サポートと確信度を考慮した比率規則マイニング

濱本　雅史† 　北川　博之†,††

† 筑波大学 システム情報工学研究科　〒 305–8573 茨城県つくば市天王台 1–1–1
†† 筑波大学 計算科学研究センター　〒 305–8573 茨城県つくば市天王台 1–1–1
E-mail: †hamamoto@kde.cs.tsukuba.ac.jp, ††kitagawa@cs.tsukuba.ac.jp

あらまし　近年様々なデータマイニング手法が検討されているが、本研究では特に比率規則を抽出する問題を考える。比率規則は属性間で成り立つ線形関係であり、データの理解補助、欠損値の補完など様々な応用が可能である。既存の比率規則に関する研究では、サポートや確信度のような相関規則マイニングで用いられる概念が存在しない。そのため部分的に強く成り立つ線形関係を比率規則として捉えることができないことや、得られた比率規則に対する客観的な尺度が得られないことなどの問題点がある。そこで本研究では比率規則に対して相関規則マイニングと対応した定式化を行いサポートと確信度の概念を導入する。それを元にサポートおよび確信度を最大とする比率規則を得る手法を提案し、実データと人工データを用いた実験により提案手法の妥当性を示す。
キーワード　比率規則, 相関規則マイニング, データマイニング, ハフ変換

Mining Ratio Rules Based on Support and Confidence

Masafumi HAMAMOTO† and Hiroyuki KITAGAWA†,††

† Graduate School of Systems and Information Engineering, University of Tsukuba, Tennohdai 1–1–1, Tsukuba, Ibaraki, 305–8573 Japan
†† Center for Computational Sciences, University of Tsukuba, Tennohdai 1–1–1, Tsukuba, Ibaraki, 305–8573 Japan
E-mail: †hamamoto@kde.cs.tsukuba.ac.jp, ††kitagawa@cs.tsukuba.ac.jp

Abstract　Extraction of rules from a large dataset is important in data mining. This paper examines the problem of extracting Ratio Rules. Ratio Rules are linear relationships in numeric attributes applicable to understanding data, filling missing attribute values, and related issues. Existing research for Ratio Rules, however, does not consider concepts used in association rule mining. This prevents us from extracting a Ratio Rule having a strong linear relationship in part. This also prevents us from measuring objective goodness of each Ratio Rule. We formulated Ratio Rule mining in analogy to association rule mining, and introduce support and confidence concepts to Ratio Rules. We propose a Ratio Rule extraction method based on support and confidence, and show the appropriateness of our proposed method using real and synthetic data.
Key words　Ratio Rule, association rule mining, data mining, Hough transformation

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。例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 マイニングといったことが挙げられる。

本研究では特に**比率規則** [8] を抽出する問題を考える。比率規則は属性間における、属性値の典型的な割合を表したものである。言い換えると、属性間で成り立っている線形関係が比率規則となる。

具体例として表 1 のような、“身長” と “体重” の 2 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のが図 1 である。この図を見ると、黒い直線で表されたような線形関係を全体的に持っていることがわかる。また直線の傾きから、“身長” と “体重” における増分の比率を得ることが出来る。このような直線が比率規則を表す。比率規則は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。

既存の手法 [7] [8] は、全体の傾向を比率規則として捉えよう

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。

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

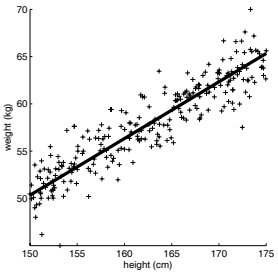

図 1　表 1 のデータに対する比率規則の例。実線が比率規則を表す。

とするため、複数の線形関係が混在したり、特定の区間でのみ線形関係が成り立つデータには適していない。例えば図 2 のように、$0.0 \leqq X \leqq 0.7$ において成り立っている線形関係と、$0.25 \leqq X \leqq 1.0$ において成り立っている線形関係が異なるとする。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 [8] 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表していないため妥当とは言い難い。

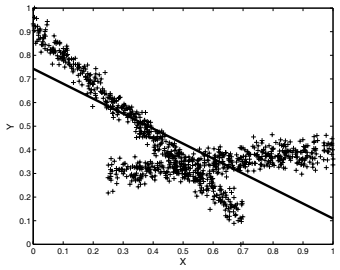

図 2　複数の比率規則が成り立つ例。実線は Korn らの手法で得られた比率規則を表す

もうひとつの既存の比率規則の問題点として、サポートや確信度のように相関規則マイニングで用いられる概念が存在しないことがある。図 2 の例では、$0 \leqq X \leqq 0.2$ および $0.7 \leqq X \leqq 1.0$ の区間には単一の線形関係しか存在せず、かつほとんどのタプルがその線形関係に従っている。言い換えれば確信度が非常に高いことになる。しかし $0.2 \leqq X \leqq 0.7$ の区間には 2 つの線形関係が混在しているので、属性 X の値が分かっても属性 Y の値の推定はより難しくなり、確信度が低いといえる。このような状況の場合、ユーザがサポートや確信度の基準を与え、それに応じて適当な比率規則を抽出することが有用である。

本論文ではまず比率規則の定義の定式化を行い相関規則マイニングと対応付けする。この定義を元に、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込み、1 次元数値属性相関規則マイニング [2] を用いた最適区間抽出、抽出された比率規則の統合の 3 フェーズから成る手法を提案する。この手法は入力タプル数に対して線形で比率規則を求めることが可能である。

本稿は以下のように構成される。2 章では比率規則の定式化を行い、相関規則マイニングの概念を導入する。3 章において提案手法を示す。4 章で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。5 章では本提案手法の関連研究を述べる。最後にまとめと今後の課題について述べる。

2. 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 1 章で示した既存の研究 [7] [8] と異なり、より一般性を持った定義となっている。そこへ相関規則マイニングで用いられる概念を導入して、目的となる比率規則の抽出手法を考える。

2.1　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。

特に本論文では、2 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, 1] となるよう正規化されているものとする。

以下、比率規則を抽出する対象とする 2 属性を X, Y とし、それぞれの属性値を $x, y (0 \leqq x, y \leqq 1)$ と表現する。

2.2　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。従って 2 属性 x_t, y_t を持つタプル t について、**t が $y_t = ax_t + b$ $(a, b \in \mathcal{R})$ を満たすとき、t はパラメータ a, b から成る比率規則に従う**と考えることが自然である。すなわち 2 属性 X, Y で張られる空間中の直線として比率規則を捉えることである。

しかし、単純に直線 $y = ax + b$ 上に複数のタプルが存在する状況は少ない。またパラメータ a, b の取り得る値はどちらも $(-\infty, \infty)$ の範囲における任意の実数であり、計算機中で扱う上で便利でない。そこで前者の問題には、パラメータに対する**許容誤差**を設定し、許容誤差内で異なる直線上に存在するタプルも、同一の比率規則に従うとする。後者の問題については、付録にて説明した **Hough 変換** [4] により変数変換を行う。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, 1] を取るよう正規化されているので、ρ, θ の値はそれぞれ $0 \leqq \rho \leqq \sqrt{2}$, $-\pi/2 \leqq \theta \leqq \pi$ であると見なすことができる。

一方で、複数のタプルが同じ比率規則に従っていても、そのタプルの密度は区間によって異なる。非常に多数のタプルが従う区間であれば、その比率規則が成り立つと確信を持って言えるが、ほとんどタプルが存在しない区間であれば、比率規則が

そもそも成り立つのか不明である。従って比率規則の定義には X または Y いずれかの属性値がどの区間に含まれるかを示す必要がある。

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 $t(x_t, y_t)$ $(x_t \in I, I \subseteq [0,1])$ 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 $|\epsilon_t| \leqq \epsilon, |\delta_t| \leqq \delta$

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。以下では誤解の無い限り、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。

2.3 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規則マイニング [2] と対応付けし、以下のように諸概念を定義する。

- 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**信頼度**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 $minconf$ を満たした上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 $minsup$ を満たした上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言える。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言える。

3. 提 案 手 法

本章では 2 章に挙げた最適確信度/サポート比率規則を抽出する手法を提案する。

3.1 基本的なアルゴリズム

最適確信度/サポート比率規則を求めようとする場合、一番の問題は最適確信度/サポート区間を求めることにある。しかしすべてのタプルについて比率規則 $RR_{[0,1]}(\rho_0, \theta_0)$ に従うかどうかの判定が成されていれば、パラメータ (ρ_0, θ_0) については 1 次元数値属性相関規則マイニング [2] 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られる。いま数値属性 X について、X の定義域中における区間 $I = [s, t](0 \leqq s \leqq t \leqq 1)$ 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、という事柄が **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。ここで条件 C を “比率規則が成り立つかどうか” と当てはめることで、1 次元数値属性相関規則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確信度/サポート区間抽出手法として、ここでは Fukuda らによる手法 [2] 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成されているとき、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 は全タプル数) で求めることができる。本論文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は図 3 のように考えられる。ρ と θ は許容誤差 ϵ, δ に基づきそれぞれ 2ϵ と 2δ 間隔の離散値となるようにする。ρ, θ の各離散値は ρ_i, θ_j(ただし $i = 1, \cdots, R$ $(R = \lceil (3\pi/2 + \epsilon)/2\epsilon \rceil)$, $j = 1, \cdots, T$ $(T = \lceil (\sqrt{2} + \epsilon)/2\epsilon \rceil)$) と表現する。ここで $\rho_1 = 0, \theta_1 = -\pi/2$ である。

```
for each (ρ_i, θ_j) do
    for each タプル t do
        t が RR_[0,1](ρ_i, θ_j) に従うか判定
    end
    最適確信度/サポート区間 I を
    1次元数値属性相関規則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 do
        RR_I(ρ_i, θ_j) を出力
    end
end
```

図 3 比率規則を求める基本的なアルゴリズム

3.2 基本的な手法の問題点と改良

上で述べた基本的な手法には 2 つの問題がある。一つはすべての (ρ_i, θ_j) の組に対して毎回全タプルを読み込み、1 次元数値属性相関規則マイニングを行う必要があるため、非常に計算量が大きくなることである。ほとんどのタプルが従わないパラメータについてもその 2 つの処理を行うため、無駄も非常に多い。もう一つの問題点は、本質的にはほぼ同一の比率規則が多数得られる可能性があることである。多数のタプルが共通して従うような比率規則は同一と考えるべきである。

この 2 つの問題を解決する方法として、最適確信度/サポート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。

3.3 枝刈りフェーズ

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数は少なくとも $support(RR_I(\rho_i,\theta_j)) \equiv support(I) \times (support(RR_I(\rho_i,\theta_j))/support(I)) \equiv support(I) \times conf(RR_I(\rho_i,\theta_j))$ 以上存在する必要がある。これは最小サポート $minsup$ と最小確信度 $minconf$ の積 $\alpha = minsup \times minconf$ 以上のタプルが比率規則に従う必要があることを表す。

そこで区間 I を任意とした形 $RR_{[0,1]}(\rho_i,\theta_j)$ において α 以上のタプルが従わないパラメータをフィルタリングする。具体的には、各タプルを通る直線のパラメータ (ρ,θ) を列挙し、そのヒストグラムから α 以上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。ただしパラメータのカウント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。従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始めから読み、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。

3.4 比率規則統合フェーズ

このフェーズでは、ある閾値以上の近さを持つ比率規則群を同一の規則として統合する。統合された比率規則群を**比率規則セット**と呼ぶことにする。２つの比率規則 $RR_{I_1}(\rho_i,\theta_j), RR_{I_2}(\rho_k,\theta_l)$ に対する近さの尺度としては、以下の式で表される **Jaccard 係数**を用いる。

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

ここで $t \in RR_I(\rho_i,\theta_j)$ は、タプル t が比率規則 $RR_I(\rho_i,\theta_j)$ に従うことを表す。この定義はすなわち、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。この値が一定以上のとき２つの比率規則は本質的に同一であるとみなして統合を行う。以下ではこの閾値を $minmerge$ と表記する。

この Jaccard 係数を単純に計算すると重複のチェックは、$RR_{I_1}(\rho_i,\theta_j)$ または $RR_{I_2}(\rho_k,\theta_l)$ いずれかの比率規則に従うタプルの全組み合わせだけ行う必要がある。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。従って記録されている順にタプルを読み込み、そのたびに比較を行うことで重複のチェックは高々 $RR_{I_1}(\rho_i,\theta_j)$ または $RR_{I_2}(\rho_k,\theta_l)$ いずれかの比率規則に従う全タプル回で押さえられる。

3.5 提案手法のまとめ

本手法の目的は単に最適確信度/サポート比率規則を抽出するのではなく、最適確信度/サポート比率規則セットを得るところにある。本提案手法は各比率規則セットを求めるため枝刈り、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 (ρ,θ) の組の数 $P \leqq RT$、最小確信度/サポートを満たす比率規則の数 $Q \leqq P$ 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 $\epsilon,\delta,minsup,minconf$ 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。

4. 実　　験

本実験では提案手法により得られる比率規則の妥当性と提案手法における計算量の性質を検討する。ここでは人工データと２種類の実データを用いる。いずれの実データも UCI の Machine Learning Repository [10] 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 CPU に Pentium III Xeon 1.0GHz を２つ、メインメモリに 2.0GB を持つ計算機を用いた。

4.1 データの概要

4.1.1 人工データの概要

本実験で扱う人工データは、全比率規則が p 個存在し、各比率規則に従うタプルを q 個とした場合を想定したように生成した。また各タプルはそれぞれ単一の比率規則に従うことを想定したので、全タプル数は pq 個である。

ある単一の比率規則に従うタプルは以下のようにして生成される。

（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 $x_i (1 \leqq i \leqq q)$ 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

ここで各パラメータは $0 \leqq \rho \leqq 1, -\pi/2 \leqq \theta \leqq \pi, 0 \leqq x_{min} \leqq x_{max} \leqq 1$ を満たし一様分布に従うよう生成する。

4.1.2 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, and Shell weight) のうち、Length と Shell weight の２属性を用いた。全タプル数は 4177 個である。

4.1.3 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。

4.2 妥当性の評価

まず提案手法により、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 にて示された例のデータである。

4.2.1 人工データ

図 4 は $(p, q) = (2, 500)$ の人工データと抽出された全比率規則セットを表す。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。各図において黒の点は各タプルを表し、色の付いた領域は得られた各比率規則セットが成り立つ領域を表す。同一色の領域は、単一の比率規則セット中に含まれる全比率規則が成立する領域を表す。

最適確信度比率規則の結果は、パラメータに $\epsilon = 0.05, \delta = 0.03, minsup = 0.2, minconf = 0.85, minmerge = 0.5$ を与えた場合の結果である。このパラメータ設定の結果、全部で 1200 個中 82 個の (ρ, θ) の組が枝刈りフェーズで残り、最終的には 3 つの比率規則セットが得られた。そのうち 2 つは図の両端、単一の線形関係のみが成り立っている部分に現れており、残り 1 つは 2 つの線形関係が交わっている部分に現れている。後者の部分は単一の線形関係と見なすことができるので、いずれの比率規則セットも妥当な結果であると考えられる。

最適サポート比率規則の結果は、パラメータに $\epsilon = 0.05, \delta = 0.03, minsup = 0.75, minconf = 0.55, minmerge = 0.5$ を与えた場合の結果である。このパラメータ設定の結果、全部で 1200 組中わずか 6 組の候補が枝刈りフェーズで残り、最終的には 2 つの比率規則セットが得られた。得られたいずれの比率規則セットも、単に線形関係を表すだけでなく、その線形関係に従うほとんどのタプルが含まれる領域を示している。

図 5 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セットを表している。最適確信度/サポート比率規則セットのいずれの結果も、パラメータは $\epsilon = 0.01, \delta = 0.1, minsup = 0.175, minconf = 0.35, minmerge = 0.01$ のように設定した結果である。全部で 1800 組の (ρ, θ) の中で枝刈りフェーズにより 26 組が残り、最終的には 4 つの比率規則セットが得られた。

この結果を見ると、いずれの種類の比率規則セットも単にタプルの分布を表現しているのではなく、X 軸について密度が高い部分、すなわち確信度が高い部分を表している。これは相関規則マイニングにおけるサポートと確信度に概念が考慮されているためである。

4.2.2 アワビデータ

図 6 はアワビデータに対する結果である。図の横軸は属性 “Length” の正規化された値を表し、縦軸は属性 “Shell weight” の正規化された値を表す。図の黒点と色のついた領域の意味は人工データと同様である。パラメータには $\epsilon = 0.008, \delta = 0.04, minsup = 0.3, minconf = 0.25, minmerge = 0.01$ を与えた。枝刈りフェーズの結果、5340 組中 60 組の候補が残り、最終的には単一の比率規則セットが得られた。

このデータでは、“Length” が 0.5 以下と 0.5 より大きな部分でタプルの分布が異なるが、本手法ではその変化に沿った比率規則が抽出されている。また最適確信度比率規則と最適サポート比率規則を比較すると、タプルの分布に対してはみ出している領域が多少異なることがわかる。この結果から、確信度とサポートをそれぞれ重視した比率規則が得られていることがわかる。

4.2.3 ワインデータ

図 7 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 $\epsilon = 0.06, \delta = 0.08, minsup = 0.5, minconf = 0.55, minmerge = 0.2$ とした。枝刈りフェーズの結果、(ρ, θ) 全 390 組中 6 組が残り、最終的に 2 つの比率規則セットが得られた。このデータでは $0.0 < Flavanoids < 0.4$ と $0.4 < Flavanoids < 0.7$ の各部分でタプルが従う線形関係が変化しているが、得られた比率規則セットは各線形関係に対応している。

以上より本提案手法は、適当なパラメータを設定することで妥当な最適確信度/サポート比率規則セットを得ることができたと考えられる。

4.3 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の処理時間について調べた。パラメータに関する実験では $(p, q) = (5, 2000)$ と $(p, q) = (10, 1000)$ となる、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。前者については妥当性の評価で用いたものと同様である。スケーラビリティの実験では上記 $(p, q) = (5, 2000)$ の人工データと同じ線形関係に従うよう、q の値のみを変化させたデータを生成した。またいずれの実験においても $minsup = minconf = minmerge = 0.1$ とした。

4.3.1 パラメータの影響

本実験では許容誤差のパラメータ 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, 1000 個になるようにしたもので、ϵ は 0.075, 0.0143, 0.00711, 0.001416, 0.0007075 の 5 通り、δ は 0.249, 0.0476, 0.02369, 0.00472, 0.002358 の 5 通りである。ϵ を変化させる場合には $\delta = 0.02369$、δ を変化させる場合には $\epsilon = 0.00711$ とした。

まず $(p, q) = (5, 2000)$ の場合の実験結果は図 8 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。各図の横軸は粒度を表すため、各パラメータの逆数とした。本提案手法において、ϵ の変化は直接計算量に関わらないが、枝刈りフェーズにおいて候補パラメータ (ρ, θ) の残る割合に影響がある (図 9 の左図)。ϵ が大きすぎても小さすぎても抽出される比率規則数が少なくなるので、ϵ が大小に偏っていないときに一番実行時間がかかったと考えられる。また δ に関しては、枝刈りフェーズで残るパラメータの割合がほぼ一定である (図 9 の右図) ので、実行時間は枝刈りの結果に依らずパラメータ θ の個数について線形で変化している。

また $(p, q) = (10, 1000)$ の場合の結果は図 10 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。

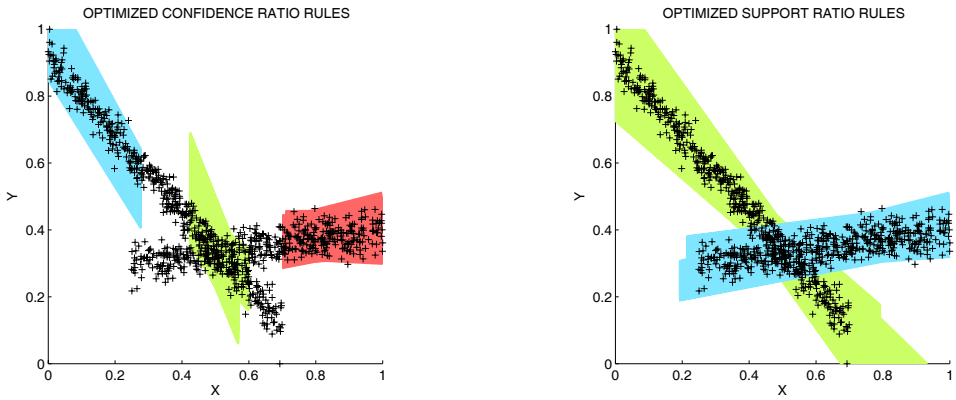

図 4　$(p, q) = (2, 500)$ の人工データに対する全比率規則

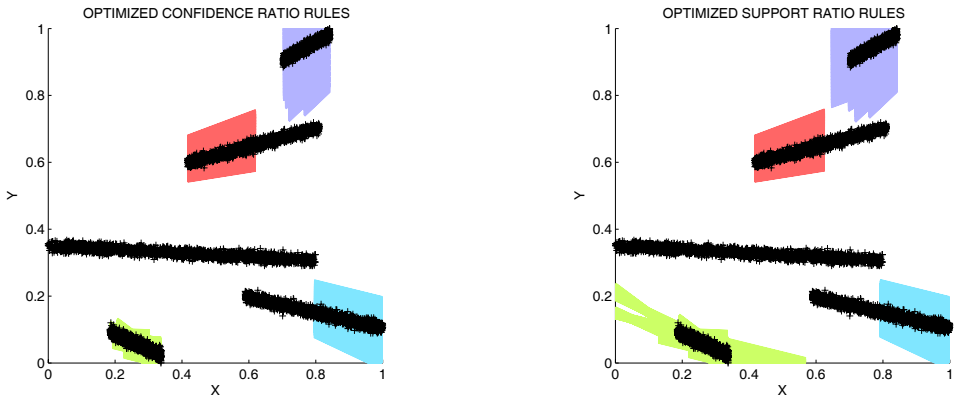

図 5　$(p, q) = (5, 2000)$ の人工データに対する全比率規則

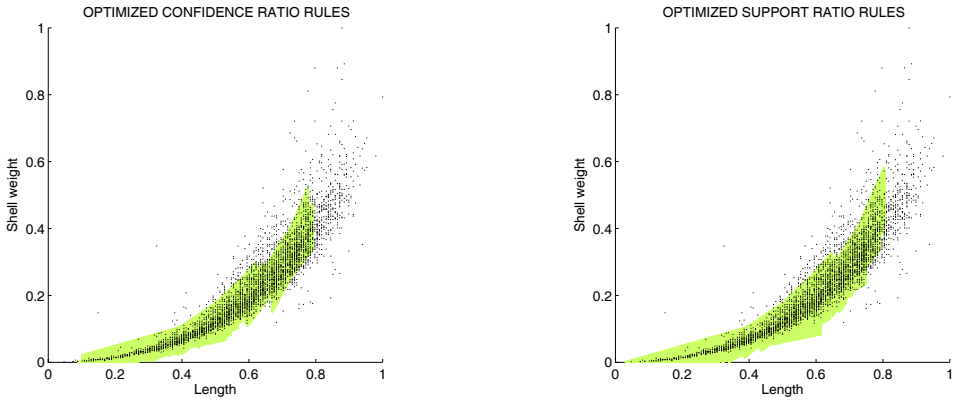

図 6　アワビデータに対する全比率規則

4.3.2　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 $\epsilon = 0.00711, \delta = 0.02369$ を与えた。

実験結果は図 11 である。3.5 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わかった。

5. 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 [8] [9]、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。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。

別の手法として非負スパースコーディング [5] を元にした手法がある [6] [7]。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。このような場合には各比率規則は直交しないが、非負スパースコーディングを用いることで妥当な比率規則を得ることができる。

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、

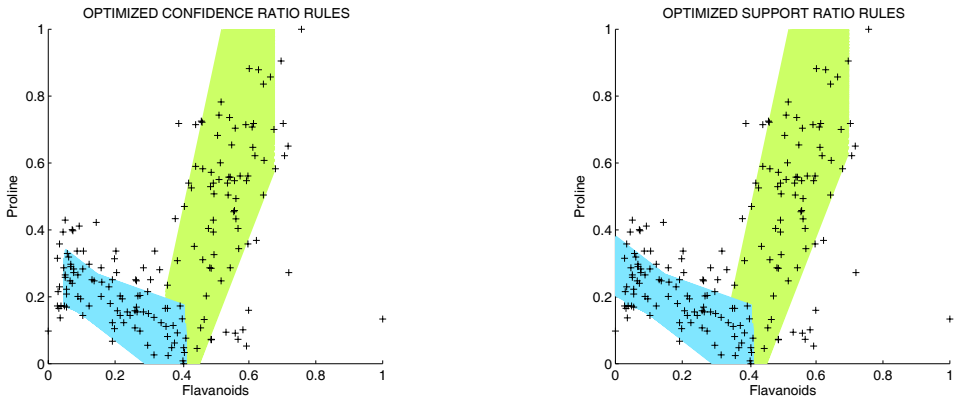

図 7　ワインデータに対する全比率規則

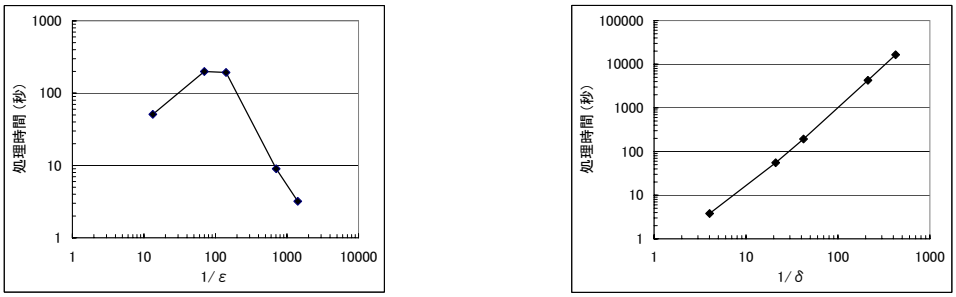

図 8　$(p, q) = (5, 2000)$ の人工データにおける実行時間

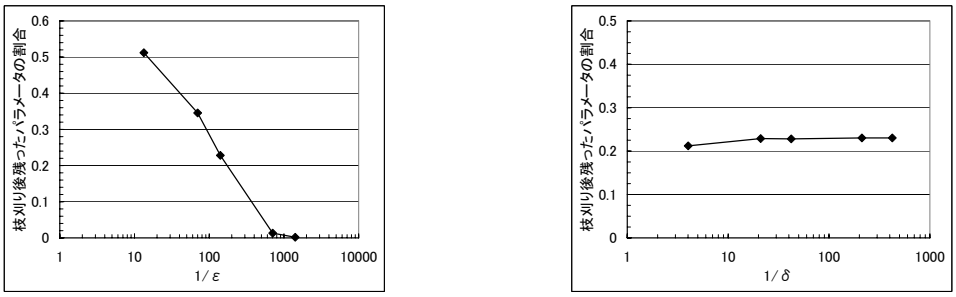

図 9　$(p, q) = (5, 2000)$ の人工データにおける、枝刈りフェーズ後残ったパラメータ (ρ, θ) の割合

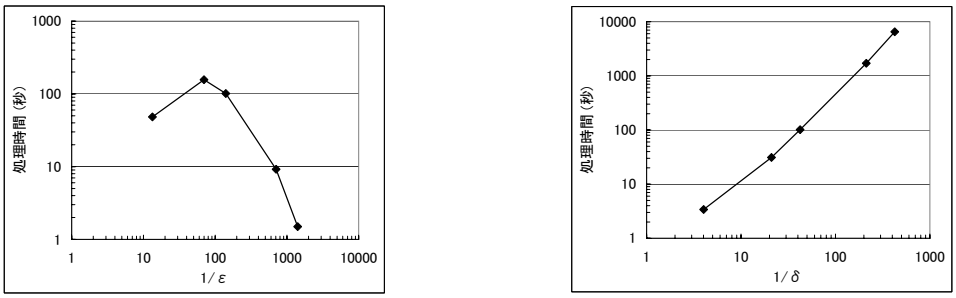

図 10　$(p, q) = (10, 1000)$ の人工データにおける実行時間

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 はタプル数、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。

これら関連研究と本研究では比率規則の定義が根本的に異なっている。本研究において比率規則は分析対象の 2 属性で張られる 2 次元空間中の線分として表されるが、関連研究では空間ベクトルとしてのみ表される。すなわち本研究の定義は既存の研究で用いられている定義をより一般化したものとなっている。

また、既に述べたように既存の研究ではサポートや確信度の概念はない。データ中から線形関係を抽出する問題は、回帰分析や主成分分析のような多変量解析の対象ともなっている [3]。既存研究 [8] [9] はこの流れのものである。既述のように、これらにはサポートや確信度といった概念はない。また、回帰分析

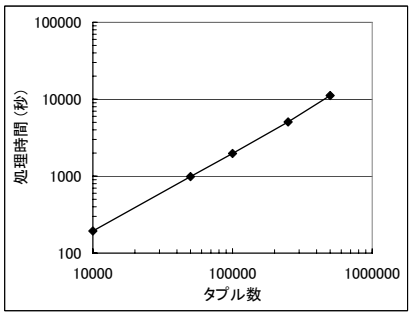

図 11　スケーラビリティの実験結果

や主成分分析では線形関係を抽出する対象データの選択はユーザにゆだねられている。これに対して、本研究では各比率規則とそれに従うデータの部分集合の抽出が一体として行われる点が特徴的である。

6. おわりに

本論文では比率規則抽出する手法を提案した。特に相関規則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。この提案手法を人工データと２種類の実データによる実験で妥当性を確認し、人工データによる実験により大規模なデータについても適用可能であることを確認した。

今後の課題としては、与えられたデータに応じてパラメータを自動的にチューニングするような手法の開発、分散処理を導入して高速に比率規則を抽出する手法の開発、３属性以上の間における比率規則の抽出手法の検討などが考えられる。

謝辞

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

文　献

[1] R. Duda and P. Hart, “Use of the hough transformation to detect lines and curves in pictures,” Communications of the ACM, vol.15, no.1, pp.11–15, 1972.

[2] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, “Mining optimized association rules for numeric attributes,” Proc. ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Montreal Quebec, Canada, pp.182–191, 1996.

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer-Verlag, New York, 2001.

[4] P. Hough, “Methods and means for recognizing complex patterns,” U.S. Patent 3,069,654.

[5] P. Hoyer, “Non-negative sparse coding,” Proc. Workshop on Neural Networks for Signal Processing, Martigny, Switzerland, pp.557–565, 2002.

[6] 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.

[7] 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.

[8] 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.

[9] F. Korn, A. Labrinidis, Y. Kotidis, and C. Faloutsos, “Quantifiable data mining using ratio rules,” VLDB Journal, vol.8, pp.254–266, 2000.

[10] D. Newman, S. Hettich, C. Blake, and C. Merz, “UCI repository of machine learning databases,” 1998.

付　録

直線 $y = a_0x + b_0$ 上にある点群から、その直線を検出する問題を考える。一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 $(-\infty, \infty)$ の値を取るため 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 [1] [4] は、無限の区間を取る２パラメータ a, b を、有限の区間を取る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 A·1 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。

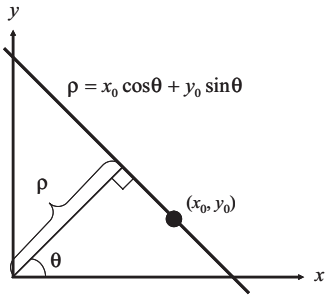

図 A·1　Hough 変換における各パラメータの関係

本論文では属性 X, Y はドメインが区間 $[0, 1]$ 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 $[0, \sqrt{2}], [-\pi/2, \pi]$ 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 $0 \leqq \rho \leqq \sqrt{2}, -\pi/2 \leqq \theta \leqq \pi$ 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 [1] は以下の通りである。

（１）　ρ, θ で張られる２次元空間を分割する。分割幅はユーザが与えるものとする。

（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