多次元比率規則の抽出手法

濱本雅史† 北川博之†,‡
†筑波大学大学院システム情報工学研究科 ‡筑波大学計算科学研究センター

1 はじめに

近年情報技術の発展により，計算機で扱うデータの量が爆発的に増大している．そのためユーザが求める情報をデータ中より発見することは困難になっている．そこで膨大なデータより必要な情報を発見するデータマイニングは，重要な技術課題となっている．データマイニングには様々な研究課題が存在するが，本研究では特に数値データに含まれる比率規則 [4] を抽出する問題を考える．比率規則とは数値属性間における増分の比率を表した規則である．本質的には，データ中の線形関係を抽出することが比率規則の抽出となる．

既存の比率規則抽出手法として主なものは，主成分分析 (PCA) を用いた比率規則抽出手法 [4] である．しかしこれは全体を近似する直線を抽出することが目的であるため，複数の線形関係が混在していたり，部分的に現れる線形関係の抽出が難しい．また属性値と線形関係がどのように対応しているかが直接的に得られない．

これに対しわれわれは，部分的に成り立つ線形関係も比率規則であると一般化し，主に 2 種類の数値属性間において比率規則を線分を用いて表現した (図 1)[5, 6]．また各比率規則の特徴を表すため，数値属性に対する相関ルールマイニング [3] における，サポートと確信度の概念を導入した．サポートは全タプルに対する割合であり，確信度はある部分領域において比率規則に従うタプルの割合である．サポートと確信度を用いると，ユーザは相関ルールマイニングと同様，最小サポートと最小確信度を適当に与えることで，それぞれの意図を反映した結果を得ることが出来る．

しかし現実には 3 種類以上の数値属性から成り立つ比率規則も存在する．本論文ではこれまでの 2 次元データに対する比率規則を拡張し，3 次元以上のデータに対応させる手法について検討する．

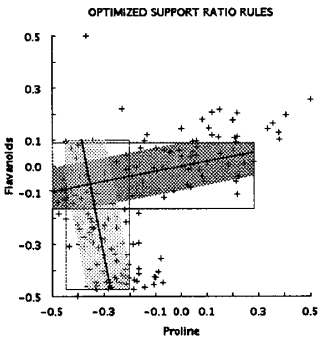

図 1: 2 次元データに対する比率規則の例.

2 比率規則

比率規則ははじめに述べたように，数値属性間の線形関係を表したものである．2 次元データに対してわれわれは，ある矩形領域 R 内に含まれるタプル t について，直線 $\rho = x\cos\theta + y\sin\theta$ との距離が許容幅 ϵ 以内であるならば t は比率規則に従う，という定義を行った [5, 6]．言い換えると，比率規則は 2 次元空間中の線分とその周辺領域が満たす性質として定義される．また各比率規則の定量的な特徴を表すため，サポートと確信度を定義した．このサポートあるいは確信度を，ユーザが与えた最小値を満たしたうえで最大とする比率規則の抽出手法を，これまで提案してきた．

2 次元データに対する比率規則を多次元に拡張する際，まず問題になるのが比率規則を表す線分の表現である．2 次元では直線をハフ変換 [1] により $\rho = x\cos\theta + y\sin\theta$ という式で表すことができたが，3 次元以上ではこの表現を直接用いることはできない．そこで多次元での直線表現手法を考える．まず直線から原点に下ろした垂線について，その垂線の長さ ρ と，垂線の単位方向ベクトル $\mathbf{a}(a_1, \cdots, a_m)$ を与える．2 次元では垂線を与えると一意に直線が定まったが，3 次元以上では直線を含む超平面のみ定まる．そのため，超平面上での直線の方向を表す単位方向ベクトル $\mathbf{b}$ も与える．このとき垂線の足と点 X の距離が，X を表す位置ベクトルと $\mathbf{b}$ の内積 $X \cdot \mathbf{b}$ で表される．

以上より，m 次元における直線を $X = \rho\mathbf{a} + (X \cdot \mathbf{b})\mathbf{b}$

Mining Multi Dimensional Ratio Rules
Masafumi Hamamoto†, Hiroyuki Kitagawa†, ‡
†Graduate School of Systems and Information Engineering, University of Tsukuba
‡Center for Computational Sciences, University of Tsukuba
hamamoto@kde.cs.tsukuba.ac.jp, kitagawa@cs.tsukuba.ac.jp

というように表す．この直線に対し2次元の比率規則と同様，タプルtと直線との距離が許容幅ϵ内である場合，tは多次元比率規則に従うとする．これを定式化したものが図2である．m次元の比率規則と分かるときは簡潔に$RR_{\mathbf{R}}(\rho, \mathbf{a}, \mathbf{b})$と書く．

> m 個の数値属性を持つタプル $\mathbf{t} = (x_1, x_2, \cdots, x_m)$ $(x_{ti} \in R_i, R_i \subseteq \mathcal{R})$ と直線 $X = \rho\mathbf{a} + (X \cdot \mathbf{b})\mathbf{b}$ の距離が ϵ 以下のとき，$\mathbf{t}$ は m 次元比率規則 $RR_{<R_1,\cdots,R_m>}(\rho \pm \epsilon, \mathbf{a}, \mathbf{b})$ に従うと呼ぶ．ここで $\mathbf{a} = (a_1, \cdots, a_m)$，$\mathbf{b} = (b_1, \cdots, b_m)$，$|\epsilon_{ti}| \leq \epsilon$，$\Sigma_i a_i^2 = \Sigma_i b_i^2 = 1$，$a \cdot b = 0$.

図 2: 多次元比率規則の定義．

個々の多次元比率規則について2次元の場合と同様，定量的な特徴を表すためサポートと確信度を定義する．比率規則のサポートは，ユーザより与えられた全タプルに対して比率規則に従うタプルの割合である．確信度はm次元空間$< R_1, \cdots, R_m >$内のタプル中，比率規則に従うタプルの割合である．すなわちサポートは比率規則に関わるタプル数を表し，確信度はその規則の確からしさを表す．そのため，サポートの値が大きな比率規則は全体的に成り立つ規則を表し，確信度の値が大きな比率規則は成り立つ可能性が高い規則を表す．特に同一のパラメータ$\rho, \mathbf{a}, \mathbf{b}$についてサポートあるいは確信度が最大となる比率規則を最適比率規則と呼び，それを与える空間を最適空間と呼ぶ．

3 提案手法

多次元比率規則は，2次元における対称比率規則[6]の拡張と言える．そのため対称比率規則の抽出手法を拡張することで，多次元比率規則の抽出を行うことができる．

対称比率規則の抽出手法を見直すと，パラメータ候補の枝刈りを行う枝刈りフェーズ，サポートあるいは確信度を最大とする領域(最適領域)を探索する比率規則生成フェーズ，類似した比率規則を統合する比率規則統合フェーズの3フェーズからなる．このうち最適領域を求めるために，Fukudaらの2次元数値属性相関ルールマイニング手法[2]を用いる．この手法は一方の属性の区間を列挙し，もう一方の属性については1次元数値属性相関ルールマイニング手法[3]を用いる．

そこで与えられたm次元データに対する，最適比率規則を抽出するための手法として，$m-1$個の属性についてはサポート以上のタプルを含むようなすべての領域を列挙し，残りの1属性については1次元数値属性相関ルールマイニングを用いる手法が考えられる．単純な列挙の場合最適空間を求める計算量は$O(N^{2m})$であるが，この手法を用いると，最大で$(N(N-1)/2)^{m-1}$個の候補を列挙することになるため，最適空間を求めるための計算量は$O(N^{2m-1})$となる．

さらに多次元の直線を表すためのパラメータである，単位方向ベクトル$\mathbf{a}, \mathbf{b}$の候補も，多次元では増加する．$\mathbf{a}, \mathbf{b}$それぞれがm次元ベクトルであるが，単位ベクトルであることと，$\mathbf{a}$と$\mathbf{b}$が直交することから，自由度は$2m-3$である．$\mathbf{a}, \mathbf{b}$の各次元の値をT個に離散化すると，候補数はT^{2m-3}個になる．また垂線の長さを表すパラメータρがR個に離散化されると，直線の候補は合計でRT^{2m-3}個となる．

以上から，計算量は$O(RT^{2m-3}N^{2m-1})$ということになる．高次元かつ多数のタプルに対して厳密な解を探索するには多くの計算が必要となるため，現実的にはサンプリングを行い近似解を求めることが考えられる．

4 おわりに

本論文ではこれまでの2次元データに対する比率規則を，多次元に拡張した多次元比率規則とその抽出手法を提案した．今後の研究として，この提案手法に対する実験と評価を行うほか，本論文で示したように次元数が増えると計算量が膨大になるため，より効率的な抽出手法の検討が考えられる．

謝辞

本研究の一部は，科学研究費補助金特定領域研究(#19024006)による．

参考文献

[1] 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).

[2] 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, Canada, pp. 13–23 (1996).

[3] 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).

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

[5] 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).

[6] 濱本雅史, 北川博之：対称比率規則の抽出手法, 日本データベース学会 Letters, Vol. 6, No. 1, pp. 73–76 (2007).