2001年度日本オペレーションズ・リサーチ学会
秋季研究発表会

データからの知識獲得における常識ルールと例外ルールについて

(申請中) 京都大学 *原口和也 HARAGUCHI Kazuya
01001374 京都大学 茨木俊秀 IBARAKI Toshihide

1 はじめに

近年、大量のデータが簡単にしかも安価に蓄えられるようになり、そこから意味のある知識を抽出するための科学的手法の重要性が増している.

本研究で扱う個々のデータ $\boldsymbol{a}=(a_1,a_2,\ldots,a_n)$ は n 個の属性に対して値を持つ n 次元ベクトルとして表現される. 各 $\boldsymbol{a}$ は正もしくは負いずれかの決定クラスに属し, 正データ集合を $\boldsymbol{P}$, 負データ集合を $\boldsymbol{N}$ とすると $\boldsymbol{P}\cup\boldsymbol{N}=\boldsymbol{S}$, $\boldsymbol{P}\cap\boldsymbol{N}=\emptyset$ が成り立つ. 各 $\boldsymbol{a}$ のクラスを $\Omega(\boldsymbol{a})$ と書き, 1 (正クラス) あるいは 0 (負クラス) の値をとる.

本研究では、ルール形式の知識をデータ集合 $\boldsymbol{S}$ から抽出することを試みる. ルール $R=(\pi,\omega)$ は条件部 π と結論部 ω から成り、「条件部が成り立つようなデータは決定クラス ω に属する」ことを表現し、支持度 μ と信頼度 ν によって確率的に評価される.

成立することが「常識的」で「当たり前」なルールを常識ルールと呼ぶ. これに対し、1つの常識ルールに何らかの条件が付加するとそれが覆るようなルールを例外ルールと呼ぶ[2]. 常識ルールと、それに対する例外ルールのルール対を発見するのが本研究の目的である. 以下ではこの目的にC4.5とFDAという2つのアルゴリズムを用い、アルゴリズムの適用の仕方によって発見されたルール対にどのような違いがあるかを考察する.

2 用語の定義

データ集合 $\boldsymbol{S}$ の持つ n 個の属性は、数値属性もしくは記号属性のいずれかに分類される. 数値属性は大小関係を持つ数値で表現され (例：身長、体温、検査値), 記号属性は大小関係を持たない記号で表現される (例：色、匂い, 味).

数値属性に対してはカット点 $x=T(i,c)$ $(a_i\geq c$ かどうかを判定), および区間 $x=I(i,c,c')$ $(c\leq a_i\leq c'$ かどうかを判定) と呼ばれる2種類の特徴が定義され、記号属性に対しては部分集合 $x=S(i,A)$ $(a_i\in A$ かどうかを判定) と呼ばれる特徴が定義される. 属性 i の特徴 x はそれぞれの性質に従いデータ $\boldsymbol{a}$ の i 番目の属性値 a_i に対して $\{0,1,*\}$ のいずれかを出力する. すなわち、$x[\boldsymbol{a}]\in\{0,1,*\}$ であって、$x[\boldsymbol{a}]=1(0)$ は特徴 x が成立する (成立しない) ことを示し、$x[\boldsymbol{a}]=*$ はどちらとも言えないことを意味する.

各特徴 x はブール変数とみなすことができるが, その否定を $\bar{x}$ と記す. ルール R の条件部 π は特徴 (あるいはその否定) の連言表現から成り, $\pi=w_1w_2\ldots w_k$ のように書ける. (ここで, 各 w_i はある特徴 x あるいはその否定 $\bar{x}$である.) $\pi[\boldsymbol{a}]=w_1[\boldsymbol{a}]w_2[\boldsymbol{a}]\ldots w_k[\boldsymbol{a}]$ は 0, 1 あるいは $*$の値をとる. $\pi[\boldsymbol{a}]=1$ のとき,「$\boldsymbol{a}$は Rの条件部を満たす」という. R の結論部 ω は0あるいは1の値をとる. $\boldsymbol{a}$ が R の条件部を満たし, かつ$\Omega(\boldsymbol{a})=\omega$ であるとき,「$\boldsymbol{a}$はルール Rを満たす」という.

ルールの性能を評価するため, 常識ルール $R^C=(\pi_C,\omega_C)$ に対し常識支持度 μ^C および常識信頼度 ν^C, 例外ルール $R^E=(\pi_E,\omega_E)$ に対し例外支持度 μ^E および例外信頼度 ν^E を定義する. ただし $\omega_E=1-\omega_C$ である. ルール R^C の条件部を満たすデータ集合を $\boldsymbol{S}_C$, $\boldsymbol{S}_C$ の中にあってルール R^E の条件部を満たすデータ集合を$\boldsymbol{S}_E$ とすると, 評価指数は次のように定式化される.

$$\mu^C = \frac{|\boldsymbol{S}_C|}{|\boldsymbol{S}|},\quad \nu^C = \frac{|\{\boldsymbol{a}\in\boldsymbol{S}_C\mid\Omega(\boldsymbol{a})=\omega_C\}|}{|\boldsymbol{S}_C|}$$

$$\mu^E = \frac{|\boldsymbol{S}_E|-|\{\boldsymbol{a}\in\boldsymbol{S}_E\mid\Omega(\boldsymbol{a})=\omega_C\}|}{|\boldsymbol{S}_C|-|\{\boldsymbol{a}\in\boldsymbol{S}_C\mid\Omega(\boldsymbol{a})=\omega_C\}|}$$

$$\nu^E = \frac{|\{\boldsymbol{a}\in\boldsymbol{S}_E\mid\Omega(\boldsymbol{a})=\omega_E\}|}{|\boldsymbol{S}_E|}$$

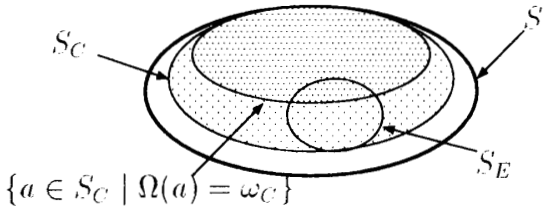

図1: データ集合 S における S_C, S_E の分布

3 ルール対発見戦略

データ集合から有用なルールを発見するにあたって本研究では (1) データ集合の決定木をC4.5 [3] によって構成し、生成された木の部分ルールの中からある程度高い支持度と信頼度を持つものを選ぶ, (2) 正負データを分離

するために必要な極小特徴集合である支持集合を近似解法 FDA [1] で求めておき，選ばれた特徴を組み合わせてルールを構成する，という2つのアプローチをとる．これらの適用の仕方によって，次の4つの戦略が可能である．

戦略1 R^CもR^Eも C4.5 を使って求める．

戦略2 R^Cは FDA，R^Eは C4.5 を使って求める．

戦略3 R^Cは C4.5，R^Eは FDA を使って求める．

戦略4 R^CもR^Eも FDA を使って求める．

4 実験結果

乳癌の診断結果に関するデータ集合 (総数 683，正例 239，負例 444) からルール対の発見を行なった．このデータ集合の属性数は 9 で，すべて数値属性である．

図 2 から図 4 に，発見されたルールを支持度 (横軸，support) と信頼度 (縦軸，confidence) にしたがってプロットしたグラフを示す．破線で結ばれたルールは，常識ルール (common rules) とそれから発見された例外ルール (exceptional rules) を表している．戦略 1 によって 6 のルール対が発見され，これをプロットしたグラフが図 2 である．

FDA ではデータに関する誤差を許容しながら支持集合の発見を行うことができ，この誤差 (単位は%) は入力として与えられる．今回は許容誤差 0% と 15% の場合について戦略 2 による実験を行った．この結果それぞれについて 28，308 のルール対が発見され，これをプロットしたグラフが図 3 および図 4 である．なお，支持度，信頼度ともに等しいルール同士は同じ点にプロットされている．

グラフの右上に位置する点に相当するルールが「優れたルール」と評価される．図 2 と図 3(あるいは図 4) を比較することにより，C4.5 と FDA はともに同じ程度に優れたルールを発見していることがわかる．ここで FDA で発見されるルールの数が多いのは，特徴の組み合わせを総当たりで調べることによる．一方，C4.5 から発見された常識ルールと比べて，FDA で許容誤差を与えずに発見された常識ルールからは優れた例外ルールを発見できるとは言えない．しかし，許容誤差を与えることによって劣っているものも数多く見つけるが，優れたものもいくつか見つけることができる．

5 おわりに

ここでは紙面の都合上，単一のデータ集合に戦略 1 および戦略 2 を適用した結果を載せただけにとどまった．しかし実際には，もっと多くの戦略を多くのデータ集合に適用してその傾向を考察している．

参考文献

[1] Mii, S.: *Feature determination algorithms in the analysis of data*, 京都大学修士論文, 2001.

[2] Suzuki, H.: 共通データからの仮説駆動型例外ルール発見, 人工知能学会誌, **15**(5), pp.782-789(2000).

[3] J.R.Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, 1993.

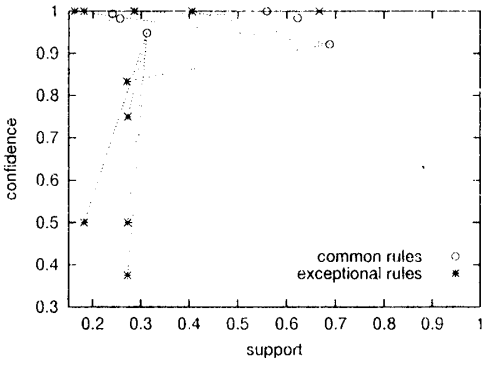

図 1: 戦略 1 によって得られたルールの評価指数 (縦軸が支持度，横軸が信頼度)

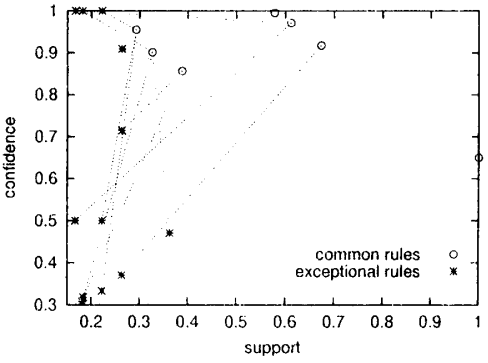

図 2: R^Cの発見時，許容誤差 0% を与えて戦略 2 によって得られたルールの評価指数

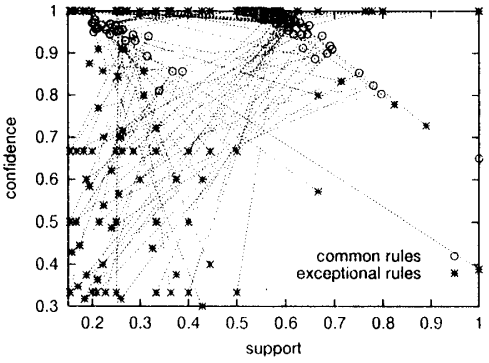

図 3: R^Cの発見時，許容誤差 15% を与えて戦略 2 によって得られたルールの評価指数