サポートと確信度をもとにした比率規則による線形関係抽出

濱本雅史[†] 北川博之[†,††]

数値属性を持つデータから得られる線形関係は，欠損値の補完，予測，外れ値検出など多数の応用が可能であり，その抽出は重要な技術課題である．本論文では線形関係を表した比率規則の抽出手法として，相関ルールマイニングで用いられるサポートと確信度の概念を取り入れた手法を提案する．線形回帰および既存の比率規則抽出手法では，線形関係を直線や超平面として表すため，表現可能な線形関係に制限があり，また同一のデータより得られる結果はユーザの興味によらず一定である．本論文では線分とその周辺領域内のデータが満たす性質として比率規則を定式化し，この定義をもとにサポートと確信度の概念を導入することで既存の手法の問題を解決する．提案手法はユーザより与えられた最小サポートと最小確信度を満たし，かつサポートまたは確信度を最大とする比率規則をタプル数に対し線形時間で抽出する．この提案手法の拡張としてクラスタリングと組み合わせることで，局所性を持った比率規則を抽出する手法も加えて提案する．人工データと実データを用いた実験で，提案手法がユーザの意向に応じた結果を出力することを示す．

Extracting Linear Relationships by Ratio Rules Based on Support and Confidence

Masafumi Hamamoto[†] and Hiroyuki Kitagawa[†,††]

Extracting linear relationships among numeric attributes is an important problem because it is applicable to filling in missing attribute values, forecasting values, detecting outliners, and related issues. This paper proposes a method to extract Ratio Rules, which represent linear relationships among numeric attributes, with support and confidence factors in analogy to association rule mining. Linear regression and existing Ratio Rule mining techniques are concerned with linear relationship extraction. However, their expressive power is limited since they represent a linear relationship as a line or a hyperplane. Moreover, they are not able to reflect the user's intention. In this paper we formulate a Ratio Rule as a line segment and its neighborhood, and then solve problems in existing methods by introducing support and confidence concepts. Our proposed method extracts Ratio Rules maximizing support or confidence, which satisfy the minimum support and confidence given by the user, in linear time for the number of tuples. We also propose a method to extract Local Ratio Rules, which hold in local areas, by combining with a clustering method. Experimental results for synthetic and real data show our proposed method works well.

1. はじめに

近年，大量のデータから重要な情報を抽出するデータマイニング手法として様々なものが検討されている．たとえば，相関ルールマイニング，クラスタリング，分類，テキストマイニング，時系列マイニング，Web マイニングなどがあげられる[6]．このような多種多様なデータマイニング手法のなかで，本論文では特に**比率規則**[11] を抽出する問題を考える．比率規則は属性間における属性値の典型的な線形関係を表したものである．

具体例として，**表 1** のような “身長” と “体重” の 2 つの数値属性を持つ学生データを考える．このデータをそれぞれの属性で張られる 2 次元空間へ射影したものが **図 1** である．この図から，黒い直線で表されたような線形関係を全体的な傾向として持っていることが分かる☆．比率規則は Korn らが示しているように[11]，単にデータを理解する補助になるだけでなく，欠損値の埋め合わせ，予測，外れ値検出，可視化など

† 筑波大学大学院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

☆ 例では 1 つの比率規則のみ示されているが，線形回帰とは異なり得られる規則は複数同時に存在しうる．

表 1 身長と体重の 2 属性を持つ学生データ例．いずれの属性も欠損値はないものとする
Table 1 Students data with height and weight attributes. Assume both attributes have no missing value.

学生 ID	身長（cm）	体重（kg）
S0001	157	51.1
S0002	174	68.0
S0003	164	60.7
...	...	...

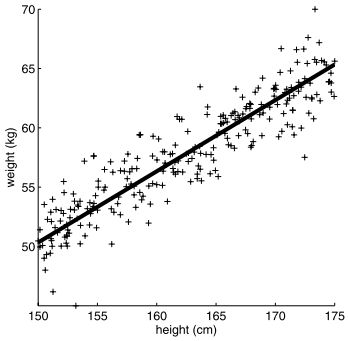

図 1 表 1 のデータに対する比率規則の例．実線が比率規則を表す
Fig. 1 Ratio Rule for Table 1. Black solid line represents a Ratio Rule.

様々な応用が可能である．

既存の比率規則抽出手法として，各タプルが複数の比率規則の線形結合で表されると考え，行列計算を用いて比率規則をとらえる手法がある[10),11)]．いずれの手法も行列分解により得られる特徴ベクトルを比率規則として表している．それゆえ一部の区間でのみ成立するような線形関係などはとらえることが難しい．また得られた比率規則に対し各タプルが従うかどうかの判定はユーザにゆだねられており，与えたデータと得られた結果の対応関係をとらえにくいという問題もある．

たとえば**図 2** のように，2 種類の異なる線形関係が成り立っているとする．このようなデータの場合，Korn らにより提案された主成分分析を使う手法[11)]では図中の黒い直線で表された結果が比率規則として得られる．この比率規則はいずれの線形関係も直接的に表していないため，妥当とはいい難い．また，このデータは負の相関関係を持つので，Hu らにより提案された非負行列分解を使う手法[10)]は適用ができない．また，与えられたデータ中には $0 \leq X \leq 0.2$ および $0.7 \leq X \leq 1.0$ の区間には属性 X と属性 Y との間に単一の線形関係しか存在せず，ほとんどのタプルがその線形関係に従っているという有益な情報が含まれている．しかし，たとえ既存の手法で妥当な線形関係が得られても，得られた結果は任意の属性値で線形関係が成り立つことを仮定しているので，このように一部の属性値間でのみ成り立つ線形関係を表現することができない．

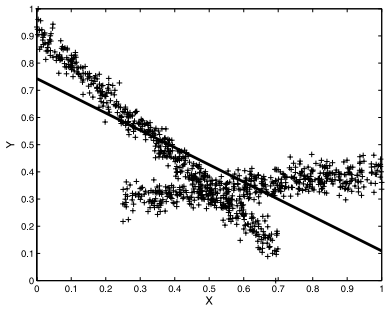

図 2 複数の比率規則が成り立つ例．実線は主成分分析を用いた手法で得られた比率規則を表す
Fig. 2 An example where multiple Ratio Rules exist. Black solid line represents a Ratio Rule extracted by Principal Component Analysis.

このような問題を解決するためのアイディアとして，我々は比率規則を直線ではなく線分およびその周辺領域内のデータが満たす性質として定義し，領域内に含まれているタプルはその比率規則に従うと定義する．この定義を行うことで，全体的に成り立つ線形関係だけでなく，部分的にのみ成り立つ線形関係を表現することが可能となる．さらに，比率規則とそれに従うタプルの対応づけを線形関係の抽出と同時に行うことができる．一方で，ユーザによって得るべき線形関係が変化しうることも考慮に入れる必要がある．ユーザが非常に強い線形関係が成り立つ部分（**図 3** の左図において黒点で示された部分）のみに興味がある場合や，多少他の線形関係が混在しても全体的に成り立つ線形関係（図 3 の右図において黒丸の部分と十字の部分の 2 種類）を知りたい場合などが考えられる．我々は相関ルールマイニングの諸概念を比率規則に導入し，ユーザがサポートや確信度の基準を与えることで，適当な比率規則を抽出することを提案する．

本論文では，まず比率規則の定式化を行い，相関ルールマイニングとの関係を示す．これをもとに，得るべき比率規則を**最適確信度比率規則・最適サポート比率規則**の 2 種類に分類する．これらを求める手法として，候補パラメータの絞り込み，1 次元数値属性相関ルールマイニング[5)]を用いた最適区間抽出，抽出された比率規則の統合の 3 フェーズからなる手法を提案する．この手法は入力タプル数に対して線形の時間で比率規則を求めることが可能である．

一方，本手法では基準となる属性の値の分布によって得られる線形関係が大きく変化する場合がある．た

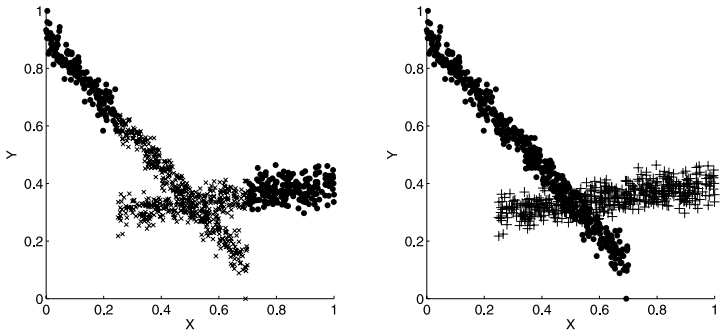

図 3 ユーザの興味により得るべき線形関係が異なる例．同じデータであっても左図では線形関係の強さに興味が置かれ，右図では線形関係全体に興味が置かれている

Fig. 3 An example where target linear relationships depend on the user's intention. For the same data, the left figure focuses on the strongness of linear relationships and the right figure focuses on the overall linear relationships.

とえば，データセット中にタプルが密集する領域が存在する場合，その領域を通る様々な線分に対応する比率規則が出力されてしまう可能性がある．逆に他のより疎な領域で線形関係が成立していても，密な領域を通らないために線形関係が抽出できない場合がある．この問題に対する解決策として，比率規則の概念を拡張し，与えられたデータセットを複数のクラスタに分割し，個々のクラスタ内の比率規則を**局所比率規則**として抽出することを考える．本論文では局所比率規則の抽出手法も加えて提案する．

本論文の以下は次のように構成される．2 章では本研究の関連研究について述べる．3 章では比率規則の定式化を行い，相関ルールマイニングの諸概念を導入する．4 章において比率規則抽出の提案手法を示す．5 章では比率規則の概念を拡張した局所比率規則について述べ，その抽出手法を提案する．6 章では人工データおよび実データを用いた実験を行い，手法の妥当性と性質を確かめる．7 章では議論として，既存の技術を組み合わせた手法との比較を行い，提案手法の特徴を示す．最後にまとめと今後の課題について述べる．

2. 関 連 研 究

数値データからの知識抽出に関する研究は，様々なものが行われている．特にデータベース的な観点では，相関ルールマイニング[1] と対応付けし，“身長 $\in [160, 165]$ ならば体重 $\in [55, 60]$ が成り立つ” といったような，数値属性に関する相関ルールを求める研究が行われている[4),5),13),14)]．Fukuda らの手法ではルールの前提部が 1 属性の場合[5] および 2 属性の場合[4] に，サポート，確信度，ゲインを最適にするルールを抽出する．Srikant らの手法[13] では最適性は持たないものの，任意の属性数に対して条件を満たすルールをすべて抽出する．

これに対し，“身長：体重 $= 3:2$” のように数値属性間で成り立つ増分の比率に注目したものが比率規則[11] である．比率規則は多次元空間上での直線として表すことができるため，前に示した数値属性の相関ルールよりも欠損値の埋め合わせ，予測，外れ値検出などの応用がしやすい利点がある．

比率規則の抽出手法として提案されているものとして以下の 2 手法がある．

(1) Korn らによる手法[11),12)]．この手法は主成分分析を用い，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して定義している．得られた比率規則は各属性の平均値を表す点を通る直線として表すことができる．アルゴリズムとしては，主成分ベクトルをまず計算し，その寄与率が一定以上の主成分ベクトルを比率規則として採用する．ただし主成分分析の意味を考慮すると，第 1 主成分に対応する比率規則が全体の主要な分布を表す．このとき各比率規則は直交するという制約を持っている．この手法は**図 4** のように複数の線形関係が混在する場合，第 1 主成分に対応する比率規則として図の直線のような結果が得られ，いずれの線形関係も直接的にとらえることができない．

(2) Hu らによる手法[9),10)]．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．このとき与えられた各タプルが非負値からなるベクトルの線形和として表されていると仮定し，そのベクトルを比率規則として，非負行列分解を用いて抽出する．この手法では各比率規則は互いに直交するという制約はないが，原点を通

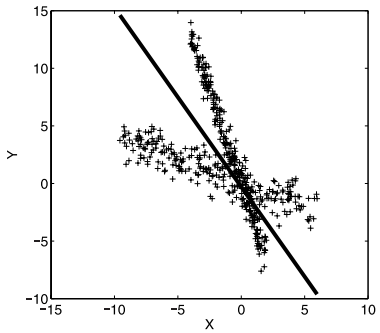

図 4 主成分分析を用いた手法ではうまくいかない例．直線は得られた比率規則を表す

Fig. 4 A example where a method using Principal Component Analysis fails. Solid line represents the extracted Ratio Rule.

るという制約を持っている．

また統計的な観点から，線形関係を抽出する問題は，回帰分析，主成分分析，独立成分分析のような多変量解析の対象ともなっている[7]．線形回帰分析では一般的に，属性値との誤差が最小になるような推定を行う直線を抽出する．しかしデータ中に複数の線形関係が存在することを想定していない．

これら既存の手法に対する本手法の特徴として，以下の点があげられる．

- 比率規則を線分およびその周辺領域内のデータが満たす性質として定義した点．既存の手法ではいずれも大域的に成り立つ関係のみが得られるのに対し，本提案手法では局所的に成り立つ線形関係を抽出することが可能となる．また，これによって，比率規則を抽出すると同時にそれに従うタプルも抽出することができる．
- 相関ルールマイニングと対応付けしてサポートと確信度の概念を導入した点．ユーザより与えられる最小サポートと最小確信度によって，ユーザの意図に沿った比率規則が抽出可能となる．
- 条件を満たす比率規則をすべて列挙可能である点．回帰分析も含め，既存の手法では得られる線形関係の数に制約を持つが，本手法では最小サポートと最小確信度を満たす比率規則をタプル数に対して線形の時間で列挙できる．
- 与えられたデータをクラスタに分割し，各クラスタ中の比率規則を抽出する拡張を行った点．全タプルを対象にするのではなくクラスタ内のタプルを抽出対象にすることで，タプルの分布による線形関係抽出に対する影響を抑えることができる．

3. 問題設定

本章では抽出すべき比率規則の定式化を行う．本論文で扱う比率規則は前章で述べた既存の研究と異なり，より一般性を持った定義となっている．また，相関ルールマイニングで用いられる概念を導入して，ユーザの意図に沿った比率規則の抽出手法を考える．

3.1 対象とするデータ

本論文が対象とするデータは，1 章の表 1 であげたように数値属性を持つタプルの集合である．ただし各属性には欠損値は存在しないと仮定する．

本論文では，2 種類の数値属性間における比率規則を抽出する問題を扱う．各属性値は連続な実数値を想定するが，本論文ではドメインが区間 $[-0.5, 0.5]$ となるよう正規化されているものとする．

以下，分析対象とする 2 属性を X，Y とし，それぞれの属性値を x，y（$-0.5 \leq x$，$y \leq 0.5$）と表現する．

3.2 比率規則の定義

比率規則は前章で述べたように，属性間の線形関係を表したものである．もし与えられたデータが全区間にわたり均一の線形関係を持つならば，2 属性 X，Y で張られる空間中の直線 $y = ax + b$（a，$b \in \mathcal{R}$）として比率規則を考えることが自然である．

しかし，この定義には 3 つの問題がある．1 点目は，知りたいことは多数のタプルが厳密な意味で直線 $y = ax + b$ 上に存在するということではなく，近似的にこのような線形関係が成立するということなので，この点を考慮に加える必要がある．2 点目はパラメータ a，b のとりうる値はどちらも区間 $(-\infty, \infty)$ における任意の実数であり，y 軸に平行あるいはそれに近い直線を表す場合，いずれのパラメータも無限大に発散してしまう．3 点目は，一般的には全区間にわたり線形関係が成り立つとは限らない点である．すなわち属性 X がある区間に含まれる場合のみ線形関係が成り立つ場合も考える必要がある．

そこで 1 点目の問題には，パラメータに対する**許容幅**を設定し，許容幅内の任意の直線上に存在するタプルは同一の比率規則に従うとする．2 点目の問題については，付録で説明した **Hough 変換**[8] により有限区間の変数へ変換を行う．Hough 変換を用いると直線 $y = ax + b$ は $\rho = x\cos\theta + y\sin\theta$（ただし $\rho = b\sin(\tan^{-1}(-1/a))$，$\theta = \tan^{-1}(-1/a)$）と表現される．属性値 x，y が区間 $[-0.5, 0.5]$ をとるよう正規化されているので，ρ，θ の値はそれぞれ有限の区間 $[0, \sqrt{2}/2]$，$[0, 2\pi]$ でおさえられる．3 点目の問題

については，比率規則を直線ではなく線分として表すことを行う．これは比率規則の定義中に，比率規則が成り立つ属性値の区間を示すことで行う．

以上の点をふまえ，比率規則を次のように定義する．

タプル $t(x_t, y_t)$（$x_t \in I, I \subseteq [-0.5, 0.5]$）**が以下の式を満たす値** ϵ_t，δ_t **を持つとき，** t **は比率規則** $RR_{x\in I}(\rho \pm \epsilon, \theta \pm \delta)$ **に従う．**

$$\rho + \epsilon_t = x_t \cos(\theta + \delta_t) + y_t \sin(\theta + \delta_t)$$

ただし $|\epsilon_t| \leq \epsilon, |\delta_t| \leq \delta$

この定義上，属性 X と Y は対称ではないことを注意しておく．以下では誤解のない限り，比率規則 $RR_{x\in I}(\rho \pm \epsilon, \theta \pm \delta)$ はパラメータを省略した形 $RR_I(\rho, \theta)$ として表現する．

3.3 比率規則の種類

比率規則 $RR_I(\rho, \theta)$ について，数値属性に対する相関ルールマイニング[5] と対応付けし，以下のような諸概念を定義する．

- 比率規則に対する**サポート**は $RR_I(\rho, \theta)$ に従うタプルの，全タプルに対する割合とし $support(RR_I(\rho, \theta))$ で表す．また区間 I に対するサポートは属性値 x が区間 I に含まれるタプルの，全タプルに対する割合とし $support(I)$ と表す．
- 比率規則 $RR_I(\rho, \theta)$ に対する**確信度**は $support(RR_I(\rho, \theta))$ の $support(I)$ に対する割合 $support(RR_I(\rho, \theta))/support(I)$ とし $conf(RR_I(\rho, \theta))$ と表す．
- 抽出される比率規則に対し，ユーザから与えられる最低限満たすべきサポートおよび確信度をそれぞれ**最小サポート**，**最小確信度**と呼ぶ．以下ではそれぞれ $minsup$，$minconf$ と表す．

これらの諸概念を用いて，次の 2 種類の比率規則を定義する．

- **最適確信度比率規則**：$support(I)$ が $minsup$ を満たし，かつ $conf(RR_I(\rho, \theta))$ が $minconf$ を満たしたうえで最大となるような比率規則 $RR_I(\rho, \theta)$．最大値を与える区間 I を最適確信度区間と呼ぶ．
- **最適サポート比率規則**：$conf(RR_I(\rho, \theta))$ が $minconf$ を満たし，かつ $support(I)$ が $minsup$ を満たしたうえで最大となるような比率規則 $RR_I(\rho, \theta)$．最大値を与える区間 I を最適サポート区間と呼ぶ．

最適確信度比率規則を抽出することは，一定数以上のタプルが比率規則に従うという条件の下，比率規則に従うタプルの割合が最大となる区間を発見する問題といえる．また最適サポート比率規則を抽出することは，一定割合のタプルが比率規則に従うという条件の下，なるべく多くのタプルが比率規則に従うような区間を発見する問題といえる．

以下の章では，この 2 種類の比率規則をまとめて最適比率規則と呼び，最適確信度区間と最適サポート区間をまとめて最適区間と呼ぶ．

4. 提 案 手 法

本章では 3 章にあげた最適比率規則を抽出する手法を提案する．3 章で述べた比率規則の定義において，ρ，θ はそれぞれ区間 $[0, \sqrt{2}/2]$，$[0, 2\pi]$ 内の任意の値をとる．しかし以下ではユーザより与えられた許容幅により，それぞれ 2ϵ，2δ 間隔の離散値 ρ_i，θ_j（$i = 1, \cdots, R$，$j = 1, \cdots, T$）として考える．すなわち比率規則 $RR_I(\rho_i, \theta_j)$ が，パラメータ ρ_i，θ_j とその許容幅内の全比率規則を代表することになる．

4.1 基本的なアルゴリズム

最適比率規則を求めようとする場合，一番の問題は最適区間を求めることにある．パラメータ (ρ_0, θ_0) を持つ比率規則に対する単純な最適区間抽出手法としては，各タプルについて比率規則 $RR_{[-0.5,0.5]}(\rho_0, \theta_0)$ に従うかどうかの判定を行い，その後考えうるすべての区間についてサポートと確信度を計算することで，条件を満たす区間を得ることができる．ただしこの場合考えうる区間の数は全タプル数を N とすると，最大で任意のタプルの組合せ数 $N(N-1)/2$ となるので現実的ではない．

本提案手法では，最適区間の抽出を 1 次元数値属性相関ルールマイニング[5] における最適確信度/サポート区間の抽出問題と同様に考える．いま数値属性 X について，X の定義域中における区間 $I = [s, t]$（$-0.5 \leq s \leq t \leq 0.5$）を考えたとき，条件 $X \in I$ を満たすならば条件 C を満たす，という規則が **1 次元数値属性相関ルール**と呼ばれ $(X \in I) \Rightarrow C$ と表記される．ここで “比率規則が成り立つかどうか” を条件 C と見なすことで，1 次元数値属性相関ルールマイニングの概念を比率規則の抽出に利用することができる．

1 次元数値属性相関ルールにおける最適区間抽出手法として，ここでは Fukuda らによる手法[5] を用いる．この手法は各タプルの属性 X がソートされており，かつ各タプルが条件に従うかどうかの判定がなさ

```
for each (ρ_i, θ_j) do
    for each タプル t do
        t が RR_[−0.5,0.5](ρ_i, θ_j) に従うか判定
    end
    最適区間 I を
    1 次元数値属性相関ルールマイニングで求める
    if I, RR_I(ρ_i, θ_j) がそれぞれ
    minsup, minconf を満たす then
        RR_I(ρ_i, θ_j) を出力
    end
end
```

図 5 比率規則を求める基本的なアルゴリズム
Fig. 5 A basic algorithm to generate Ratio Rules.

れているとき，最小サポート/確信度を満たす最適確信度/サポート区間を $O(n)$（n は入力タプル数）で求めることができる．本提案手法では入力データはすでに属性 X でソートされているものと仮定する．

基本的なアルゴリズムを**図 5** に示す．このアルゴリズムは $O(RTN)$ で実行可能である．

しかしこの基本的なアルゴリズムには 2 つの問題がある．1 つはすべての (ρ_i, θ_j) の組に対して毎回全タプルを読み込み，最適区間の抽出を行う点である．このアルゴリズムはほとんどのタプルが従わない候補についてもタプルの読み込みと区間の抽出を行う．そのため，パラメータ ρ および θ を細かく離散化した場合，その分実行時間が単調に増加する．もう 1 つの問題点は，本質的にはほぼ同一と見なせる比率規則が多数得られる可能性があることである．パラメータ ρ や θ がごくわずか異なるのみの比率規則には多数のタプルが共通して従うと考えられ，そのような比率規則群は統合する方が適切である．

この 2 つの問題を解決する方法として，最適区間の抽出と比率規則の出力処理（まとめて**比率規則生成フェーズ**と呼ぶ）の前後に，**枝刈りフェーズ**と**比率規則統合フェーズ**を用意する．以下ではこの 2 つのフェーズについて説明する．

4.2 枝刈りフェーズ

前節で述べたように，すべての比率規則について最適区間の抽出を行うことは非常に無駄が大きい．そこで，どのような区間をとっても条件を満たさない場合を考え，これを枝刈りによって除くことを行う．

最小サポートと最小確信度を満たす比率規則 $RR_I(\rho_i, \theta_j)$ が存在する場合，その比率規則に従うタプルの割合 $support(RR_I(\rho_i, \theta_j))$ は以下の式を満たす．

$$\begin{aligned} &support(RR_I(\rho_i, \theta_j)) \\ &\quad \equiv support(I) \times \\ &\qquad (support(RR_I(\rho_i, \theta_j))/support(I)) \\ &\quad \equiv support(I) \times conf(RR_I(\rho_i, \theta_j)) \end{aligned}$$

区間のサポート $support(I)$ の最小値は最小サポート $minsup$，比率規則の確信度 $conf(RR_I(\rho_i, \theta_j))$ の最小値は最小確信度 $minconf$ であるので，この式はその 2 つの積 $\alpha = minsup \times minconf$ 以上の割合のタプルが比率規則に従う必要があることを表す．

枝刈りフェーズでは $RR_{[-0.5,0.5]}(\rho_i, \theta_j)$ において α 以上の割合のタプルが従わないパラメータ (ρ_i, θ_j) を枝刈りする．具体的には，まず各タプルを通る直線 $\rho_i = x\cos\theta_j + y\sin\theta_j$ を列挙する．そして全タプルにおけるパラメータの組 (ρ_i, θ_j) のヒストグラムを作成する．このヒストグラムから α 以上の割合のタプルが従うパラメータを得る．各タプルについて，各 θ に対応する ρ は定数時間で計算可能であるので，全ヒストグラムは $O(TN)$ で作成できる．

ヒストグラムを作成する際，単に各パラメータ (ρ_i, θ_j) のカウンタを用意するだけでなく，各パラメータに従うタプルを記録する．これは各タプルに対して比率規則に従うかどうかの判定が必要だからである．ただし，θ を動かしてパラメータをカウントするのと同時にタプルを記録した場合，計 TN 個のエントリが必要となり，タプル数が多数のときにメモリ使用量が非常に大きくなる．したがって，はじめにパラメータのカウントのみを行い，その後再度タプルをはじめから読み，閾値以上カウントがあったパラメータに対してのみタプルを記録する．このとき入力データは属性 X でソートされていることを仮定し，タプルは X でソートされた順に記録される．

4.3 比率規則統合フェーズ

比率規則統合フェーズでは，本質的に類似した最適比率規則群を**比率規則集合**へ統合する．2 つの比率規則 $RR_{I_1}(\rho_i, \theta_j)$，$RR_{I_2}(\rho_k, \theta_l)$ に対する類似尺度としては，以下の式で表される **Jaccard 係数**を用いる．

$$\frac{|RR_{I_1}(\rho_i, \theta_j) \cap RR_{I_2}(\rho_k, \theta_l)|}{|RR_{I_1}(\rho_i, \theta_j) \cup RR_{I_2}(\rho_k, \theta_l)|}$$

ここで $|RR_I(\rho_i, \theta_j)|$ は，比率規則 $RR_I(\rho_i, \theta_j)$ に従うタプル数を表す．したがって類似度は，2 つの比率規則の両方に従うタプルの，いずれかの比率規則に従うタプルに対する割合である．この値が閾値以上のとき 2 つの比率規則は同一の比率規則集合に統合する．以下ではこの閾値を $minmerge$ と表記する．

この Jaccard 係数の分子項を単純に計算すると，$RR_{I_1}(\rho_i, \theta_j)$ に従うタプルと $RR_{I_2}(\rho_k, \theta_l)$ に従うタ

プルの全組合せだけチェックを行う必要がある．しかし枝刈りフェーズにおいて各比率規則に従うタプルは属性 X でソートされた順に記録されている．このことを利用すれば，分子項の計算は各比率規則に従うタプルを一度ずつ読むだけで完了できる．すなわちこのフェーズは，比率規則生成フェーズで生成された比率規則数を Q とすると，$O(Q^2N)$ で実行が可能である．

4.4 提案手法のまとめ

本提案手法は，枝刈り，比率規則生成，比率規則統合の 3 フェーズから構成され，最適比率規則集合を得る．いま全タプル数 N，パラメータ ρ，θ の各個数 R，T，枝刈りにより残るパラメータ組 (ρ, θ) の数 $P \leq RT$，最小確信度と最小サポートを満たす比率規則の数 $Q \leq P$ とすると，本手法における各フェーズの計算量はそれぞれ $O(TN)$，$O(PN)$，$O(Q^2N)$ で表される．T，P，Q の値はユーザにより与えられるパラメータ ϵ，δ，*minsup*，*minconf* により異なるが，タプル数 N についてはいずれのフェーズも線形時間で処理可能である．

5. 拡張：局所比率規則抽出

5.1 比率規則と局所比率規則

3 章で述べたサポートの定義において “全タプル” が何を表すのかはユーザの意図により変化しうる．考えられる意図としては，与えられたデータに含まれる全タプルであるか，クラスタなどのある部分集合内に含まれる全タプルであるか，という 2 種類がある．前者は大域的な情報であるが，後者はデータ中の局所的な情報を表しており，得られた比率規則が他の部分集合で成り立つとは限らない．このような比率規則を**局所比率規則**と呼ぶ．ここでこれまで述べてきた比率規則は局所比率規則の特殊な場合，すなわち部分集合が与えられた全タプルを表す場合である．サポートと確信度は，比率規則と局所比率規則では異なった値を考えることができる．以下では局所比率規則に関するサポートと確信度をそれぞれ**局所サポート**，**局所確信度**と呼び，ユーザより与えられる閾値を**最小局所サポート** *minlsup* および**最小局所確信度** *minlconf* と呼ぶ．

与えられたデータの分布によっては，局所比率規則を用いた方がデータの性質を上手くとらえられる場合がある．例として，タプルの分布が密なクラスタと疎なクラスタが共存する場合を考える．比率規則ではタプルが密なクラスタと疎なクラスタを同一の空間で扱うため，得られる結果が密なクラスタの振舞いに大きく影響される．そのため疎なクラスタで成り立つ線形関係がとらえにくく，逆に密なクラスタでは強い線形関係が成り立たないにもかかわらず比率規則として得られることが考えられる．一方で局所比率規則では，密なクラスタと疎なクラスタを分けて考えることで，各クラスタで成り立つ線形関係を個別にとらえることができる．

```
/* クラスタ生成フェーズ */
与えられたデータより
クラスタ {C_1, C_2, ...} を生成
for each |C_p| ≥ mincard × N do
    /* 枝刈りフェーズ */
    パラメータ組 (ρ, θ) の候補を枝刈り
    /* 比率規則生成フェーズ */
    for each 残りの候補 (ρ_i, θ_j) do
        最適局所サポート/
        最適局所確信度区間 I を抽出
        if LRR_I(ρ_i, θ_j) が minlsup および
        minlconf を満たす then
            LRR_I(ρ_i, θ_j) を生成
        end
    end
    /* 比率規則統合フェーズ */
    for each 類似した組
    LRR_{I_1}(ρ_i, θ_j), LRR_{I_2}(ρ_k, θ_l) do
        LRR_{I_1}(ρ_i, θ_j) と LRR_{I_2}(ρ_k, θ_l) を
        同じ集合 S_q に統合
    end
end
局所比率規則集合 {S_1, S_2, ...} を出力
```

図 6 局所比率規則抽出の提案手法
Fig. 6 Proposed method to extract Local Ratio Rules.

5.2 提 案 手 法

局所比率規則を抽出するための提案手法を**図 6** に示す．提案手法は 4 フェーズから構成される．まず局所比率規則を抽出するために与えられたデータよりクラスタを構築する**クラスタ生成フェーズ**を行う．得られた各クラスタに対しては，比率規則と同一のアルゴリズムで局所比率規則を抽出することができる．最終的に得られる結果は局所比率規則の集合である**局所比率規則集合**となる．クラスタ生成フェーズについて次の項で説明する．

5.2.1 クラスタ生成フェーズ

このフェーズでは与えられたデータをクラスタリングによってクラスタに分割する．どのようにクラスタを作成するかによって，最終的に得られる結果が異なってくる．

この後の 3 フェーズではクラスタリングの結果得られた各クラスタから局所比率規則を抽出する．ただしクラスタリングの結果として，非常にタプル数

が少ないクラスタが生成される可能性がある．極端な例をあげると，1 タプルからなるクラスタが生成された場合，その 1 タプルを通るすべての局所比率規則がサポート 1，確信度 1 で得られる．しかし得られた情報はユーザにとって有益とはいえない．そこでクラスタに含まれるタプル数に関して**最小濃度** $mincard$（$0 < mincard \leq 1$）を設定し，メンバ数を $mincard \times N$ 以上持つクラスタから局所比率規則の抽出を行う．ここで N は与えられたデータセット中の全タプル数を表す．

本提案手法はクラスタリング手法を特定するものではないが，6 章の実験では密度ベースのクラスタリング手法である OPTICS [2] を用いた．

6. 実　　験

本実験では，提案手法により得られる比率規則および局所比率規則の妥当性と，提案手法の処理時間に関する性質を検討する．ここでは人工データと 3 種類の実データを用いる．実データはいずれも UCI Machine Learning Repository ☆から入手可能である．

以下の実験では C 言語で実装されたアルゴリズムを用いた．実験環境として用いた計算機は，Pentium III Xeon 1.0 GHz を 2 基有し，メインメモリサイズは 2.0 GB である．

6.1 データの概要

6.1.1 人工データの概要

本実験で扱う人工データは，比率規則数を p 個とし，各比率規則に対して q 個のタプルを生成した．全タプル数は pq 個である．

ある 1 つの比率規則に従うタプルは以下のようにして生成した．

(1) パラメータ ρ，θ と区間 $I = [x_{min}, x_{max}]$ をランダムに生成．
(2) 区間 I 内で一様に分布するよう，属性値 x_i $(1 \leq i \leq q)$ を生成．
(3) 各 x_i に対し属性値 $y_i = (\rho - x_i \cos\theta)/\sin\theta$ を生成．
(4) 各 y_i に平均 0，分散 0.1 で正規分布するノイズ値を加える．
(5) x，y それぞれ区間 $[-0.5, 0.5]$ をとるよう正規化．

ここで各パラメータは $0 \leq \rho \leq 1$，$-\pi \leq \theta \leq \pi$，$0 \leq x_{min} \leq x_{max} \leq 1$ を満たし一様分布に従うよう生成する．

☆ http:///www.ics.uci.edu/∼mlearn/MLRepository.html

6.1.2 アワビデータ

このデータにはアワビの体長，身の重さ，性別などが記録されている．今回は連続値で表される 7 属性（Length，Diameter，Height，Whole weight，Viscera weight，Shell weight）のうち，Length と Shell weight の 2 属性を用いた．全タプル数は 4,177 個である．

6.1.3 ワインデータ

このデータは 3 つの異なる品種のワインについて，アルコールやリンゴ酸など 13 項目が調べられた化学分析データである．本実験では “Flavanoids” と “Proline” の 2 属性を用いた．全タプル数は 178 個である．

6.1.4 自動車データ

このデータには 1985 年にアメリカへ輸入された自動車に関する，価格・燃費・重量など計 26 項目の数値およびカテゴリデータが記録されている．本実験ではそのうち，高速道路の燃費とエンジンの圧縮率の 2 属性を対象とした．全タプル数は 205 個である．

6.2 妥当性の評価

まず各データについて妥当な比率規則が得られるか検討した．人工データを生成する際のパラメータ (p, q) には $(2, 500)$ と $(5, 2000)$ の 2 種類を与えた．前者は 1 章の図 2 で示された例のデータである．以下の各図において黒の点は各タプルを表し，濃いグレーの線分は得られた各比率規則の許容幅を含まない形 $(RR_{x \in I}(\rho \pm 0, \theta \pm 0))$，薄いグレーの領域は許容幅も含め得られた各比率規則が成り立つ領域を示す．薄いグレーの色の濃さは比率規則集合ごとに変えており，同一の比率規則集合に含まれる比率規則はすべて同じ濃度で表されている．

6.2.1 人工データ

図 7 は $(p, q) = (2, 500)$ の人工データと抽出された全比率規則集合を表す．左図は最適確信度比率規則，右図は最適サポート比率規則の結果である．パラメータには，$\epsilon = 0.0325$，$\delta = 0.0325$，$minsup = 0.2$，$minconf = 0.8$，$minmerge = 0.5$ を与えた．本実験では全部で 1,176 組の候補中 86 個のパラメータ組 (ρ, θ) が枝刈りフェーズで残った．

最適確信度比率規則および最適サポート比率規則とも，最終的に 3 個の比率規則からなる比率規則集合と 1 個の比率規則からなる比率規則集合の 2 つが得られた．前者は属性 X が -0.2 以下の部分，後者は 0.2 以上の部分で成り立つ線形関係をそれぞれ表している．特に最適確信度比率規則の結果は単一の線形関係のみ成り立つ領域を適当に抽出している．

もし比率規則統合フェーズがない場合，得られるす

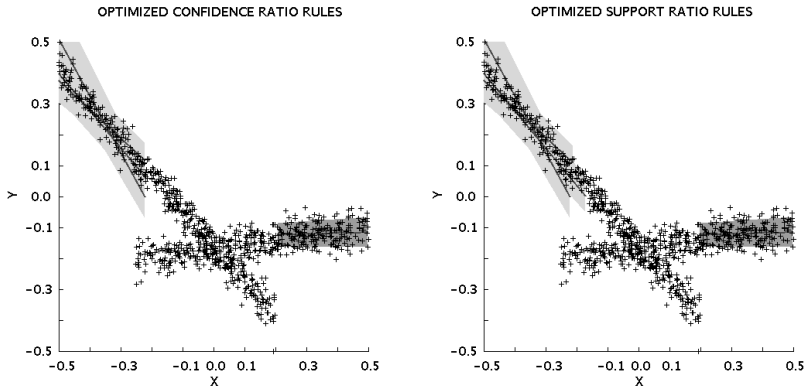

図 7 $(p, q) = (2, 500)$ の人工データに対する最適比率規則抽出結果．左図が最適確信度比率規則，右図が最適サポート比率規則であり，*minsup*，*minconf* はそれぞれ 0.2 と 0.8 である

Fig. 7 Extracted optimized Ratio Rules for synthetic data when $(p, q) = (2, 500)$. The left figure shows optimized confidence Ratio Rules, and the right figure shows optimized support Ratio Rules. *minsup* and *minconf* are 0.2 and 0.8, respectively.

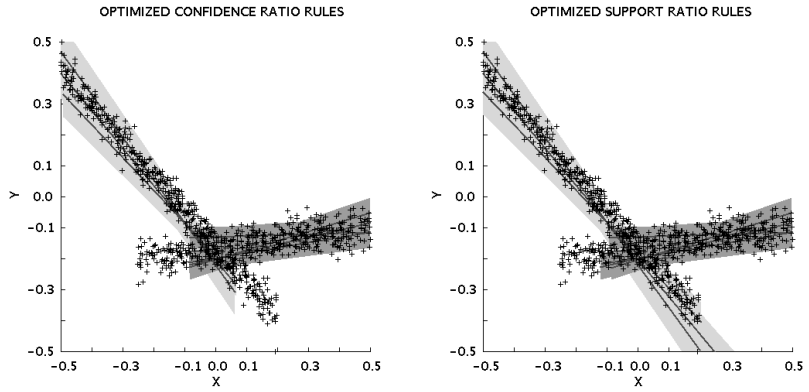

図 8 $(p, q) = (2, 500)$ の人工データに対して最小サポートと最小確信度をそれぞれ 0.6，0.5 とした場合の最適比率規則抽出結果．左図が最適確信度比率規則，右図が最適サポート比率規則である

Fig. 8 Extracted optimized Ratio Rules for synthetic data when $(p, q) = (2, 500)$. The left figure shows optimized confidence Ratio Rules, and the right figure shows optimized support Ratio Rules, when *minsup* and *minconf* are 0.6 and 0.5, respectively.

べての比率規則は一様に表示されてしまう．その場合，全体として 2 種類の比率規則が存在することは，図示して人間が判断しない限り理解が難しい．したがってこの実験結果から，比率規則統合フェーズが得られた結果の理解を補助していることが分かる．

また同じデータに対し最小確信度と最小サポートをそれぞれ 0.6，0.5 と変化させた場合の結果を**図 8** に示す．この場合枝刈りフェーズでは 1,176 組中 15 組のみ残り，最終的には 3 個の比率規則からなる比率規則集合（図の左側）と 4 個の比率規則からなる比率規則集合（図の右側）が得られた．得られた結果は図 7 の実験と異なりデータ中の線形関係を全体的に表している．特に最適サポート比率規則を見ると全体的なタプルの分布をほぼ近似した結果となっている．この結果から，最小サポートと最小確信度を変化させることでユーザの意図に沿うように得られる比率規則を変化させることができると考えられる．

図 9 は異なる人工データ，すなわち $(p, q) = (5, 2000)$ のときのデータに対する全比率規則集合を表している．左図が最適確信度比率規則集合，右図が最適サポート比率規則集合である．いずれもパラメータは，$\epsilon = 0.01$，$\delta = 0.005$，$minsup = 0.2$，$minconf = 0.4$，$minmerge = 0.5$ のように設定した結果である．枝刈りフェーズでは全部で 22,644 組の

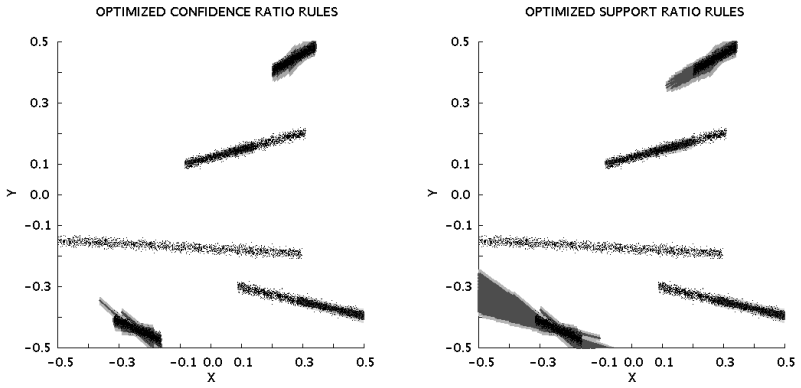

図 9 $(p, q) = (5, 2000)$ の人工データに対する最適比率規則抽出結果
Fig. 9 Extracted optimized Ratio Rules for synthetic data when $(p, q) = (5, 2000)$.

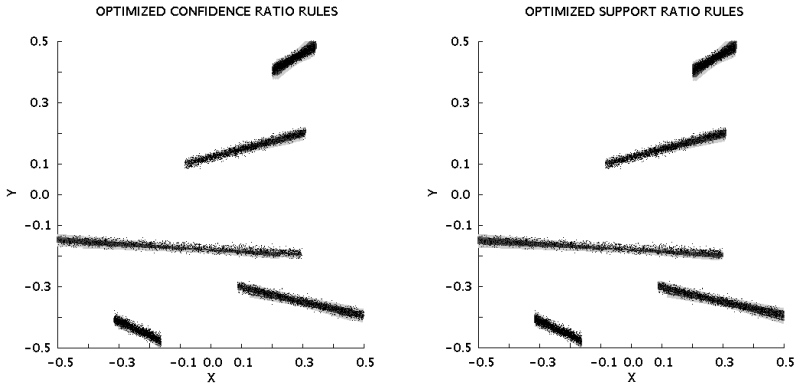

図 10 $(p, q) = (5, 2000)$ の人工データに対する最適局所比率規則の抽出結果
Fig. 10 Extracted optimized Local Ratio Rules for synthetic data when $(p, q) = (5, 2000)$.

(ρ, θ) の候補に対し 442 組が残り，最終的には 4 つの比率規則集合が得られた．得られた全比率規則数は，最適確信度比率規則と最適サポート比率規則のいずれも 189 個である．

この結果では，得られた比率規則集合は適切な線形関係をとらえているが，その一方で $-0.2 < Y < -0.1$ で成り立つ線形関係のようにとらえることができていないものも存在する．この理由として，比率規則を抽出する際の基準であるサポートは属性 X に対する全データを考慮していることがある．そのため，他のクラスタの影響で線形関係をとらえきれない場合が発生する．

これに対し局所比率規則集合を抽出した結果が**図 10** である．左図が最適確信度局所比率規則，右図が最適サポート局所比率規則である．パラメータには，$\epsilon = 0.01$，$\delta = 0.005$，$minlsup = 0.9$，$minlconf = 0.7$，$minmerge = 0.5$，$mincard = 0.5$ を与えた．また，OPTICS を用いてクラスタリングする際のパラメータには “generating distance”，“MinPts”，“clustering distance” の 3 種類があるが，それぞれ 0.1，5，0.05 とした．クラスタリングの結果，各線形関係がそれぞれ別クラスタとなる 5 クラスタが得られた．

得られた局所比率規則は，上に述べた比率規則と比較して明らかなように，データ中の線形関係を適当にとらえていることが分かる．このようにクラスタに分割して比率規則を抽出することで前述の問題を解消している．

6.2.2 アワビデータ

図 11 はアワビデータに対する結果である．人工データの場合と同様，左図が最適確信度比率規則を表し，右図が最適サポート比率規則を表す．図の横軸は属性 “Length”（貝殻の最も長い部分の長さ）を正規化した値を表し，縦軸は属性 “Shell weight”（貝殻のみを測った重さ）の三乗根を正規化した値を表す．貝の体積は長さの三乗に比例するため，重さの三乗根をとることで全体的に線形関係を持ったデータとなっている．パラメータには，$\epsilon = 0.015$，$\delta = 0.01$，$minsup = 0.3$，$minconf = 0.6$，$minmerge = 0.5$ を与えた．枝刈り

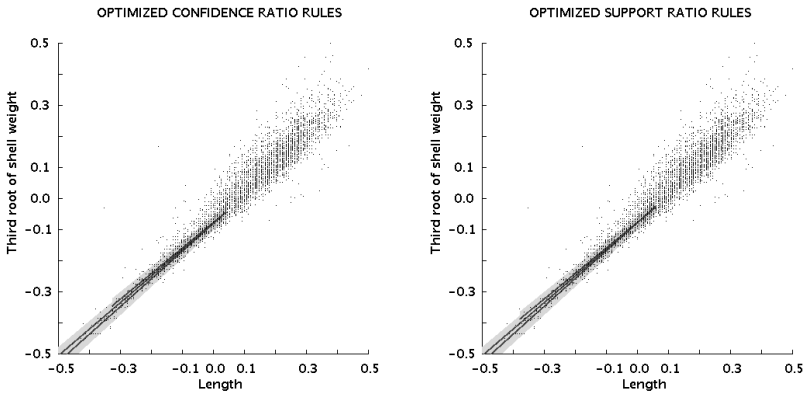

図 11 アワビデータに対する最適比率規則抽出結果
Fig. 11 Extracted optimized Ratio Rules for Abalone data.

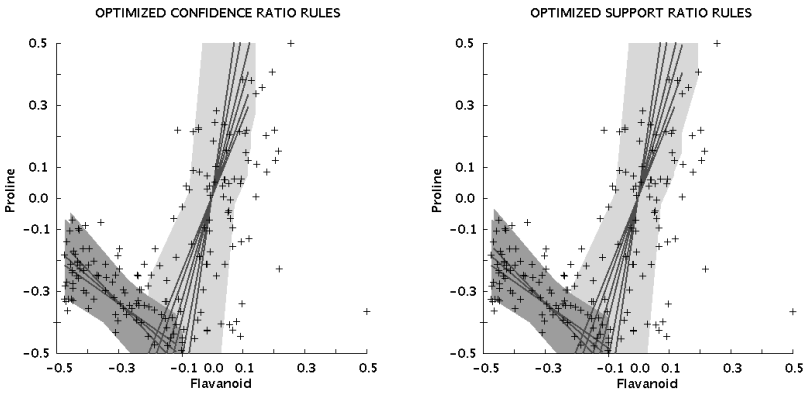

図 12 ワインデータに対する最適比率規則抽出結果
Fig. 12 Extracted optimized Ratio Rules for Wine recognition data.

フェーズの結果 7,875 組中 96 組の候補が残り，最終的には 3 個の比率規則からなる単一の比率規則集合が得られた．

このデータ全体では線形関係が成り立つものの，その分布の仕方は Length の値により異なっている．Length が小さな場合には強い線形関係が成り立ち，大きな場合にはやや弱くなっているが，得られた結果は前者の関係を適当にとらえている．

アワビデータ中には 6.1.2 項で示したように，連続値で表される 7 属性が含まれる．このいずれの 2 属性も線形関係を持つか，三乗根をとると線形関係を持つ．したがって本手法を同様に適用して線形関係を得ることができる．

6.2.3 ワインデータ

図 12 はワインデータに対する結果である．横軸は “Flavanoids”，縦軸は “Proline” のそれぞれ正規化した値を表す．パラメータは最適確信度/サポート比率規則のいずれも，$\epsilon = 0.075$，$\delta = 0.05$，$minsup = 0.5$，$minconf = 0.7$，$minmerge = 0.5$ とした．枝刈りフェーズの結果，全 384 組中 31 組が残り，最終的にはいずれの最適比率規則とも，3 つの比率規則からなる比率規則集合（図の左側）と 6 つの比率規則からなる比率規則集合（図の右側）の計 2 組得られた．このデータでは $-0.5 < Flavanoids < -0.1$ と $-0.1 < Flavanoids < 0.2$ の各部分でタプルが従う線形関係が変化しているが，得られた比率規則集合は各線形関係に対応している．

ワインデータ中に含まれる数値属性は，アワビデータの場合と比べて複雑である．そのため選んだ 2 属性によっては，線形関係を持たない場合や一部分でのみ線形関係を持つ場合がある．後者の場合には本手法を適用することで線形関係が抽出可能である．

6.2.4 自動車データ

図 13 および **図 14** は自動車データより最適比率規則を抽出した結果である．いずれの図も左図は最適確信度比率規則，右図は最適サポート比率規則を表す．

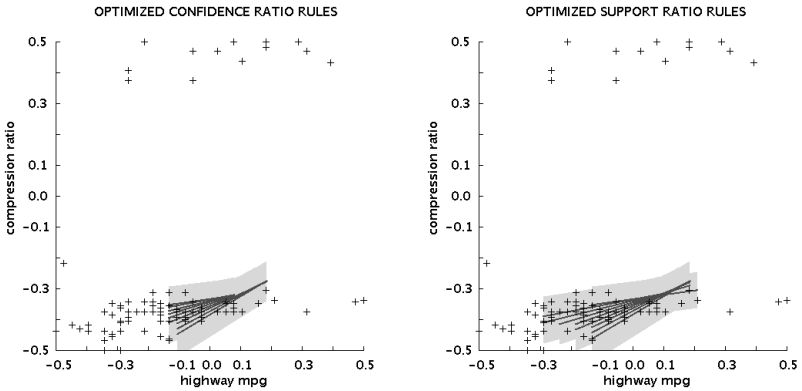

図 13 $minsup = 0.4$ の場合における，自動車データに対する最適比率規則の抽出結果
Fig. 13 Extracted optimized Ratio Rules for Automobile data when $minsup$ =0.4.

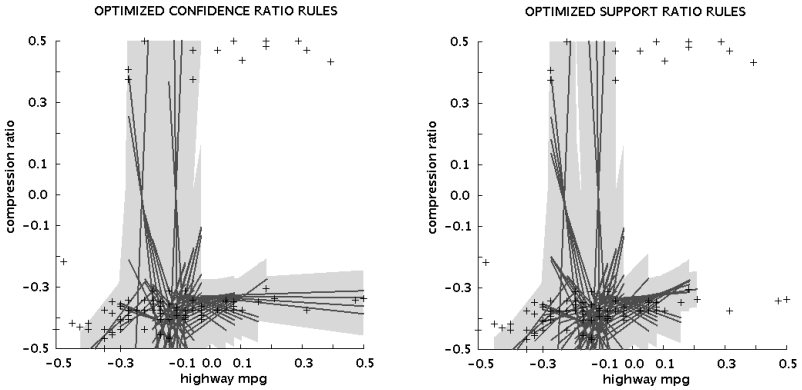

図 14 $minsup = 0.3$ の場合における，自動車データに対する最適比率規則の抽出結果
Fig. 14 Extracted optimized Ratio Rules for Automobile data when $minsup = 0.3$.

図の横軸は高速道路の燃費，縦軸はエンジンの圧縮率のそれぞれ正規化された値を表す．図を見ると圧縮率により異なる線形関係が存在するが，これはディーゼルエンジンとガソリンエンジンで圧縮率が大きく異なるためである．図 13 は，$\epsilon = 0.055$，$\delta = 0.025$，$minsup = 0.4$，$minconf = 0.85$，$minmerge = 0.3$ とした場合であり，図 14 は $minsup$ のみ 0.3 と変化させた場合である．枝刈りフェーズでは，前者は 889 組の候補が 44 組に，後者では 112 組に削減された．最終結果として，$minsup = 0.4$ の場合は 9 個の比率規則からなる単一の比率規則集合が得られた．$minsup = 0.3$ の場合は最適確信度では 58 個，最適サポートでは 55 個からなる単一の比率規則集合が得られた．

このデータでは圧縮率が -0.3 以下で，かつ燃費が -0.3 から 0.1 の領域に多数のタプルが分布している．そのため最小サポートを下げると，その領域を含む任意の比率規則が出力される傾向にあることが分かる．したがって圧縮率が -0.3 以下と 0.3 以上の部分に存在する，異なる 2 つの線形関係を個々に取り出すことは難しい．

一方，最適局所比率規則を抽出した結果が**図 15** である．左図が最適確信度局所比率規則，右図が最適サポート局所比率規則である．パラメータは，$\epsilon = 0.055$，$\delta = 0.025$，$minlsup = 0.7$，$minlconf = 0.75$，$minmerge = 0.4$，$mincard = 0.025$ である．OPTICS のパラメータは，それぞれ generating distance = 0.3，MinPts = 5，clustering distance = 0.3 とした．クラスタリングの結果，圧縮率が -0.2 以下の部分と 0.3 以上の部分の 2 クラスタが得られた．図が示すとおり，得られた 2 つの局所比率規則集合は圧縮率により異なる線形関係を個々にとらえており，妥当な結果といえる．

自動車データの他の数値属性については，車体の全長と高さのように全体的に線形関係があるもの，全長とエンジンのピーク回転数のように明確な線形関係を持たないもの，また今回のようにエンジンの種類によ

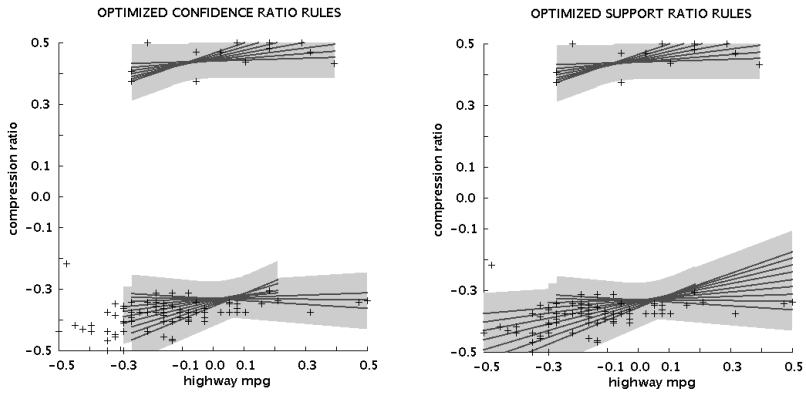

図 15 自動車データに対する最適局所比率規則の抽出結果
Fig. 15 Extracted optimized Local Ratio Rules for Automobile data.

り大きく分布が変わるものがある．全体的な線形関係を持つ場合には大域的な最適比率規則，大きく分布が変わる場合にはこの実験のように最適局所比率規則を抽出することで，それぞれ適当な結果を得ることができる．

以上の実験より，比率規則あるいは局所比率規則を適当に抽出することにより，データ中の線形関係が妥当にとらえることができる．

6.3 処理時間の評価

この実験では，パラメータ ϵ，δ とデータサイズを変化させたときに，比率規則を抽出するための処理時間について調べた．比率規則は最適確信度比率規則と最適サポート比率規則を同時に求めた．処理時間は5回実行させたときの平均値とした．パラメータを変化させる実験には，$(p,q)=(5,2000)$ と $(p,q)=(10,1000)$ の，いずれも 10,000 タプルを持つ 2 種類の人工データを用いた．前者は妥当性の評価で用いたものと同一である．データサイズを変化させた実験には上記 $(p,q)=(5,2000)$ の人工データと同様の線形関係に従うよう，q の値のみを変化させたデータを用いた．

本実験では，特に提案手法における枝刈りフェーズの効果を測るため，提案手法を “枝刈りあり”，図 5 の基本的なアルゴリズムと比率規則統合フェーズのみで枝刈りフェーズを省いたものを “枝刈りなし” として比較を行った．なお枝刈りありの場合の実行時間には枝刈りそのものの処理時間を含んでいる．

6.3.1 パラメータの影響

本実験では，許容幅のパラメータ ϵ，δ を変化させたときの処理時間の変化を調べる．パラメータ ρ，θ はその粒度，すなわちいくつに離散化されるかにより決定した．具体的には，それぞれ 10，50，100，500，1,000 個になるようにしたもので，ϵ は 0.0373，0.00715，0.00356，0.000708，0.000354 の 5 通り，δ は 0.34，0.0635，0.0316，0.00629，0.003146 の 5 通りである．ϵ を変化させる場合には $\delta=0.0316$，δ を変化させる場合には $\epsilon=0.00356$ とした．その他のパラメータは $(p,q)=(5,2000)$ のデータに対しては $minsup=minconf=minmerge=0.3$ とし，$(p,q)=(10,1000)$ のデータに対しては $minsup=minconf=minmerge=0.2$ とした．

まず，$(p,q)=(5,2000)$ の場合の実験結果は**図 16** となる．左図は ϵ を変化させた場合で，右図は δ を変化させた場合である．各図は両対数グラフとして表現してあり，縦軸は処理時間，横軸は粒度を表すため各パラメータの逆数とした．枝刈りフェーズがない場合はパラメータを細かくするにつれファイルからの読み込みが増加するので，結果としてほぼパラメータの粒度に対し線形で処理時間が増加している．これに対し枝刈りフェーズを加えた場合，許容幅が小さくなるにつれ各比率規則に従うタプル数は減少するので，枝刈りの効果が増加し処理時間は大きく変わっていないと考えられる．

また，$(p,q)=(10,1000)$ の場合の結果は**図 17** である．この場合も $(p,q)=(5,2000)$ の場合と同様の傾向が得られた．

一方許容幅の変化と，枝刈りフェーズで残る比率規則の候補数の割合および最終的に生成される比率規則数の割合の関係を**図 18** および **図 19** に示す．なお比率規則数は最適確信度，最適サポートとも同数である．各図とも，縦軸は考え得るパラメータ組 (ρ,θ) の全候補に対して，枝刈りフェーズの結果残った比率規則および最終結果として得られた比率規則の割合である．この結果から許容幅が狭くなりパラメータ組の候補が増えるほど，枝刈りおよび最終結果の割合がほぼ反比

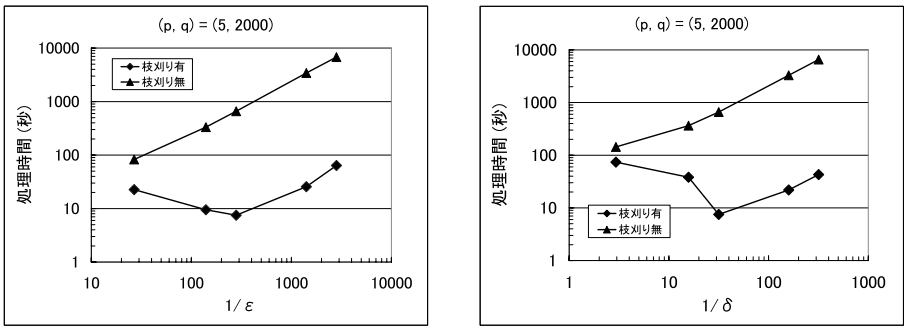

図 16 $(p, q) = (5, 2000)$ の人工データにおける比率規則抽出の実行時間
Fig. 16 Processing time of Ratio Rules extraction for synthtic data when $(p, q) = (5, 2000)$.

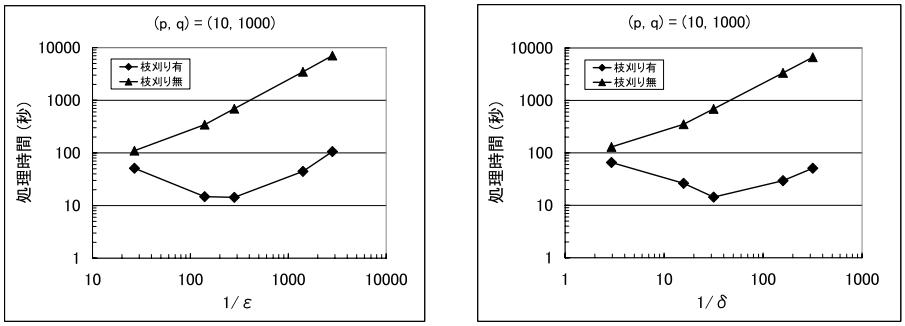

図 17 $(p, q) = (10, 1000)$ の人工データにおける実行時間
Fig. 17 Processing time of Ratio Rules extraction for synthtic data when $(p, q) = (10, 1000)$.

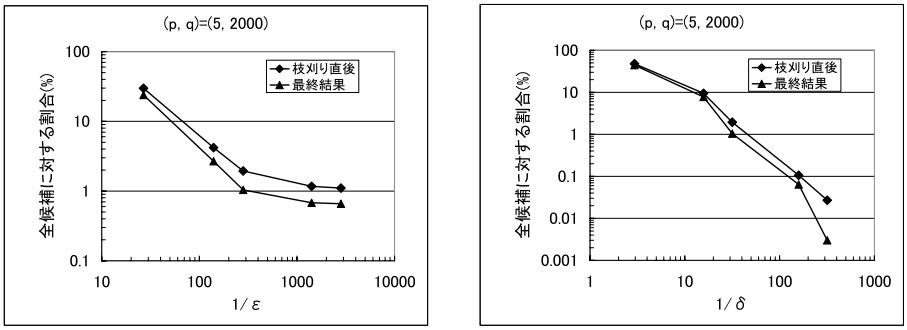

図 18 $(p, q) = (5, 2000)$ の人工データにおける比率規則の候補数および最終結果数の割合
Fig. 18 Ratio of the number of candidate or result Ratio Rules for synthtic data when $(p, q) = (5, 2000)$.

例して減少していることが分かる．4.3 節に記述したように，比率規則統合フェーズのコストは $O(Q^2N)$ と，得られる比率規則数 Q に対し二乗のコストがかかる．そのため候補比率規則数が多くなるほど膨大なコストがかかる可能性があるが，この結果からは適当な最小確信度と最小サポートに対しては Q の与える影響は限定的と考えられる．

6.3.2 スケーラビリティ

次に与えられるタプル数を変化させたときの処理時間を調べる．タプル数は 10,000，50,000，100,000，250,000，500,000 の 5 種類とし，いずれも同一の 5 つの線形関係を持つよう生成した．パラメータには，$\epsilon = 0.00356$，$\delta = 0.0316$，$minsup = minconf = minmerge = 0.3$ を与えた．

実験結果は**図 20** であり，横軸がタプル数，縦軸が

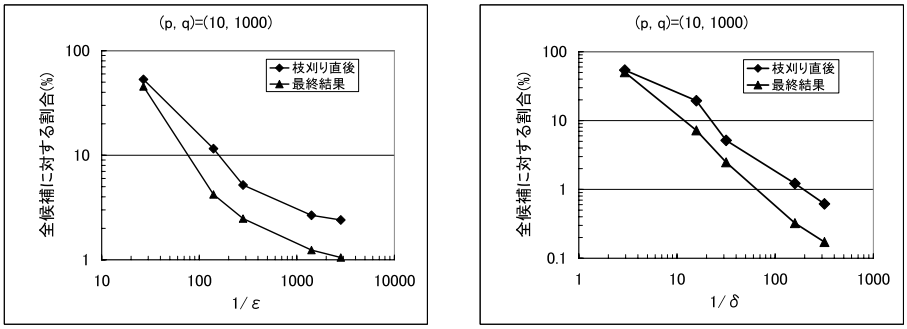

図 19 $(p, q) = (10, 1000)$ の人工データにおける比率規則の候補数および最終結果数の割合
Fig. 19 Ratio of the number of candidate or result Ratio Rules for synthtic data when $(p, q) = (10, 1000)$.

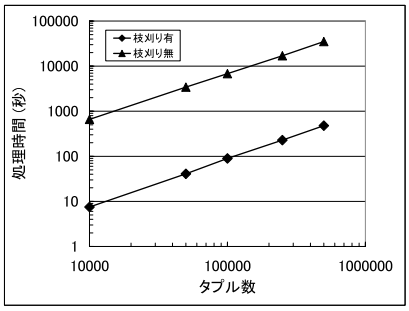

図 20 スケーラビリティの実験結果
Fig. 20 Result in the scalability experiment.

処理時間である両対数グラフで表している．4.4 節で述べたとおり，入力データサイズに対して線形の処理時間であることが結果に示されている．また枝刈りを行うことでタプル数によらず処理時間がほぼ 1/100 となっている．

7. 議　　論

素朴な疑問として，既存の線形関係抽出手法と，最適局所比率規則で用いたようにクラスタリングを組み合わせれば，妥当な結果が得られるのではないか，ということがある．本章ではクラスタリングと回帰分析を組み合わせた手法，クラスタリングと主成分分析を組み合わせた手法，および本論文で提案した最適局所比率規則を比較する．クラスタリング手法には前章の局所比率規則抽出と同様 OPTICS を用いた．

本章で用いる実験データは，前章で説明した人工データの生成方法でパラメータを $(p, q) = (4, 500)$ として得られる，**図 21** のようなデータである．図 21 では同時に，パラメータを $\epsilon = 0.025$，$\delta = 0.02$，$minsup = 0.5$，$minconf = 0.5$，$minmerge = 0.3$ として得られた最適局所比率規則を表示している．ここではタプルの分布として，図で示したように 4 種類の線形関係があるが，そのうち 2 種類が交わる場合を考える．

OPTICS のパラメータは，それぞれ generating distance = 0.05，MinPts = 5，clustering distance = 0.05 とした．この結果，図左側の線形関係が存在する部分，図上側の線形関係が存在する部分，図右下の 2 種類の線形関係が混在する部分の 3 クラスタが得られた．この場合，最適確信度局所比率規則（左図）は単独の線形関係が存在するクラスタ全体と，2 種類の線形関係が混在するクラスタの交点付近を中心とした部分に対応する．また，最適サポート局所比率規則の場合には，2 種類の線形関係が混在するクラスタにおいて，より広く 2 種類の関係をとらえていることが分かる．

これに対し，クラスタリングと線形回帰分析を組み合わせた結果が **図 22** である．単一の線形関係のみが含まれるクラスタについては，非常によくその関係をとらえているが，2 種類の線形関係が現れるクラスタへ適用した場合，いずれの線形関係もとらえることができない．

またクラスタリングと，主成分分析すなわち Korn らの比率規則抽出手法[11] を適用した結果が **図 23** である．この場合も単一の線形関係のみが含まれるクラスタについては適当な線形関係がとらえることができているが，線形関係が交わっているクラスタについては各線形関係をとらえることができていない．これは得られる線形関係がタプルの平均値を通り，かつクラスタ中の分布を最大にするような直線が得られるためである．

このことから，本提案手法は既存の線形関係抽出手法よりもより一般性を持っていると考えられる．

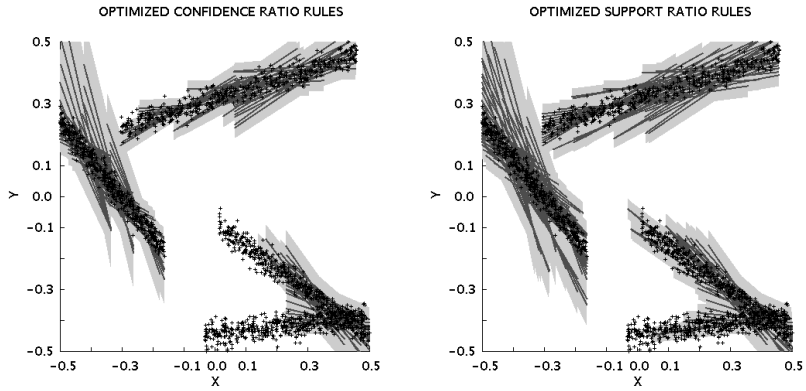

図 21 $(p, q) = (4, 500)$ の人工データに対して最適局所比率規則を抽出した結果
Fig. 21 Extracted optimized Local Ratio Rules for synthetic data when $(p, q) = (4, 500)$.

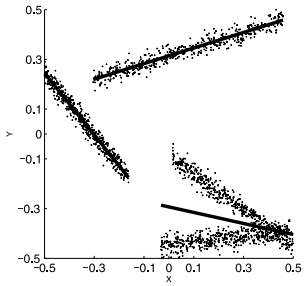

図 22 $(p, q) = (4, 500)$ の人工データに対してクラスタリングと線形回帰分析を行った結果
Fig. 22 Extracted result of a method using clustering and linear regression for synthetic data when $(p, q) = (4, 500)$.

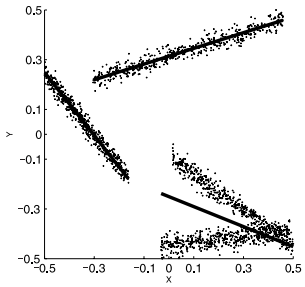

図 23 $(p, q) = (4, 500)$ の人工データに対してクラスタリングと主成分分析を行った結果
Fig. 23 Extracted result of a method using clustering and Principal Component Analysis for synthetic data when $(p, q) = (4, 500)$.

8. おわりに

本論文では，比率規則を抽出する新たな手法を提案した．特に相関ルールマイニングと対応付けし，比率規則にサポートや確信度といった概念を持ち込み，局所的に強く成り立つような比率規則を抽出する手法を提案した．提案手法は，枝刈り，比率規則生成，比率規則統合の 3 フェーズから構成され，入力タプル数に対して線形時間で比率規則を求めることができる．また，この拡張手法として局所比率規則の抽出手法を提案した．局所比率規則によりタプルの分布が偏った場合でも適当な線形関係をとらえることができた．提案手法は密度ベースのクラスタリング手法を用いてデータの局所性をとらえた後，各クラスタより局所比率規則を抽出する．これら提案手法を人工データと 3 種類の実データに適用しその妥当性を確認するとともに，人工データによる実験により大規模なデータに対する処理時間の分析を行った．

今後の課題としては，本論文中ではすべてのクラスタで共通していた最小サポートと最小確信度を，クラスタ中のタプルの密度によって適当に与えるなど，パラメータを自動的にチューニングする手法の開発がある．また本論文における比率規則の定義では，比率規則を表す領域中のタプルの分布は考慮していないが，基準となる線分からの離れ具合など領域中の分布を，比率規則の抽出に組み込むことが考えられる．ほかには 3 属性以上の間における比率規則の抽出手法の検討，本手法に適したクラスタリング手法の開発などが考えられる．

謝辞 本研究の一部は，科学研究費補助金特定領域研究（#18049005）による．

参 考 文 献

1) Agrawal, R., Imielinski, T. and Swami, A.: Mining Association Rules Between Sets of Items in Large Databases, *Proc. ACM SIGMOD International Conference on Management of Data*, Washington, D.C., pp.207–216

(1993).
2) Ankerst, M., Breunig, M.M., Kriegel, H.-P. and Sander, J.: OPTICS: Ordering Points To Identify the Clustering Structure, *Proc. ACM SIGMOD International Conference on Management of Data*, Philadelphia, PA, pp.49–60 (1999).
3) Duda, R. and Hart, P.: Use of the Hough Transformation to Detect Lines and Curves in Pictures, *Comm. ACM*, Vol.15, No.1, pp.11–15 (1972).
4) Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T.: Data Mining Using Two-Dimentional Optimized Association Rules: Scheme, Algorithms, and Visualization, *Proc. ACM SIGMOD International Conference on Management of Data*, Montreal Quebec, Canada, pp.13–23 (1996).
5) Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T.: Mining Optimized Association Rules for Numeric Attributes, *Journal of Computer and System Sciences*, Vol.58, No.1, pp.1–12 (1999).
6) Han, J. and Kamber, M.: *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco (2001).
7) Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning*, Springer-Verlag, New York (2001).
8) Hough, P.: Methods and Means for Recognizing Complex Patterns (1962). U.S. Patent 3,069,654.
9) Hu, C., Wang, Y., Zhang, B., Yang, Q., Wang, Q., Zhou, J., He, R. and Yan, Y.: Mining Quantitative Associations in Large Database, *Proc. 7th Asia-Pacific Web Conference*, Shanghai, China, pp.405–416 (2005).
10) Hu, C., Zhang, B., Yan, S., Yang, Q., Yan, J., Chen, Z. and Ma, W.-Y.: 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).
11) Korn, F., Labrinidis, A., Kotidis, Y. and Faloutsos, C.: 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).
12) Korn, F., Labrinidis, A., Kotidis, Y. and Faloutsos, C.: Quantifiable Data Mining Using Ratio Rules, *VLDB Journal*, Vol.8, pp.254–266 (2000).
13) Srikant, R. and Agrawal, R.: Mining Quantitative Association Rules in Large Relational Tables, *Proc. ACM SIGMOD International Conference on Management of Data*, Montreal Quebec, Canada, pp.1–12 (1996).
14) 福田剛志，森本康彦，徳山 豪：データマイニング，共立出版 (2001).

付 録

ある直線上にある点群から，その直線の式 $y = a_0x + b_0$ を検出する問題を考える．1つの手法として各点 (x_i, y_i) を通る直線 $y_i = ax_i + b$ のパラメータ (a, b) を列挙する手法が考えられる．この操作をすべての点について行い，得られた (a, b) のヒストグラムにおいて $a = a_0$，$b = b_0$ が最も頻度が大きくなる．

しかしパラメータ a，b はいずれも区間 $(-\infty, \infty)$ の値をとるため a，b の組は無限に存在し，列挙は非常に難しい．この問題に対し Hough 変換[3),8)] は，無限の区間をとる2パラメータ a，b を，有限の区間をとる2パラメータ ρ，θ へ変換する．

パラメータ ρ，θ の意味は**図 24** のとおりである．直線 $y = ax + b$ は $\rho = x\cos\theta + y\sin\theta$ として表される．ここで ρ は直線から原点へ引かれた垂線の長さ，θ は X 軸と垂線のなす角度を表す．パラメータ (a, b) と (ρ, θ) の関係は $\rho = b\sin(\tan^{-1}(-1/a))$，$\theta = \tan^{-1}(-1/a)$ となる．

本論文では属性 X，Y はドメインが区間 $[-0.5, 0.5]$ となるよう正規化されているので，ρ，θ のドメインはそれぞれ $[0, \sqrt{2}/2]$，$[0, 2\pi]$ におさえられる．それゆえすべての (a, b) を列挙することは，$\rho - \theta$ 空間の領域 $0 \le \rho \le \sqrt{2}/2$，$0 \le \theta \le 2\pi$ に含まれる点 (ρ, θ) を列挙することと考えられる．

a_0，b_0 に対応するパラメータ ρ_0，θ_0 を得るための古典的な手法[3)] は以下のとおりである．

(1) ρ，θ で張られる2次元空間をユーザが与える

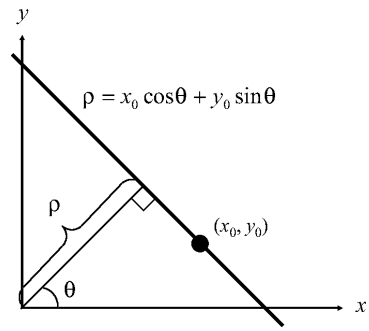

図 24 Hough 変換における各パラメータの関係
Fig. 24 Relationships among parameters in Hough transformation.

分割幅で分割する（本手法では，ρ 軸を 2ϵ 間隔，θ 軸を 2δ 間隔で等分割して $2\epsilon \times 2\delta$ のセルをカウントに用いる）．

(2) 各点 (x_i, y_i) に対して，曲線 $\rho = x_i \cos\theta + y_i \sin\theta$ が通過するセルのカウンタをインクリメントする．

(3) 最もカウント数が大きなセルに対応するパラメータ (ρ, θ) を出力する．

(平成 18 年 6 月 20 日受付)
(平成 18 年 10 月 6 日採録)

（担当編集委員　大森 匡）

濱本　雅史（学生会員）

2003 年筑波大学第三学群情報学類卒業．同年同大学大学院システム情報工学研究科入学，現在に至る．データマイニング，多変量解析に興味を持つ．日本データベース学会，ACM 各学生会員．

北川　博之（フェロー）

1978 年東京大学理学部物理学科卒業．1980 年同大学大学院理学系研究科修士課程修了．日本電気（株）勤務の後，1988 年筑波大学電子・情報工学系講師．同助教授を経て，現在，筑波大学大学院システム情報工学研究科教授，ならびに計算科学研究センター教授．理学博士（東京大学）．異種情報源統合，XML とデータベース，WWW の高度利用，データマイニング等の研究に従事．著書『データベースシステム』（昭晃堂），“The Unnormalized Relational Data Model”（共著，Springer-Verlag）等．日本データベース学会副会長，ACM，IEEE-CS，日本ソフトウェア科学会各会員．