例としては、野球選手の成績情報から“打数”と“ヒット数”は４：１である、という情報を発見することが挙げられる。

このように属性間における、属性値の典型的な割合を表したものは比率規則と呼ばれる [4]。比率規則は単にデータの内容を理解するだけでなく、欠損値の埋め合わせ、予測、外れ値検出、可視化など様々な応用が可能である。

データ全体の特徴が単一の比率規則で表すことができるとは限らない。具体例として以下のような状況が考えられる。

- “パン”と“バター”について５：２で金を費やすグループと２：３で費やすグループがある
- “パン”と“バター”は５：２で売れる一方、“パン”と“ジャム”と“マーガリン”も３：２：１で売れる

本稿ではデータ中に含まれる複数の比率規則を総称して多重比率規則と定義する。多重比率規則を発見するために、比率規則におけるアプリオリ特性を考える。この文脈でのアプリオリ特性とは、ある属性集合 R からなる比率規則がある場合、必ず R の真部分集合 S からなる比率規則が存在するということである。上記した例を用いると、“パン”:“ジャム”:“マーガリン”=３：２：１という比率規則が存在する場合、必ず“パン”:“ジャム”=３：２、“パン”:“マーガリン”=３：１、“ジャム”:“マーガリン”=２：１、という３種類の比率規則が存在するということである。

† 筑波大学 システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba
†† 筑波大学 計算科学研究センター
Center for Computational Sciences, University of Tsukuba
††† Carnegie Mellon University

このアプリオリ特性を用いると、どのような比率規則も 2 属性からなる比率規則から導くことができる。2 属性の比率規則を発見する手法として、データをバケットと呼ぶ小さな領域に分割し、バケット内のデータ数に関するヒストグラムを用いる手法を提案する。この手法はデータ数 n、属性数 m のデータに対し $O(nm^2)$ で比率規則を発見することができる。

本稿は以下のように構成される。2 章で関連研究について述べる。3 章では比率規則を説明し、本稿が特に扱う多重比率規則について述べる。4 章において提案手法を示し、5 章で人工データを用いた実験を行うことでその有効性を確かめる。最後にまとめと今後の課題について述べる。

2 関連研究

比率規則の発見に関する既存の手法は主に 2 種類ある。

一つは主成分分析を用いた手法であり [4][5]、全体の分布を最大にする軸である主成分ベクトルを比率規則とする。この手法は、全体を一つの主要な比率規則で表し、続いてそれを補足する比率規則でデータを表現する。従って本稿の多重比率規則のように主要な比率規則が複数存在する場合、個々の規則を発見することは難しい

別の手法として非負スパースコーディング [1] を用いた手法がある [2][3]。この研究では多重比率規則を対象としている。しかしこの手法では与えられたデータが非負の実数で表され、かつ比率規則が負の相関を持たないことを仮定している。従って “失業率” と “経済成長率” が $2:-1$ である (失業率が上がるほど経済成長率が下がる) といったような比率規則は得ることができない。また複数得られる比率規則に対し、各データがなるべく少数の規則と対応することが考慮されている。

このどちらの手法も、各データは比率規則の線形和によって表されるという仮定を元に、行列計算で比率規則を発見している。すなわち入力データを $X = [\mathbf{x_1}, \cdots, \mathbf{x_n}]$ とするとき、各列ベクトルが比率規則を表す行列 $R = [\mathbf{r_1}, \cdots, \mathbf{r_k}]$ と、データと比率規則の対応度合を表す行列 $V = [\mathbf{v_1}, \cdots, \mathbf{v_n}]$ により $X \approx RV$ となるよう表される。ここで $\mathbf{x}, \mathbf{r}, \mathbf{v}$ はそれぞれ列ベクトル、n はデータ数、k はユーザもしくはシステムが定める比率規則数を表す。

これらと比較して本稿で提案する手法は次のような特徴がある。まず既存の手法はいずれも与えられた全属性における比率規則を出力するのに対し、本提案手法では一部の属性のみで成り立っている比率規則が出力可能である。また外れ値が多少含まれていても、本提案手法で用いるヒストグラムへの影響は小さいので、提案手法は外れ値に対し堅牢であるといえる。

一方個々の手法と比較すると、Korn らの主成分分析を用いる手法に対して、本手法では局所的な関係が発見できると考えられる。主成分分析を用いた場合は全体の分散が重視されてしまうため、一部のデータのみ成り立つ比率規則が見つからない可能性がある。また Hu らの非負スパースコーディングを用いる手法とは異なり、本手法は負の相関を持ったデータに対しても正の相関のデータと全く同様に比率規則を得ることができる。

3 比率規則

本章では比率規則と多重比率規則について具体例とともに述べる。

まず本稿が扱うデータとして、表 1 のような構造をもったデータを想定している。データはタプルの集合であり、各タプルは 2 種類以上の属性値からなっている。また属性値に欠損値はないものとしている。

はじめに述べたように、比率規則は “パン” と “バター” といった属性間における属性値の典型的な比率を表したものである。具体例として図 1 のように 2 次元空間上に分布しているデータを考える。このデータ全体の分布を表すベクトルとして図の破線のようなベクトルが得られ、ベクトルの各成分から属性値の典型的な比率がわかる。したがって比率規則を発見することは、データの分布を表す特徴的なベクトルを発見することとみなすことができる。

ここで本稿では、比率規則を表すベクトルと一定の近さを持つデータは、その比率規則に従うと表現する。また逆にあるベクトルの周囲に一定量のデータが分布している場合、そのベクトルが表す比率規則が成り立っていると呼ぶ。

比率規則は単にデータの性質を捉えることができるだけでなく、以下のような応用に用いることができる [4]。

- 欠損値の埋め合わせ
- 属性値の予測
- 外れ値検出
- 可視化

欠損値の埋め合わせと属性値の予測は、すでにわかっている属性値から残りの属性を比率規則によって導くことである。また比率規則と各データの近さを定義することで、極端に比率規則から離れているデータを外れ値として検出す

顧客番号	パン (円)	バター (円)	マーガリン (円)	チーズ (円)
N0001	200	400	250	100
N0002	100	220	400	300
N0003	500	900	100	100
...	...	...	...	...

表 1: 本稿が想定するデータ例

ることができる。可視化に関して、各比率規則は上で述べたようにベクトルとして表される。従って２つないし３つの比率規則で張られる空間にデータを射影し、高次元のデータを２次元または３次元空間で表現できる。

一方で図 5 のように２方向に分布しているデータには、破線で表されるような２つの比率規則が含まれていると考えられる。このようにデータ中に含まれている比率規則を総称して多重比率規則と呼ぶ。多重比率規則の各比率規則を適切に発見することは、データの性質の分析を補助するだけでなく、上記した応用面でも有益である。

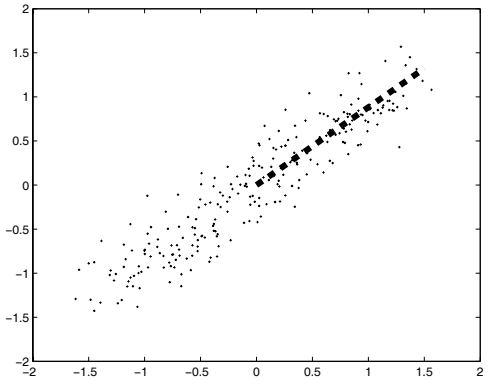

図 1: ２次元空間でのデータと比率規則

4 提案手法

本章では提案手法について述べるが、まず提案手法の本質である、比率規則におけるアプリオリ特性について述べる。

4.1 比率規則のアプリオリ特性

比率規則は２種類以上の属性によって構成される。しかし３属性以上で構成される比率規則は、その比率規則を構成する、任意の２属性からなる比率規則が必ず成り立たなくてはならない。

これを相関ルールマイニングのアプリオリ特性に当てはめると、次のように表すことができる。属性集合 P からなる比率規則 R_P と、属性集合 Q からなる比率規則 R_Q があり、$P \subset Q$ とする。データ集合全体に対する、比率規則 R に従うデータ数の割合を R のサポートと呼び $support(R)$ と表現すると、常に $support(R_P) \geq support(R_Q)$ が成り立つ。このときデータ全体に対して比率規則 R に従う最小のデータ数の割合を最小サポートと呼ぶ。また R_P が比率規則として成り立たないならば、P を含むすべての比率規則は成り立たない。

具体例として、３種類の属性 ‘A’,‘B’,‘C’ を持つデータに対する各比率規則の関係を図 2 に示す。この図において垂直方向は全タプルを表している。また比率規則は、一定量のデータが従っている規則のみを実線で示している。

まず ‘A’,‘B’ の２属性からなる比率規則 A:B=2:1 と ‘A’ と ‘C’ からなる比率規則 A:C=2:5 が成り立っているとする。このときどちらの比率規則にも従うデータは属性 ‘B’ と ‘C’ の属性値が B:C=1:5 となり、そのサポートが一定量以上ならば B:C=1:5 という比率規則が成り立つと言える。さらにこれら３つの比率規則に従うデータは必然的に A:B:C=2:1:5 という関係を持っている。この関係を持つデータは A:B=2:1 かつ A:C=2:5 かつ B:C=1:5 であることが必要十分条件である。よって比率規則 A:B:C=2:1:5 は２属性の比率規則から導くことができる。

一方比率規則 A:B=3:5 と A:C=1:2 が成り立っているとする。同時に２つの比率規則に従うデータについて比率規則 B:C=5:6 のサポートが最小サポートより小さい場合、A:B:C=3:5:6 のサポートはアプリオリ特性より最小サポート以下であるので比率規則とはならない。

４属性以上の比率規則についても、以上の性質が成り立つことは容易に理解できる。

4.2 提案手法

前節で述べたように３属性以上の比率規則は２属性の比率規則から導くことができる。従ってまず２属性の比率規則を発見する手法を述べ、次に３属性以上の比率規則を導く手法を述べる。

ただし問題を簡単化するため、各タプルは原点を中心にして分布していることを仮定する。

A:B=2:1 かつ A:C=2:5 となるデータ数が少ない場合でも、B:C=1:5 となる比率規則が存在する可能性がある。これは B:C=1:5 という関係において属性 ‘A’ の値は任意だからである。

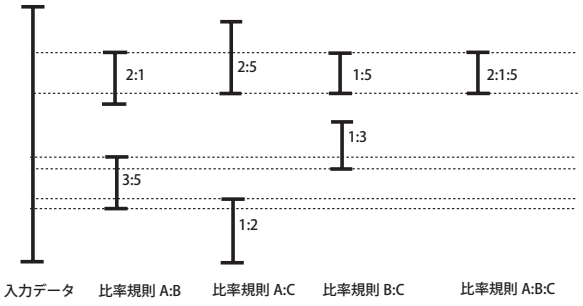

図 2: 入力データと各比率規則の関係

4.2.1 2 属性の比率規則

多重比率規則中の各比率規則の周囲には多数のデータが分布し密になっている。そこで点の分布のヒストグラムを作成し、最小サポート以上のデータが含まれている方向が求める比率規則と考える。単純な手法として、中心から一定の角度ごとにデータ数を数え上げる手法を用いる。ここで中心から一定角度の領域をバケットと呼び、定めた一定角度をバケット幅と呼ぶことにする。データが密な方向を見つけるには、各バケット内のデータ数をカウントし、適当に定めた最小サポート以上のデータを含むバケットを見つければよい。この結果得られたバケットに対応する比率規則を与えることで、比率規則を抽出することができる。

具体例を図 3 に示す。この例ではバケット幅を 22.5 度としている。この図では X 軸に対し 0〜22.5 度、45〜67.5 度、180〜202.5 度、225〜257.5 度の 4 バケットが多数のデータを含んでいる。しかし 180〜202.5 度、225〜257.5 度のバケットはそれぞれ 0〜22.5 度、45〜67.5 度のバケットと原点対象であるので同一のバケットとみなす。従ってこのデータからは、0〜22.5 度と 45〜67.5 度の 2 バケットから比率規則が与えられる。

一方バケット幅が狭い場合、隣接している複数のバケットのデータ数が最小サポートを超えることがある。これは本来わずかな差であるにも関わらず、それぞれ異なる比率規則として区別してしまうためである。この場合個々のバケットに対応する比率規則を出力すると比率規則数が膨大になることが考えられる。これを防ぐため、ヒストグラムが極大となるバケットで代表する方法や各バケットに対応する比率規則の平均を求める方法が考えられる。以下では前者の方法を用いることにする。

アルゴリズムを図 4 に示す。ここで入力されるバケット幅 θ と最小サポート δ はそれぞれ $0 < \theta < 90$(度), $0 < \delta < 1$ とする。まず (1)(2) においてデータをバケットに分割しヒストグラムを作成する。(3) では最小サポートを満たすバケットを抽出する。(4) は上で述べたように、

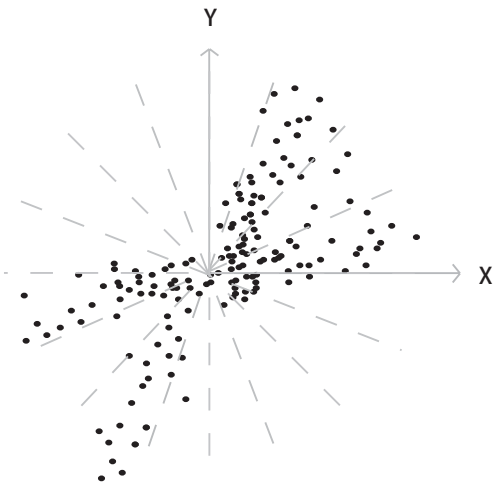

図 3: バケット幅 22.5 度の例

重要でない比率規則を捨てる作業である。

データ数が n 個で属性数が m 種類の場合、2 属性の組み合わせは ${}_mC_2$ 通りである。この各組み合わせについて $O(n)$ で比率規則を抽出することができるので、2 属性からなる比率規則は $O(nm^2)$ で抽出が可能である。実際には $n \gg m$ であることが想定されるので、計算量としては大きくないと考えられる。

4.2.2 3 属性以上の比率規則

4.1 節で述べたように、3 属性以上の比率規則は 2 属性の比率規則より導くことが可能である。p 属性の比率規則 $a_1 : a_2 : \cdots : a_p = b_1 : b_2 : \cdots : b_p$ が抽出されている場合を考える。これに属性 a_{p+1} を加えた $p+1$ 属性からなる比率規則 $a_1 : a_2 : \cdots : a_{p+1} = b_1 : b_2 : \cdots : b_{p+1}$ が成り立つ必要十分条件は、$i = 1, \cdots, p$ について $a_i : a_{p+1} = b_i : b_{p+1}$ が成り立つことである。つまり 2 属性の比率規則がすでに抽出されていれば、データを再走査することなく 3 属性以上の比率規則を導くことができる。

5 実験

本章では人工データを用いた実験を行い、提案手法の有効性を検討する。以下比率規則は、各比率を構成する値の二乗和が 1 となるように表現する。また各属性は $a_1, a_2, \cdots$ と表す。

各バケットに対応する比率規則は、そのバケットの二等分線を表す大きさ 1 のベクトルとした。

入力：$x_1, \cdots, x_n$(入力データ), θ(バケット幅), δ (最小サポート)

(1) $x_1, \cdots, x_n$ を適当なバケットに挿入

(2) 各バケット内のデータ数をカウント

(3) $n\delta$ 個以上のデータを持つバケット群を抽出

(4) 連続して抽出されたバケット群に関して

最も多くのデータを含むバケット以外を削除

(5) 抽出された各バケットに対応する比率規則を出力

図 4: ２属性の比率規則抽出アルゴリズム

5.1 人工データの生成手法

本章で用いる人工データを生成する際、次のような仮定を元にした。

- 各タプルはいずれか１つの比率規則のみに従う。ただし３属性以上の比率規則に関しては、その比率規則およびそれを構成する各比率規則に従う。
- 各データは、どれだけ比率規則に従うかという強度と、データの振れを表すノイズの２つで構成される。

２番目について、強度・ノイズともに平均0の正規分布に従うと仮定した。強度のノイズは各実験で適当に変化させ、ノイズの分散は 0.4 と固定した。

5.2 実験１

まず提案手法の妥当性やパラメータの影響を調べるため、簡単なデータで実験を行う。用いたデータは２次元データで、表2で表される２つの比率規則が含まれる場合を想定している。各比率規則に従うデータ数は１０００個とし、比率規則１の強度の分散は 1.5、比率規則２の強度の分散は２と設定した。

以上のデータに対しバケット幅を 5 で固定させ、最小サポートを 0.05,0.07,0.03 の３種類設定したときの結果が表 3 から 5 である。なお最小サポート 0.08 以上の場合は最小サポートを超えるバケットは存在しなかった。またバケット幅が５度の場合における、入力データと抽出された規則の関係を図 5 に示し、各バケットのデータ数のヒストグラムを図 6 に示した。

最小サポート 0.05 の場合は比率規則の数も正しく推定できており、図 5 からもわかるとおり比率規則自体も妥当であることがわかる。一方表 4 と 5 から、最小サポートが高すぎる場合には得られる比率規則が少なくなり、最小サポートが低すぎるとそれほど妥当とは言えない比率規則が得られてしまうことがわかった。図 6 を見ると、比率規則１に比べ比率規則２のほうがよりデータが集中していることがわかる。これは分散が異なり、かつノイズが同じである影響だと考えられる。つまりデータが中心に近いほど、含まれるバケットがノイズにより変化してしまうということである。

次にバケット幅を 10 度と２度にし、最小サポートをそれぞれ 0.05,0.03 に設定した場合の結果が、それぞれ表 6 と表 7 である。バケット幅が１０度のとき、比率規則１はバケット幅５度のときよりもはじめに設定した比率規則に近い値を出力している。しかし比率規則２については、バケット幅５度のときより離れてしまっている。これはバケット幅を固定長にいる影響だと考えられる。一方バケット幅が２度の場合、最小サポートを下げても比率規則が１つしか得られないだけでなく、比率規則自体も元の規則から多少離れてしまっている。これは極端にバケットが狭すぎるために全体の傾向をとらえることが難しくなってしまった影響だと考えられる。この結果から、より正確な値を得ようとしてバケット幅を極端に狭くすると、逆に比率規則の妥当性を悪化させる可能性があることが示されている。

	比率規則１	比率規則２
a_1	0.9701	0.3714
a_2	0.2425	0.9285

表 2: 設定した比率規則。整数で表すと比率規則１は４：１，比率規則２は２：５となる

5.3 実験２

次に３属性以上の比率規則を発見できるか実験した。実験データとして、表 8 にあるような４属性の比率規則に従う４次元データを用

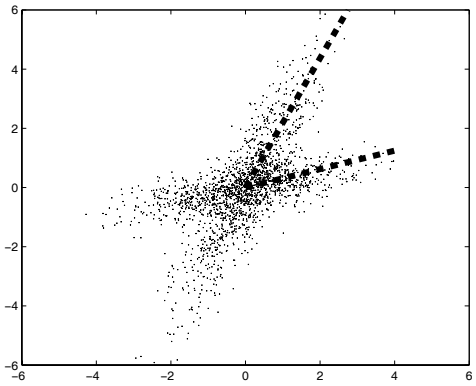

図 5: 入力データとバケット幅 5 度、最小サポート 0.05 で抽出された比率規則。破線が抽出された比率規則を表す

	比率規則 1	比率規則 2
a_1	0.9537	0.3827
a_2	0.3007	0.9239

表 3: バケット幅 5 度、最小サポート 0.05 で抽出された全比率規則

いた。各比率規則に従うデータの強度の分散は 3 とし、データ数は各１０００個とした。このデータに対しバケット幅を 5 度、最小サポートを 0.04 と設定した。

実験結果は表 9 の通りである。比率規則１では a_1, a_4 がそれぞれ過小・過大評価されているものの、各属性の大小関係は正しく捉えられていることがわかる。一方比率規則２では大小関係および各成分がほぼ正しく出力されていることがわかる。

以上から、提案手法は３属性以上の比率規則も妥当な結果を得られることがわかる。

6 おわりに

本稿ではデータ中の多重比率規則を抽出するための手法として、データをバケットに分割し最小サポート以上のバケットに対応する比率規則を与える手法を提案した。また比率規則におけるアプリオリ特性を考慮することで、３属性以上からなる比率規則を効率的に導出する手法を提案した。この提案手法に対し人工データを用いて実験を行い、妥当な結果が得られることがわかった。

今後の課題として、以下のようなことが挙げられる。

	比率規則 1
a_1	0.3827
a_2	0.9239

表 4: バケット幅 5 度、最小サポート 0.07 で抽出された全比率規則

	比率規則 1	比率規則 2	比率規則 3
a_1	0.9537	0.7934	0.3827
a_2	0.3007	0.6088	0.9239

表 5: バケット幅 5 度、最小サポート 0.03 で抽出された全比率規則

- 最小サポートやバケット幅の決定手法
- 動的なバケット幅を持つ手法の検討
- 偏りがあるデータへの対処手法
- より効率的な比率規則抽出手法の検討
- 実データへの適用と評価

１点目と２点目に関しては、データの分布の疎密を考慮して最小サポートやバケット幅を変える手法が考えられる。また最小サポートを使わない手法として Korn らの手法 [4] で行われているように、主成分分析を行い総分散の８から９割程度の主成分数を比率規則数として用いる手法が考えられる。

３点目の例として図 7 にあるような分布のデータが考えられる。このデータでは x,y 軸とも平均値が原点になるため、単純に本提案手法を適用することはできない。しかし左右の各クラスタごとにデータを分割し比率規則を抽出するアプローチが考えられる。

４点目に関しては 4.1 節および図 2 で示したように、データを再走査しなくとも比率規則 A:B と A:C から B:C を導くことができる場合がある。そのためには A:B かつ A:C に従うデータを検索する必要がある。したがって各データと比率規則を効率的に管理するデータ構造を検討することでより高速に比率規則が抽出できると考えられる。

	比率規則 1	比率規則 2
a_1	0.9659	0.4226
a_2	0.2588	0.9063

表 6: バケット幅 10 度、最小サポート 0.05 で抽出された全比率規則

	比率規則 1
a_1	0.3584
a_2	0.9336

表 7: バケット幅 2 度、最小サポート 0.03 で抽出された全比率規則

謝辞

本研究の一部は、科学研究費補助金基盤研究 (B)(#15300027)、特定領域研究 (2)(#16016205) による。

参考文献

[1] P. Hoyer. Non-Negative Sparse Coding. *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Martigny, Switzerland, pp. 557–565, 2002.

[2] C. Hu, B. Zhang, S. Yan, Q. Yang, J. Yan, Z. Chen, and W-Y. Ma. Mining Ratio Rules Via Principal Sparse Non-Negative Matrix Factorization. *Proc. 4th IEEE International Conference on Data Mining*, Brighton, U.K., pp. 407-410, 2004.

[3] C. Hu, Y. Wang, B. Zhang, Q. Yang, Q. Wang, J. Zhou, R. He, and Y. Yan. Mining Quantitative Associations in Large Database. *Proc. 7th Asia-Pacific Web Conference*, Shanghai, China, pp. 405-416, 2005.

[4] F. Korn, A. Labrinidis, Y. Kotidis, and C. Faloutsos. Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining. *Proc. 24th International Conference on Very Large Data Bases*, New York, pp. 582–593, 1998.

[5] F. Korn, A. Labrinidis, Y. Kotidis, and C. Faloutsos. Quantifiable Data Mining Using Ratio Rules. *VLDB Journal*, vol. 8, pp. 254–266, 2000.

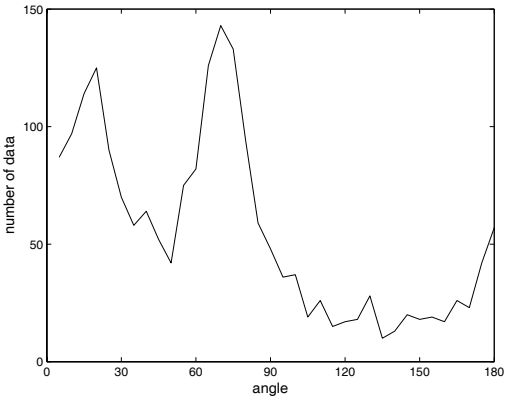

図 6: バケット中のデータ数に関するヒストグラム (バケット幅 5 度)

	比率規則 1	比率規則 2
a_1	0.1826	0.7303
a_2	0.3651	-0.5477
a_3	0.5477	0.3651
a_4	0.7303	-0.1826

表 8: 実験 2 で設定した比率規則。それぞれ整数で表すと 1 : 2 : 3 : 4, 4 : -3 : 2 : -1 となる

	比率規則 1	比率規則 2
a_1	0.1158	0.7238
a_2	0.2795	-0.5554
a_3	0.3672	0.3767
a_4	0.8796	-0.1604

表 9: 実験 2 :実験結果

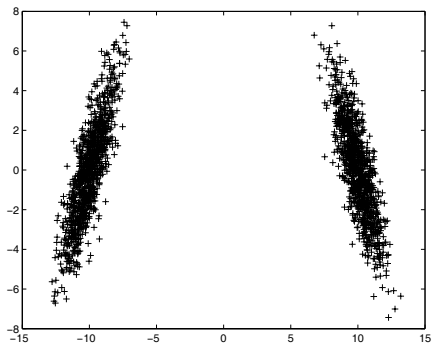

図 7: 偏りがあるデータ例