比率規則による線形関係マイニングシステムの開発

濱　本　雅　史 †

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．線形関係を抽出する既存の手法として、比率規則を抽出する手法がある．既存の比率規則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．これに対し筆者らは，比率規則を線分とその周辺領域内のデータが満たす性質として定義し，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してきた．本論文では，この比率規則を抽出するためのシステム RRMiner の開発手法について述べ，実際に人工データおよび実データより抽出が行えることを示す．

Development of Linear Relationships Mining System by Ratio Rules

Masafumi Hamamoto †

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. A method to extract linear relationships is the Ratio Rule mining. In the existing Ratio Rules, 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. We have formulated a Ratio Rule as a line segment and its neighborhood, and then solved problems in existing methods by introducing support and confidence concepts. In this paper we describe the development of a system named RRMiner for extracting our Ratio Rules. We also show the system is applicable both synthetic and real data.

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[5]．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**[10] 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[10]，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，可視化など様々な応用が可能である．

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
...	...	...

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[9),10]．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

† 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

☆ 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．

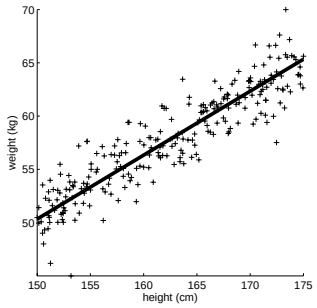

図 1　表 1 のデータに対する比率規則の例．実線が比率規則を表す
Fig. 1　Ratio Rule for Table 1. Black solid line represents a Ratio Rule.

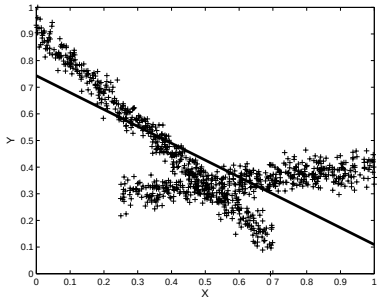

図 2　複数の比率規則が成り立つ例．実線は主成分分析を用いた手法で得られた比率規則を表す
Fig. 2　An example where multiple Ratio Rules exist. Black solid line represents a Ratio Rule extracted by Principal Component Analysis.

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[10] 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[9] 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．

このような問題を解決するためのアイディアとして，筆者らは論文[13] において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した．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．筆者ら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した．

本論文ではこの比率規則を抽出し，活用するためのシステム RRMiner の開発について述べる．まず 2 章で線形関係の抽出に関する関連研究について触れ，3 章において本システムが扱う比率規則の定式化について述べる．続いて 4 章で比率規則を抽出するアルゴリズムについて述べる．このアルゴリズムは，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[4] を用いた最適区間抽出，抽出された比率規則の統合の 3 フェーズからなり，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．そして 5 章において RRMiner の構成について述べ，6 章で実際にこのシステムで人工データおよび実データから妥当な比率規則を抽出し，ユーザに提示できることを示す．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[3),4),12),14)]．Fukuda らの手法ではルールの前提部が 1 属性の場合[4] および 2 属性の場合[3] 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[12] 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[10] 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．

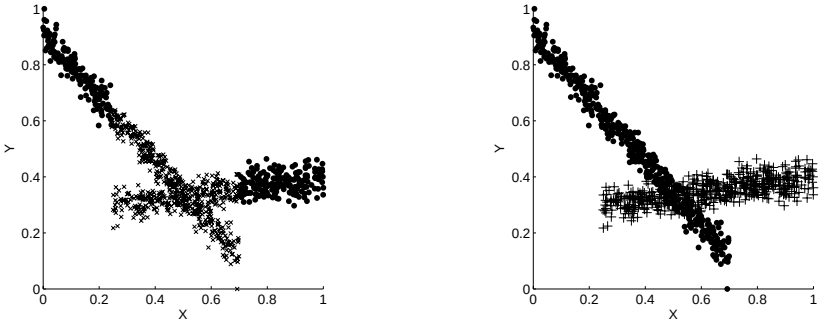

図 3 ユーザの興味により得るべき線形関係が異なる例．同じデータであっても左図では線形関係の強さに興味が置かれ，右図では線形関係全体に興味が置かれている

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.

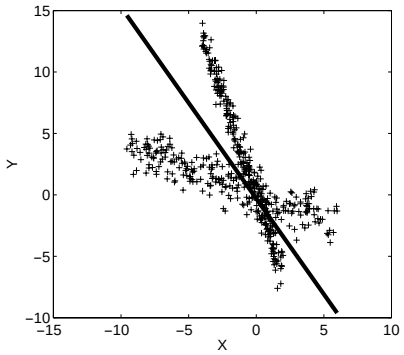

図 4 主成分分析を用いた手法ではうまくいかない例．直線は得られた比率規則を表す

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

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[10),11)]．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，第一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，第一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[8),9)]．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[6)]．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して，筆者らが論文[13)]にて提案した比率規則の特徴として，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．

3. 比 率 規 則

本章では抽出すべき比率規則の定義を説明する．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，ユーザの意図を反映させることができる．

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 で挙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**[7] 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．

以上の点より，比率規則は次のように定義される．

> **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[4] 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. 比率規則抽出手法

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，$\theta_j (i = 1, \cdots, R,\ j = 1, \cdots, T)$ として考える．すなわち比率規則 $RR_I(\rho_i, \theta_i)$ が，パラメータ ρ，θ が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[4] 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 次元数値属性相関規則**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 次元数値属性相関規則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[4] 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純に増加する．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．

```
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.

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．

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

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 アルゴリズムのまとめ

本手法は，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. 比率規則マイニングシステム：RRMiner

本章では，これまでの章にて説明した比率規則を抽出し活用するためのシステム RRMiner の構築について述べる．

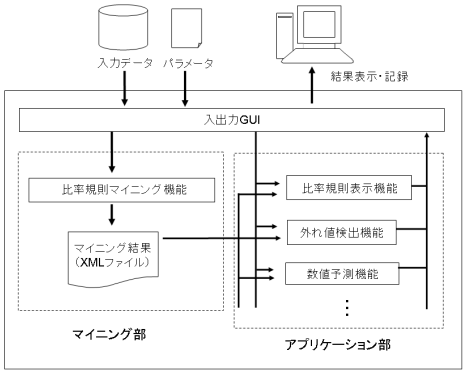

図 6 RRMiner のシステム構成図
Fig. 6 System diagram of RRMiner.

5.1 システム構成

本システムの目的は比率規則を単に抽出するだけでなく，その結果を用いた応用も行うことである．そのためシステムの構成としては大まかに，比率規則の抽出を行うマイニング部と，得られた比率規則の応用を行うアプリケーション部の 2 つに分けられる．これに加え，ユーザとの入出力のやりとりを行う GUI を備えたものが RRMiner の構成である (**図 6**).

マイニング部では，ユーザより与えられたデータとパラメータから 4 章で説明したアルゴリズムにより比率規則を抽出し，その結果を XML ファイルとして書き出す．アプリケーション部では抽出された比率規則を用いて，新たにユーザより与えられるデータ☆へ様々な応用を行う．比率規則の応用には比率規則を可視化して表示するほか，Korn らの論文[11] で示されているように，外れ値検出や数値予測などの様々なものが考えられる．この応用については次節にて説明する．

実際に実装された RRMiner のスクリーンショットは**図 7** である．GUI 部分およびアプリケーション部は Java(JDK1.5)，マイニング部は C 言語で実装されている．OS は現在のところ，WindowsXP と Linux(Fedora Core 5) にて動作を確認している．

5.2 比率規則の応用

前節で述べたように，マイニング部で得られた比率規則はアプリケーション部で様々な用途に応用される．以下すでに実装済みである比率規則表示機能と，今後の課題であるそのほかの機能について説明する．

比率規則表示機能

本システムでは，与えられたデータに対して比率規則がどのように得られたかを表示する機能を持つ．具体的な画面を**図 8** に示す．

中央の図において，黒の点は各タプルを表し，濃い

☆ 応用の種類により，比率規則を抽出するときに用いたデータと同一の場合もありうる

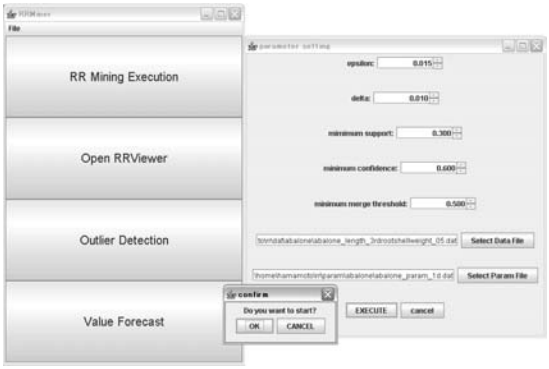

図 7 RRMiner の実行画面
Fig. 7 Screenshot of RRMiner.

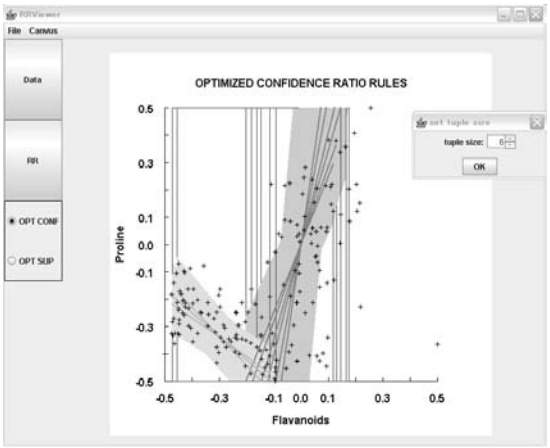

図 8 RRMiner の比率規則表示機能画面
Fig. 8 Screenshot of Ratio Rule viewer in RRMiner.

色の線分は得られた各比率規則の許容幅を含まない形（$RR_{x \in I}(\rho \pm 0, \theta \pm 0)$），薄い色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．色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色で表されている．各比率規則を囲うように描かれている長方形の幅は，比率規則が成り立つ区間 $I \subset X$ である．すなわち，全タプル数に対する長方形内に含まれるタプル数が区間 I のサポートであり，長方形内に含まれるタプル中で比率規則に従うタプル数が比率規則の確信度となる．

そのほかの機能

比率規則表示機能以外の応用として，外れ値検出機能や数値予測機能が考えられる．

外れ値検出は，得られた比率規則に従わないタプルを外れ値として検出する機能である．外れ値として検出されたタプルをデータから取り除くことで，タプル全体を比率規則の組み合わせで近似することが可能であり，データのクリーニングや情報圧縮が可能となる．また逆に外れ値として得られたタプルを調査することで，データの不正な入力や特異なデータの発見が可能となる．

外れ値検出を行う上での課題として，比率規則に従わないタプルにも分布の差があることがある．つまり，いずれの比率規則からも大きく離れている明らかな外れ値と，許容幅がわずか足りずに外れ値となってしまったものを，同様に扱っていいかという問題である．これについては最も近い比率規則からの距離を用い，ランキング表示をしたり，ある閾値以上のタプルのみを外れ値とするなどの工夫が考えられる．

数値予測機能は，ある属性値が欠けている場合などに比率規則から推定する手法である．属性値 x がもし比率規則 $RR_I(\rho, \theta)$ の区間 I に含まれる場合，属性値 y は $(\rho - x\cos\theta)/\sin\theta$ の周辺であると推定を行う．この際比率規則の確信度によって，推定された値のもっともらしさを測ることができる．

ただし与える属性値によっては，対応する比率規則が複数になる場合や，逆に全く存在しない場合もありうる．前者の場合，最も確信度が高い比率規則を選んだり，複数の比率規則でそれぞれ推定した値の平均値をとるなどの手法が考えられる．後者の場合，最も区間 I が近くなる比率規則を当てはめることなどが考えられる．

6. 実　　験

本章における実験では，本システムより得られる比率規則の妥当性を，前章で説明した RRMiner の比率規則表示機能によって示す．ここでは人工データと 2 種類の実データを用いる．実データはいずれも UCI の Machine Learning Repository☆から入手可能である．

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．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．

(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．

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

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

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．

6.2 実 験 結 果

6.2.1 人工データ

まず人工データについての実験結果を示す．人工データを生成する際のパラメータ (p, q) には $(2, 500)$ を与えた．これは 1 章の図 2 で示された例のデータである．

図 9 は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無い場合，得られるす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，$minmerge$ を 0.35 と変化させた場合の結果を**図 10** 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 9 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．

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．

以上の実験より，本システムを用いて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を妥当にとらえることができることが分かる．

7. お わ り に

本論文では，比率規則によって線形関係を抽出するシステム RRMiner について述べた．このシステムによって得られる比率規則は，相関ルールマイニングと対応付けされており，サポートや確信度といった概念を持つため，局所的に強く成り立つような線形関係を

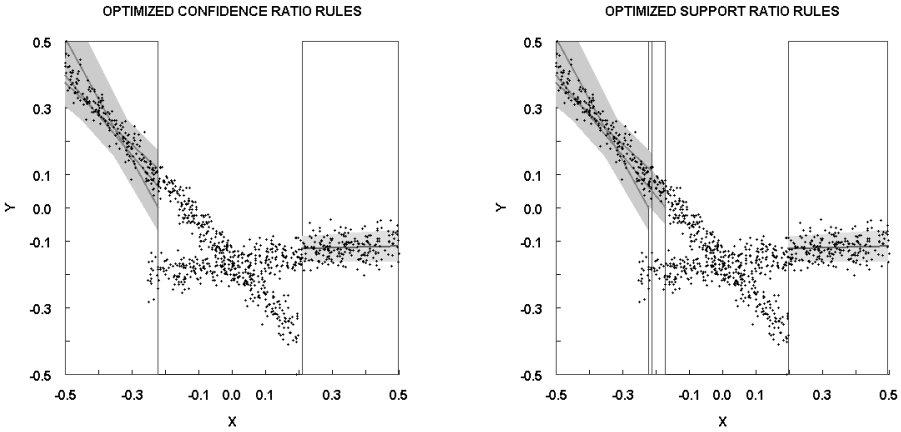

図 9 $(p, q) = (2, 500)$ の人工データに対する最適比率規則抽出結果．左図が最適確信度比率規則，右図が最適サポート比率規則であり，$minsup$，$minconf$ はそれぞれ 0.2 と 0.8 である

Fig. 9 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.

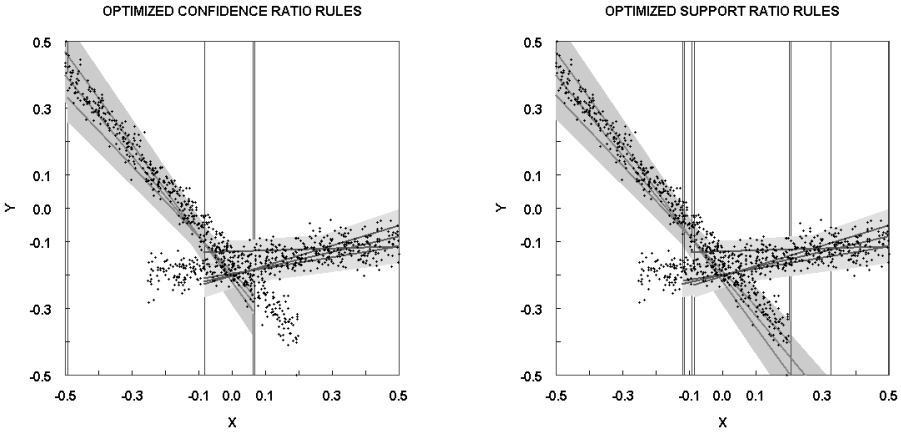

図 10 $(p, q) = (2, 500)$ の人工データに対して最小サポートと最小確信度をそれぞれ 0.6，0.5 とした場合の最適比率規則抽出結果．左図が最適確信度比率規則，右図が最適サポート比率規則である

Fig. 10 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.

とらえることができる．本システムの中心部分である比率規則の抽出アルゴリズムは，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得ることができる．また得られた比率規則に対する応用として，本システムは比率規則を可視化する機能を持つ．本論文ではこの機能について人工データと 2 種類の実データに適用した結果を示し，その妥当性を確認した．

今後の課題として，本論文で述べた比率規則の可視化以外の応用手法，パラメータの自動設定手法，3 属性以上の間における比率規則の抽出手法の検討および実装があげられる．また本論文で挙げた以外の実データに対する実験と考察も課題としてあげられる．

謝辞 本システムを開発するにあたりご指導頂いた，

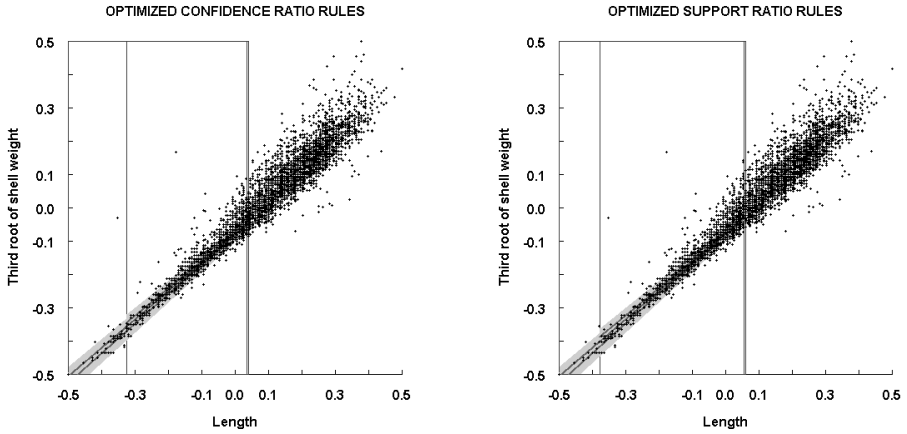

図 11 アワビデータに対する最適比率規則抽出結果
Fig. 11 Extracted optimized Ratio Rules for Abalone data.

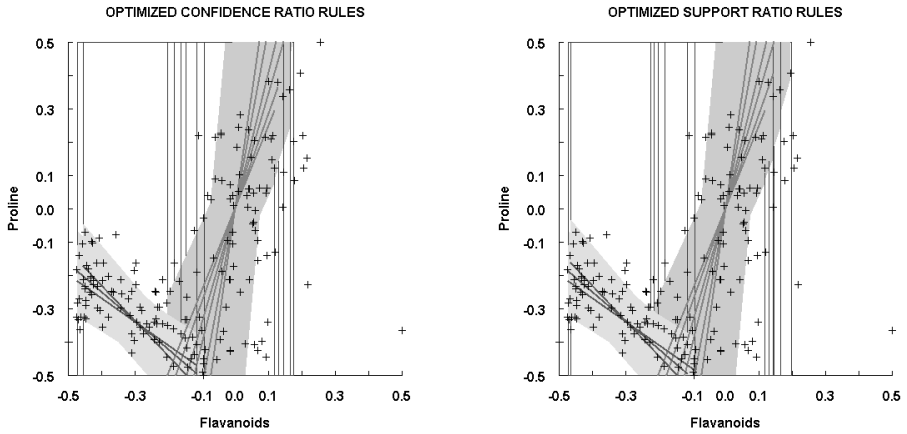

図 12 ワインデータに対する最適比率規則抽出結果
Fig. 12 Extracted optimized Ratio Rules for Wine recognition data.

筑波大学大学院システム情報工学研究科の北川博之教授に感謝いたします.

本研究の一部は，魅力ある大学院教育イニシアティブ「実践 IT 力を備えた高度情報学人材育成プログラム」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) Duda, R. and Hart, P.: Use of the Hough Transformation to Detect Lines and Curves in Pictures, *Communications of the ACM*, Vol.15, No.1, pp.11–15 (1972).

3) 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).

4) 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).

5) Han, J. and Kamber, M.: *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco (2001).

6) Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning*, Springer-Verlag, New York (2001).

7) Hough, P.: Methods and Means for Recognizing Complex Patterns (1962). U.S. Patent 3,069,654.
8) 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).
9) 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).
10) 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).
11) Korn, F., Labrinidis, A., Kotidis, Y. and Faloutsos, C.: Quantifiable Data Mining Using Ratio Rules, *VLDB Journal*, Vol.8, pp.254–266 (2000).
12) 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).
13) 濱本雅史，北川博之：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，情報処理学会論文誌:データベース，Vol.47, No.SIG19(TOD32), pp.54–71 (2006).
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[2),7)] 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 13** 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．

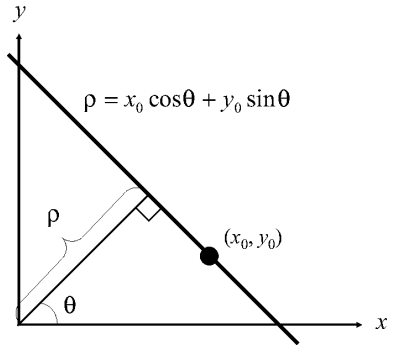

図 13　Hough 変換における各パラメータの関係
Fig. 13　Relationships among parameters in Hough transformation.

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 $[0, \sqrt{2}/2]$，$[0, 2\pi]$ に押さえられる．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[2)] は以下の通りである．

(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．