DEWS2007 L4-5

対称比率規則の抽出手法

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　数値属性間で成り立つ線形関係を表した比率規則は，データの理解補助，欠損値の補完などの幅広い有用な応用が可能である．われわれの先行研究である，2 属性の数値データに対する比率規則には，相関ルールマイニングと対応づけられたサポートと確信度の概念が導入されている．しかしその定義において 2 属性は非対称に扱われるため，属性の役割を入れ替えて抽出された比率規則は一致せず，それらが表す線形関係も異なりうる．本論文では 2 属性を対象に扱うような比率規則である，対称比率規則の定義を行い，その抽出手法を提案する．この提案手法について人工データと実データを用いた実験を行い，提案手法の妥当性を示す．
キーワード　データマイニング，比率規則，知識発見，線形関係

Mining Symmetric Ratio Rules

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　Ratio Rules represent linear relatioships among numeric attributes, and are applicable to data understanding support, filling in missing values, and related issues. Our privious work introduced some concepts such as support and confidence, which are used in association rule mining, to Ratio Rules for two attributes. However, extracted Ratio Rules or linear relationships represented by them do not coincide when the attributes are swapped, because two mining target attributes are dealt asymmetrically in the definition. In this paper we define the Symmetric Ratio Rules, which deal target attributes symmetrically, and propose a method to extract them. We also show appropriateness of the proposed method using synthetic and real data.
Key words　data mining, Ratio Rules, knowledge discovery, linear relationship

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．例えば相関規則マイニング，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 マイニングなどが挙げられる [1]．これに対し本研究では特に**比率規則** [2] 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．この直線のように，データの線形関係を表すものが比率規則となる．比率規則は [2] で示されているように，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．

既存の手法として，各タプルは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捉える手法がある [2] [3]．いずれの手法も，比率規則を行列分解により得られる特徴ベクトルとして表す．言い換えると原点を通る直線とし

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
...	...	...

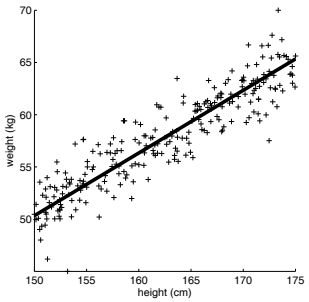

図 1　表 1 のデータに対する比率規則の例．実線が比率規則を表す.

て比率規則が表される．それゆえ一部分でのみ成立する線形関係などは捉えることが難しい.

一方でわれわれは先行研究において，比率規則を線分とその周辺領域として定義し，比率規則の抽出を相関ルールマイニングの問題と対応づけた [4]．このアプローチでは 2 種類の数値属性 X，Y について，各タプルが比率規則の表す領域に含まれるかを調べる．このとき領域に含まれるタプルは比率規則に従うと考える．そ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，比率規則に従うタプル数などの関係から定義した．最終的に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，かつサポートあるいは確信度が最大となるような比率規則を抽出する.

ただしこの先行研究における定義では，対象となる 2 種類の数値属性は非対称に扱われる．これは比率規則の成り立つ区間を表すのに，一方の属性における区間のみを用いているためである．こうすることでマイニングが高速に行える利点などを持つが，数値属性を入れ替えた結果は一致せず，得られた比率規則が表す線形関係も一致しない．また同一区間中に複数の線形関係が存在する場合，それらを個々に抽出するためには最小確信度を下げる必要があるため，より不要な比率規則が多数生成される可能性がある.

本論文では比率規則の抽出対象となる 2 属性を対称に扱う，**対称比率規則**を定義する．この対称比率規則は比率規則が成り立つ部分を，抽出対象の 2 属性両方を用いて表す．これにより抽出対象の 2 属性の役割によらず同様の線形関係を表すことが出来る．この対称比率規則について諸概念の定義および抽出手法の提案を行い，実験によりその手法の妥当性を示す.

本論文は以下のように構成される．2 章では関連研究について述べ，それらと比較した対称比率規則の特徴を示す．3 章では先行研究である非対称比率規則の定義をもとに，対称比率規則とその関連概念を定義する．4 章では条件を満たす対称比率規則の抽出手法を提案する．5 章では人工データと実データを用いた実験結果を示す．最後にまとめと今後の課題について述べる.

2. 関 連 研 究

比率規則の抽出に関するアプローチには主に 2 種類ある．一つは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抽出する方法である．具体的には，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 $R = [\mathbf{r_1}, \cdots, \mathbf{r_k}]$，タプル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 $X \approx RV$ となるような行列分解を行う．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．比率規則は空間ベクトルとして表されるので，いずれの規則も原点を通るという制約がある.

具体的な行列分解手法としては，主成分分析を用いる手法 [2] [5] と，非負行列分解 [6] を元にした手法 [3] [7] がある．主成分分析を用いた手法では，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．非負スパースコーディングを元にした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.

もうひとつのアプローチとしてはわれわれの先行研究である，2 次元空間中の線分とその周辺領域として比率規則を捉える手法がある [4]．得られる線形関係は 2 属性間のみに限られるが，比率規則を線分として考えることで前者のアプローチより一般化された定義となっている．また数値属性に関する相関ルールマイニング [8] と対応付けられた，サポートや確信度の概念を導入している．これにより得られる規則がどのような性質を持っているかの説明づけがなされるほか，サポートや確信度の最小値を調節することで，ユーザの意向を結果に反映させることができる.

ただしこの定義上，扱う 2 属性は非対称である．つまり 2 属性 $< X, Y >$ から得られる規則と，役割を入れ替えた 2 属性 $< Y, X >$ から得られる規則は同一とならない．またマイニング対象の 2 属性のうち，一方の分布は考慮されない．この理由については次章で述べる.

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が [9]，回帰分析や主成分分析では線形関係に従う対象データの選択はユーザにゆだねられている．行列計算によるアプローチはこの流れを汲んでいる．これに対して線分によるアプローチは，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．

本論文では先行研究における非対称性の問題を取り除くため，定義の拡張を行った．定義される対称性を持った比率規則は，属性の役割を入れ替えても同様の線形関係を表すことができる．また抽出対象の 2 属性両方に関する分布を考慮するため，多数の線形関係が同一の区間に含まれる場合に，得られる比率規則に過不足が起こりにくくなっている．

3. 比 率 規 則

本章では対称比率規則の定式化を行う．以下ではまず本研究の対象データおよび先行研究の非対称比率規則を説明し，続いて対称比率規則とその関連概念を定義する．

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．

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 $[-0.5, 0.5]$ 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 $x, y(-0.5 \leqq x, y \leqq 0.5)$ と表現する．

3.2 非対称比率規則の定義

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．したがって分析対象の 2 属性 X, Y で張られる空間を考えたとき，直線 $y = ax + b$ $(a, b \in \mathcal{R})$ として比率規則を考えることが最も単純である．

しかし，直線 $y = ax + b$ 上に厳密な意味で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るため，Y 軸にほぼ平行な直線を取り扱う際に a, b が無限大に発散する．そこで前者の問題には，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 **Hough 変換** [10] により，パラメータが有限区間を取るよう変数変換を行う．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 $[-0.5, 0.5]$ を取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 $[0, \sqrt{2}/2]$, $[0, 2\pi]$ で押さえられる．

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ず，属性 X または Y がある区間に含まれる場合のみ線形関係が成り立つと考えられる．よって比率規則の定義には属性がどの区間に含まれるかを示す必要がある．非対称比率規則では属性 X だけに注目する．

以上の点から，非対称比率規則は図 2 のように定義される．

以下では誤解の無い限り，非対称比率規則 $RR_{x\in I}(\rho\pm\epsilon, \theta\pm\delta)$ 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 対称比率規則

前節で述べた非対称比率規則の定義では，線分を表す際に 1 属性のみに注目している．しかし 2 属性 X と Y は対称ではなく，両者の役割を入れ替えると得られない規則がある．例として X 軸に平行な規則 $RR_I(\rho, 0)$ を考える．この規則は 2 属性の役割を入れ替えると，Y 軸に平行で $y \in I$ を満たす区間で成り立つ規則となるべきであるが，非対称比率規則ではこの規則を表すすべが無い．

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

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

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

図 2　非対称比率規則の定義

この原因は，定義中に属性 Y のどの区間で成り立つか示されていないためである．そこで非対称比率規則の定義を拡張し，属性を入れ替えても同様に表すことのできる比率規則を，対称比率規則として図 3 のように定義する．

タプル $t(x_t, y_t)$ $(x_t \in I, y_t \in J, I, J \subseteqq [-0.5, 0.5])$ **が以下の式を満たす値** ϵ_t, δ_t **を持つとき，** t **は対称比率規則** $SRR_{<x\in I, y\in J>}(\rho \pm \epsilon, \theta \pm \delta)$ **に従う．**

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

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

図 3　対称比率規則の定義

以下誤解のない限り，短縮した形 $SRR_{<I,J>}(\rho, \theta)$ として対称比率規則を表現する．

3.4 対称比率規則の諸概念

非対称比率規則では，相関ルールマイニングと対応付けられたサポートや確信度などの概念を持っている．これらの概念は対称比率規則にも同様に導入できる．以下にその諸概念を定義する．

- サポート：対称比率規則 $SRR_{<I,J>}(\rho, \theta)$ のサポートを，全タプルに対する，対称比率規則 $SRR_{<I,J>}(\rho, \theta)$ に従うタプルの割合と定義し $support(SRR_{<I,J>}(\rho, \theta))$ で表す．また領域 $< I, J >$ のサポートは，全タプルに対する，領域 $< I, J >$ に含まれるタプルの割合と定義し $support(< I, J >)$ で表す．
- 確信度：対称比率規則 $SRR_{<I,J>}(\rho, \theta)$ の確信度を，$support(< I, J >)$ に対する $support(SRR_{<I,J>}(\rho, \theta))$ の割合 $support(SRR_{<I,J>}(\rho, \theta))/support(< I, J >)$ で定義し，$conf(SRR_{<I,J>}(\rho, \theta))$ で表す．
- 最小サポート，最小確信度：抽出される対称比率規則に対して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．以下ではそれぞれ $minsup$, $minconf$ で表す．
- 最適確信度対称比率規則：$support(< I, J >)$ が $minsup$ を満たし，かつ $conf(SRR_{<I,J>}(\rho, \theta))$ が $minconf$ を満たした上で最大となるような対称比率規則 $SRR_{<I,J>}(\rho, \theta)$．最大値を与える領域 $< I, J >$ を最適確信度領域と呼ぶ．

- 最適サポート対称比率規則：$conf(SRR_{<I,J>}(\rho,\theta))$ が $minconf$ を満たし，かつ領域 $<I,J>$ が $minsup$ を満たした上で最大となるような対称比率規則 $SRR_{<I,J>}(\rho,\theta)$．最大値を与える領域 $<I,J>$ を最適サポート領域と呼ぶ．

最適確信度対称比率規則は，領域 $<I,J>$ に含まれるタプル数が一定数以上である条件のもと，領域 $<I,J>$ に含まれれば確実に成り立つような対称比率規則である．これを見つけることで，より強い線形関係がわかる．また最適サポート対称比率規則は，領域 $<I,J>$ 中で対称比率規則に従うタプルの割合が一定以上である条件のもと，なるべく多くのタプルが含まれる領域 $<I,J>$ を持つ対称比率規則である．これを見つけることで，より広い領域で成り立つ線形関係がわかる．

以下，最適確信度対称比率規則と最適サポート対称比率規則，および最適確信度領域と最適サポート領域をまとめてそれぞれ最適対称比率規則，最適領域と呼ぶ．

4. 提案手法

本章では最適対称比率規則を抽出するするための手法を提案する．

本手法ではすべてのタプル（全タプル数を N とする）がメインメモリに乗ることを仮定し，その仮定のもとで最適領域の厳密な解を求める．またパラメータ ρ,θ はそれぞれ 2ϵ，2δ ごとの離散値 $\rho_1,\rho_2,\cdots,\rho_R$ および $\theta_1,\theta_2,\cdots,\theta_T$（$\rho_1=0,\theta_1=0$）として扱う．

求める最適対称規則は，サポートや確信度が全体で最も大きい1つではなく，各パラメータ組 (ρ_p,θ_q) ごとに求める．

4.1 単純な手法

単純な手法として，考え得るすべての対称比率規則 $SRR_{<I,J>}(\rho,\theta)$ について，その対称比率規則に従うタプル数と領域 $<I,J>$ に含まれるタプル数をカウントする手法が考えられる．具体的には，ρ,θ の二重ループの内部で，考え得るすべての区間 I,J の組み合わせ $<I,J>$ について，各タプルが対称比率規則 $SRR_{<I,J>}(\rho,\theta)$ に従うか，および領域 $<I,J>$ に含まれるかを判定する．

このとき区間 I,J はいずれも $O(N^2)$ 個存在する．従ってこの手法の計算量は $O(RTN^4)$ と非常に膨大なため，適用は現実的に難しい．

4.2 提案手法

われわれが提案した非対称比率規則の抽出手法[4]では，単純な列挙の代わりに Fukuda らにより提案された1次元数値属性相関ルールマイニング[8]を用いることで，計算量を $O(RTN^2)$ から $O(RTN)$ へと減らした．このアプローチと同様，2次元数値属性相関ルールマイニング[11]を用いることで，最適区間を $O(RTN^3)$ で求めることができる．

2次元数値属性相関ルールマイニングでは，“2種類の数値属性 X,Y が各数値属性で張られる2次元空間中の領域 L に含まれるならば，条件 C を満たす” という2次元数値属性相関ルールを考える．このとき全体に対して L に含まれるタプルの割合を L のサポート，L に含まれるタプル中条件 C を満たすタプルの割合を確信度とし，サポートあるいは確信度を最大とする領域 L を見つけることを行う．本研究においては L として矩形領域を，C として “対称比率規則を満たす” という条件をそれぞれ与えることで，2次元数値属性相関ルールマイニングを適用することが出来る．

```
/* 枝刈りフェーズ */
パラメータ組 (ρ, θ) の候補の枝刈り
属性 Y に対するソート処理

/* 対称比率規則生成フェーズ */
for each 残りの候補 (ρ_i, θ_j) do
    候補集合 = {}
    for each 区間 J_k do
        最適サポート/最適確信度区間 I を抽出
        if SRR_<I,J_k>(ρ_i, θ_j) が minsup および
        minconf を満たす then
            SRR_<I,J_k>(ρ_i, θ_j) を候補集合に追加
        end
    end
    候補集合中，確信度・サポートが最大の対称比率規則を出力
end
/* 対称比率規則統合フェーズ */
for each 類似した組
SRR_<I_a,J_b>(ρ_i, θ_j), SRR_<I_c,J_d>(ρ_k, θ_l) do
    RR_<I_a,J_b>(ρ_i, θ_j) と RR_<I_c,J_d>(ρ_k, θ_l) を
    同じ集合 S_q に統合
end
対称比率規則集合 {S_1, S_2, ...} を出力
```

図4　提案する対称比率規則抽出手法

2次元数値属性相関ルールマイニングのアルゴリズムでは2種類の属性のうち，片側（以下属性 X とする）については先に述べた1次元数値属性相関ルールマイニングを用いることで，$O(N)$ で最適な区間を得ることができる(注1)．もう片側（以下属性 Y とする）についてはすべての候補を列挙する必要がある．列挙が必要な属性 Y について，その候補数は $N(N-1)/2$ 個であるため，候補数の削減が望まれる．そこで定義に照らし合わせ，次節で述べる2種類の枝刈りを行い候補数を削減する．

また対称比率規則をそのまま結果として出力すると，規則数が多い場合ユーザの結果の理解が困難になると考えられる．そこで非対称比率規則の抽出手法[4]でも用いられている，類似した対称比率規則の統合を行う．最終的には，類似した規則が含まれる集合である，対称比率規則集合が結果として出力する．

以上をまとめたアルゴリズム全体を，図4に擬似コードとして示す．アルゴリズムは，枝刈り，対称比率規則生成，対称比率規則統合の3フェーズから構成される．枝刈りフェーズは上に述べたように候補数の削減を行う．対称比率規則生成フェーズでは条件を満たす最適対称比率規則を生成する．対称比率規則統合フェーズでは，類似した規則を同じ集合に統合する．

以下，枝刈りフェーズと対称比率規則統合フェーズについて説明する．

(注1)：入力タプルが数値属性でソートされている場合．本論文でもユーザより与えられるデータは X でソートされているものとみなす．

4.2.1 枝刈り手法

（1） 対称比率規則に従うタプル数による枝刈り

この枝刈りは非対称比率規則の抽出手法 [4] で用いられている枝刈りとほぼ同様に，パラメータ ρ,θ に対する枝刈りを行う．抽出すべき対称比率規則 $RR_{<I,J>}(\rho,\theta)$ に従うタプル数は，与えられた最小サポート $minsup$ と最小確信度 $minconf$ に対して次の式を満たす．

$$\begin{aligned} & N \times support(RR_{<I,J>}(\rho,\theta)) \\ & = N \times conf(RR_{<I,J>}(\rho,\theta)) \times support(<I,J>) \\ & > N \times minsup \times minconf \end{aligned}$$

左辺の最大値 $N \times support(RR_{<[-0.5,0.5],[-0.5,0.5]>}(\rho,\theta))$ は全空間中で対称比率規則に従うタプル数を表す．この値が最小サポートと最小確信度の積未満の場合，それよりも小さな空間 $<I,J>$ の比率規則は条件を満たすことはないので，このときのパラメータ組 (ρ,θ) は枝刈りできる．

（2） 一定幅未満の候補に対する枝刈り

列挙を行う側の区間が非常に狭い場合，もう一方の区間を最大限広くしても最小サポートを満たさないことが考えられる．そのような狭すぎる区間に対する枝刈りを行う．

具体的には属性 Y でソートを行い，$j-i+1 \geqq N \times minsup, i \leqq j$ を満たさない区間 $[y_i, y_j]$ を枝刈りする．本論文ではすべてのタプルがメインメモリに乗る程度のタプル数を想定しているので，手法全体におけるソート処理のコストは大きくないと仮定する．

4.2.2 条件を満たすタプルのチェック手法

2次元数値属性相関ルールマイニングを行う際，属性 X の最適区間 I を求めるためにはあらかじめチェックすべきことがある．それは列挙された属性 Y の区間 J にどのタプルが含まれ，またどのタプルが対称比率規則 $SRR_{<[-0.5,0.5],J>}(\rho_p,\theta_q)$ に従うかである．このとき各区間 J についてすべてを調べなおすのは効率が悪い．

提案手法では，最小サポートを満たす区間が最小でも $\lceil N \times minsup \rceil$ 個のタプルを含むことと，枝刈り処理のために属性 Y でソートされた結果を持つことを利用する．まず区間 J を $[y_1, y_N]$ から $[y_1, y_{\lceil N \times minsup \rceil}]$ まで列挙し，属性 X の最適区間 I を求める．続いて $[y_2, y_{\lceil N \times minsup \rceil+1}]$ より $[y_2, y_N]$ まで列挙する．以下奇数は y_N からの縮小，偶数は最小の区間からの拡大を行う．

区間 J を伸縮する際，区間内に新たに入ったり外れたりするタプル数は，縮小から拡大に変わるときを除き1つである．あらかじめ $SRR_{<[-0.5,0.5],[-0.5,0.5]>}(\rho_p,\theta_q)$ に従うタプルを調べ，その結果を持っておけば，伸縮の際に新たに最適比率規則に従うかどうかを調べなおす必要はない．

このアルゴリズムを用いると全区間数 $(N+1)N/2$ に対し，調べる区間数は $B = N - \lceil N \times minsup \rceil + 1, 0 < minsup$ として $B(B+1)/2$ 個となる．ここで B は始めに縮小を行うときに列挙される区間の個数である．

4.2.3 対称比率規則統合フェーズ

対称比率規則統合フェーズでは，2つの対称比率規則 $SRR_{<I_a,J_b>}(\rho_i,\theta_j)$,$SRR_{<I_c,J_d>}(\rho_k,\theta_l)$ の類似度を，以下の式で表される **Jaccard 係数**で測る．

$$\frac{|SRR_{<I_a,J_b>}(\rho_i,\theta_j) \cap SRR_{<I_c,J_d>}(\rho_k,\theta_l)|}{|SRR_{I_a,J_b}(\rho_i,\theta_j) \cup SRR_{I_c,J_d}(\rho_k,\theta_l)|}$$

$|SRR_{<I,J>}(\rho_i,\theta_j)|$ は，対称比率規則 $SRR_{<I,J>}(\rho_i,\theta_j)$ に従うタプル数を表す．すなわち，類似度は2つの対称比率規則の両方に従うタプルの，いずれかの対称比率規則に従うタプルに対する割合で表される．この値が閾値以上のとき2つの対称比率規則は同一の対称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．

5. 実　　験

本実験では提案手法により得られる対称比率規則の妥当性を検討する．妥当性は，与えられたタプルの分布に対しどのような比率規則が得られるかによって測る．また比較対象として，非対称比率規則を抽出した結果も示す．

5.1 データの概要

本実験では人工データと実データの2種類を用いた．実データには UCI Machine Learning Repository(注2)で配布されている Automobile データを使用した．以下，それぞれのデータについて説明する．

5.1.1 人工データ

本実験で扱う人工データは，データ中に3種類の線形関係が存在し，それぞれに 200 個のタプルが従うよう生成した．各線形関係に従うタプルは，以下の手続きにより生成を行った．

（1） パラメータ ρ,θ と区間 $I = [x_{min}, x_{max}]$ を適当に与える

（2） 区間 I 内で一様に分布するよう，属性値 $x_i (1 \leqq i \leqq 200)$ 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

5.1.2 実データ:自動車データ

このデータには 398 台の自動車に関する，燃費・重量・馬力など計 8 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，排気量と馬力の 2 属性を対象とした．ただしいずれの属性値も区間 $[-0.5, 0.5]$ を取るよう正規化した．

5.2 実験結果

実験結果として各比率規則を 2 次元空間上に示す．個々の対称比率規則は濃いグレーの線分，薄いグレーの近傍，および最適領域 $<I,J>$ を表す矩形で表現する．非対称比率規則の場合，最適領域は一方の属性しか与えられないので，もう一方は任意（すなわち区間 $[-0.5, 0.5]$）として矩形を表現する．また同一の比率規則集合に含まれるものはすべて同じ色で示し，異なるものは近傍領域の部分の濃さを変えて示す．

(注2)：http://www.ics.uci.edu/~mlearn/MLRepository.html

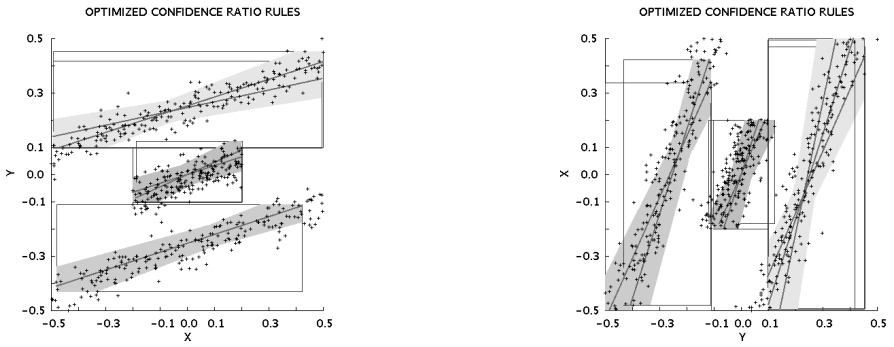

図 5　人工データより最適確信度対称比率規則を抽出した結果．属性を入れ替えてもほぼ同様の結果が得られている

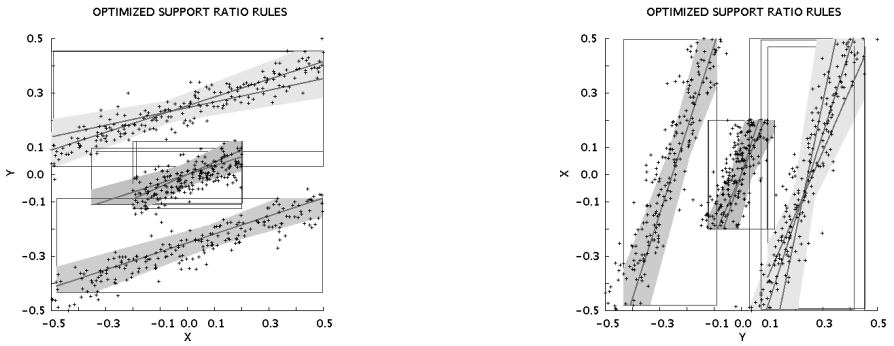

図 6　人工データより最適サポート対称比率規則を抽出した結果.

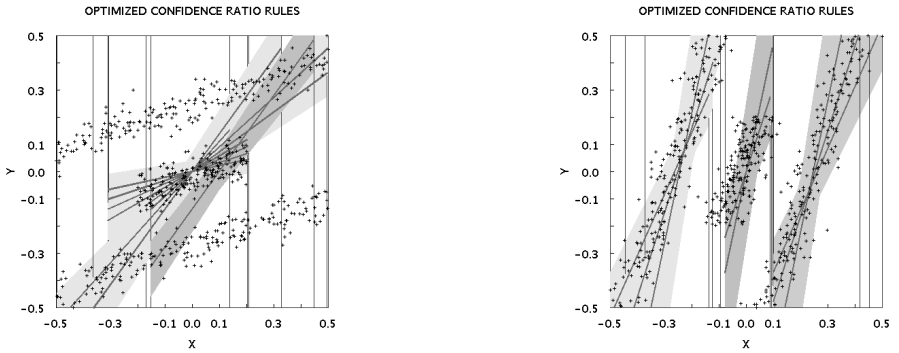

図 7　人工データより最適確信度非対称比率規則を抽出した結果

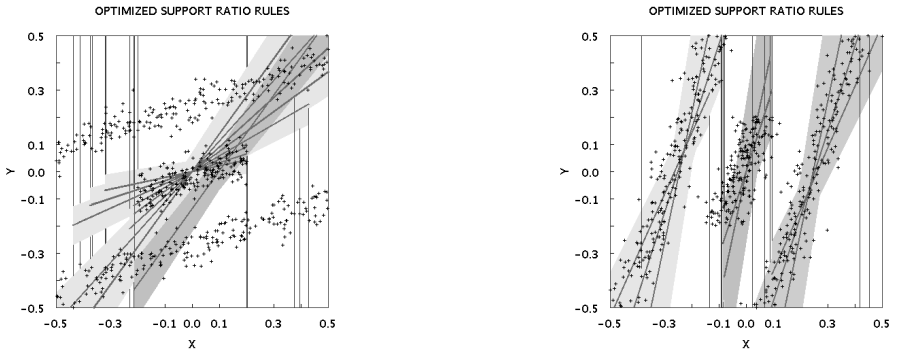

図 8　人工データより最適サポート非対称比率規則を抽出した結果

5. 2. 1　人工データ

人工データより抽出した最適確信度対称比率規則を図 5 に，最適サポート対称比率規則を図 6 に示す．また左右の図は，対称比率規則の特徴である対称性を示すため，属性の役割を入れ

替えた場合における結果である．いずれの結果も，パラメータは $\epsilon = 0.04$, $\delta = 0.0524$, $minsup = 0.3$, $minconf = 0.725$, $minmerge = 0.3$ として得られた．このとき，パラメータ (ρ, θ) の候補は 610 通りあるが，枝刈りの結果 30 通りが残った．

実験結果から，データ中に含まれている線形関係を捉え，かつ属性の役割によらずほぼ同じ結果を出力していることがわかる．一部の対称比率規則は，タプルの実際の分布よりも広く領域を捉え，ほぼ垂直に成り立つ線形関係と重なっている．これは全くタプルが存在しない部分は出力される結果に影響しないためである．また最適確信度対称比率規則と比べ，最適サポート対称比率規則は最適領域 $< I, J >$ が広くなっている．ただしその分他の対称比率規則集合に含まれるタプルがその領域に含まれている．

一方，同じデータから得られた最適確信度非対称比率規則および最適サポート非対称比率規則はそれぞれ図 7，図 8 である．左図のパラメータは $\epsilon = 0.04$, $\delta = 0.0524$, $minsup = 0.7$, $minconf = 0.3$, $minmerge = 0.3$ であり，右図は $minsup$ と $minconf$ をそれぞれ 0.3，0.7 に変えたものである．どちらの図も，水平方向の属性が最適区間を求める対象となっている．

図を見てわかるとおり，属性を入れ替えると得られる比率規則は大きく変わる．特に言えることは，垂直方向の分布を考慮しないため，タプルの分布が密な部分と疎な部分を分けられないことがある．そのため図 7 および図 8 の左図では密な部分を通る比率規則が過度に得られている．また非対称比率規則の表現に垂直方向の区間が示されないため，タプルが存在しない部分まで表現されている．

5.2.2 実データ

実データより抽出した最適確信度対称比率規則を図 9 に，最適サポート対称比率規則を図 10 に示す．人工データと同様に，左右は，属性の役割を入れ替えた場合における結果である．いずれの結果も，パラメータは $\epsilon = 0.045$, $\delta = 0.04$, $minsup = 0.375$, $minconf = 0.8$, $minmerge = 0.3$ として得られた．720 個のパラメータ組 (ρ, θ) に対する枝刈りの結果，37 個が残った．このデータは，タプルが密に分布している部分とそうでない部分にそれぞれ線形関係が存在するが，実験結果から対称比率規則はいずれも捉えていることが分かる．

これに対し最適確信度非対称比率規則の結果が図 11，最適サポート非対称比率規則が図 12 である．左図のパラメータは $\epsilon = 0.045$, $\delta = 0.04$, $minsup = 0.35$, $minconf = 0.75$, $minmerge = 0.3$ であり，右図は $minconf$ を 0.7 に変えたものである．どちらの図も，水平方向の属性が最適区間を求める対象となっている．実験結果を見ると，疎な部分の線形関係は捉えられているが，密な部分については垂直方向にタプルが無い部分にまで伸びる比率規則が得られていることがわかる．これは属性を入れ替えても同様であり，非対称比率規則はタプルの分布の疎密に影響を受けやすいといえる．

以上の結果より，対称比率規則は非対称比率規則の欠点を補い，妥当な線形関係を捉えることが出来ると考えられる．

6. おわりに

本論文では属性を対称に扱う対称比率規則を定義し，その抽出手法を提案した．既存の非対称比率規則では，抽出対象の 2 属性の役割を入れ替えると表現できない線形関係がある．一方対称比率規則は，属性の役割を入れ替えても表現される線形関係は変化しない．この対称比率規則に，サポートや確信度など相関ルールマイニングで用いられる概念を与え，サポートあるいは確信度を最大にする最適対称比率規則を抽出する手法を提案した．提案手法は，枝刈り，対称比率規則生成，対称比率規則統合の各フェーズからなる．最終的には類似した対称比率規則の集合である，最適対称比率規則集合が生成される．この提案手法について人工データと実データを用いた実験を行い，得られた結果が妥当であり，既存の非対称比率規則の問題点を解消していることを示した．

今後の研究のひとつとして，3 属性以上の高次元データに対する適用手法の検討があげられる．本論文における 2 属性に対する対称比率規則は，2 次元空間中の矩形領域を考えているが，これを直方体など高次元に拡張する必要がある．このほかの課題としては，より効果的な枝刈り手法や，本提案手法でも最悪の場合 $O(RTN^3)$ かかるためより高速なマイニング手法の検討などが挙げられる．

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．

文　献

[1] J. Han and M. Kamber: “Data Mining: Concepts and Techniques”, Morgan Kaufmann, San Francisco (2001).

[2] 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).

[3] 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).

[4] 濱本，北川：“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，**47**, SIG19(TOD32), pp. 54–71 (2006).

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

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

[7] 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).

[8] 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).

[9] T. Hastie, R. Tibshirani and J. Friedman: “The Elements of Statistical Learning”, Springer-Verlag, New York (2001).

[10] P. Hough: “Methods and means for recognizing complex patterns”, U.S. Patent 3,069,654 (1962).

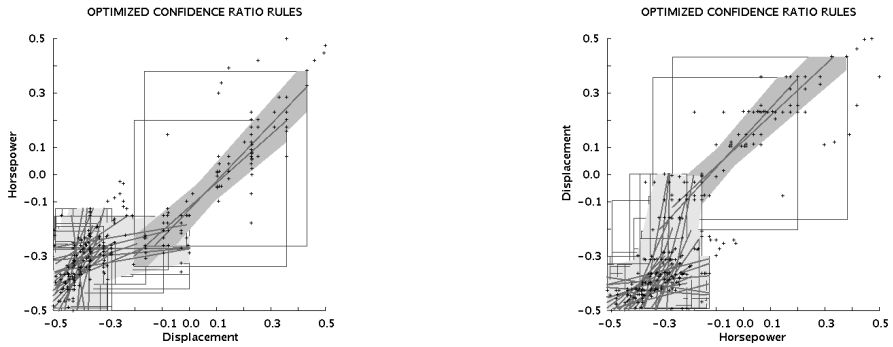

図 9　車の排気量と馬力に関するデータから最適確信度対称比率規則を抽出した結果.

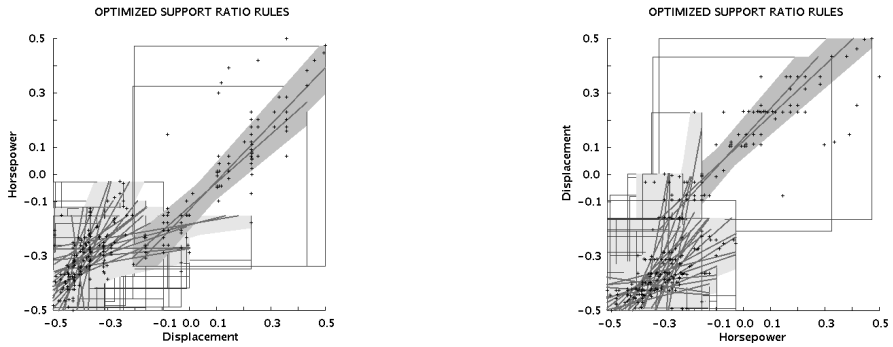

図 10　車の排気量と馬力に関するデータから最適サポート対称比率規則を抽出した結果.

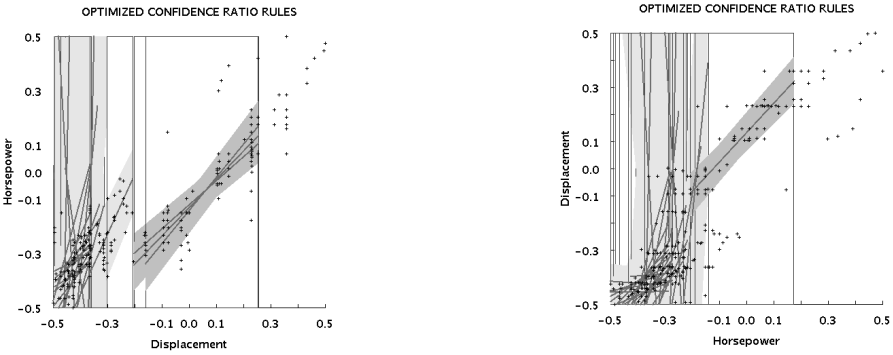

図 11　車の排気量と馬力に関するデータから最適確信度非対称比率規則を抽出した結果.

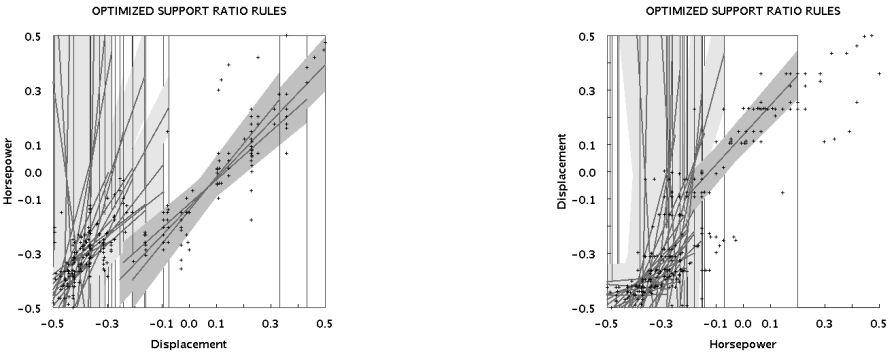

図 12　車の排気量と馬力に関するデータから最適サポート非対称比率規則を抽出した結果.

[11] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama: “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).