人　工　知　能　104－7
（1996．5．14）

情報理論的規準を用いたデータベースからの例外的知識の発見

鈴木　英之進 *1　　志村　正道 *2
suzuki@dnj.ynu.ac.jp
shimura@cs.titech.ac.jp
*1 横浜国立大学工学部電子情報工学科
*2 東京工業大学大学院情報理工学研究科計算工学専攻

概要

本論文では，データベースから有用な例外的知識を発見する手法について述べる．従来の手法は，あらかじめ与えた背景知識や領域固有の規準で例外的知識の潜在的有用性，すなわち興味深さを評価するので，有用な知識を発見し損なう可能性があることが指摘されていた．そこで，我々は興味深さを評価する情報理論的規準を定義することにより，背景知識にも領域固有の規準にも依存せずに例外的知識と関連する通常的知識を一体として発見する手法を提案する．この手法では，情報理論的規準の上限値に基づく分岐限定法を用いて発見アルゴリズムを効率化している．提案した手法に基づいてＭＥＰＲＯを構築し，実験によってその有効性を確認している．

Exceptional Knowledge Discovery in Databases based on an Information-Theoretic Criterion

Einoshin Suzuki *1　and　Masamichi Shimura *2
*1　Division of Electrical and Computer Eng.,　Yokohama National Univ.
*2　Dept. of Computer Science,　Tokyo Institute of Technology

Abstract

This paper presents an algorithm for discovering useful exceptional knowledge from databases. Previous approaches are prone to overlook useful knowledge since they employ either pre-supplied background knowledge or domain-specific criteria for evaluating the possible usefulness, i.e. the interestingness of exceptional knowledge. In order to circumvent these difficulties, we define an information-theoretic criterion which requires neither pre-supplied background knowledge nor domain-specific criteria, and propose an approach in which we obtain exceptional knowledge associated with general knowledge. Its search efficiency is improved by a branch-and-bound method based on the upper-bound for the criterion. Our MEPRO has been validated using several databases.

1 はじめに

今日，データベースの数とそこに保存されるデータの量は飛躍的に増加している．このような背景を受けて，データベースから自動的に知識を獲得する手法に関心が集まりつつある [Frawley 91, 河野 95, Matheus 93]．データベースから発見される知識は，「大型旅客機は安全な乗物である」などの多数の例について高い確率で成立する通常的知識と，「△△である大型旅客機は危険な乗物である」などの通常的知識に対する例外を表す例外的知識に分類することができる．

例外的知識は，予測できないような意外性の高い知識で，通常的知識以上に有用となることがある．このような例外的知識をデータベースから発見するシステムとしては，領域固有の階層関係や背景知識を用いる EXPLORA [Hoschka 91] や，発見した知識を評価する領域固有の規準を用いる KEFIR [Matheus 94, Piatetsky-Shapiro 94] などが提案されている．

データベース中には極めて多数の知識が存在するので，その中から有用である可能性が高い，すなわち興味深い知識を見分けることは一般に容易ではないが，これは知識発見システムに必要不可欠な機能である．特にデータベース中に埋没している例外的知識を発見する場合においては，興味深さを評価する規準を定義することが最も重要であるとされている [Piatetsky-Shapiro 94]．上述のようにこれまで提案された手法では，あらかじめ与えられた背景知識か，領域固有の規準に基づいて興味深さを評価していた．しかし，このような背景知識を用いると逆に興味深い知識を発見できない場合があることも指摘されており [Frawley 91]，また興味深さを評価する適切な領域固有の規準が存在しない場合もある．

このような問題に対処するために，例外的知識を発見する手法として，背景知識にも興味深さを評価する領域固有の規準にも依存しない情報理論的手法を提案する．最初に 2 節では，発見対象として，例外的知識を表す例外的ルールと通常的知識を表す通常的ルールから構成されるルールペアを導入する．次に 3 節では，例外的知識の興味深さを評価する規準として，平均圧縮情報量積を提案する．4 節では，発見アルゴリズムと，その中で用いられている分岐限定法を説明する．5 節では，機械学習における標準問題 [Murphy 94] を用いた実験について述べる．最後の 6 節は，結論である．

2 ルールペア

データベース中に蓄えられているレコードを例 e_i と呼ぶ時，データベース中には n 個の例 $e_1, e_2, \cdots, e_n$ が保存されているとする．この例 e_i は，m 個の属性についての属性値 $a_{i1}, a_{i2}, \cdots, a_{im}$ から構成されるタプル $< a_{i1}, a_{i2}, \cdots, a_{im} >$ で表されている．ただし，属性はすべて離散値の属性値をとると仮定する．もちろん，一般に属性は実数の属性値や構造化された属性値をとりうるが，そのような属性は実数値属性を扱う種々の方法 [Dougherty 95] や木構造属性を扱う方法 [Almuallim 95] などを前処理として用いることにより，離散値の属性値をとる属性に変換されているものとする．

本論文では，K 個の知識 $\{r_1, r_2, \cdots, r_K\}$ を興味深い順に求めることを考える．データベースから発見される知識 r_i は，次のようなルールペア $r(\mu, \nu)$ で表すことにする．

$$r(\mu,\nu) \equiv \begin{cases} Y_\mu & \rightarrow \quad x \\ Y_\mu \wedge Z_\nu & \rightarrow \quad x' \end{cases} \qquad (1)$$

ただし，$Y_\mu = y_1 \wedge y_2 \wedge \cdots \wedge y_\mu$, $Z_\nu = z_1 \wedge z_2 \wedge \cdots \wedge z_\nu$ であり，x, x', y_i, z_i はそれぞれ一つのアトムを表す．ここで，アトムとはある属性が単一の属性値をとる命題で表される事象を表している．なお，アトム x, x' 中の属性は等しく，その属性値は異なるものとする．

ルールの前提部と結論部は正の相関，あるいは因果関係を表しているので，前提部の事象が出現する時に結論部の事象が出現する条件つき確率は結論部の事象が出現する確率より大きくなる．したがって，ルールペアにおいては必ず，

$$\mathrm{p}(x|Y_\mu) > \mathrm{p}(x),\ \mathrm{p}(x'|Y_\mu \wedge Z_\nu) > \mathrm{p}(x'), \qquad (2)$$

が成立するものとする．

式 (1) のルールペアは，「Y_μ ならば x であり，Y_μ かつ Z_ν ならば x' である」と解釈できる．ルール

$Y_\mu \to x$ は比較的一般的な通常的知識を表すルールであり，このルールを 通常的ルールと呼ぶ．一方，事象 Y_μ は事象 Y_μ かつ Z_ν よりも頻繁に起こるため，ルール $Y_\mu \wedge Z_\nu \to x'$ は通常的知識に対する例外的知識を表すので，例外的ルールと呼ぶ．

3 $ACEP$: 平均圧縮情報量積

情報理論的に見ると，ルール $Y_\mu \to x$ は，本来記述長 $-\log \mathrm{p}(x)$ の符号で記述される $n\mathrm{p}(x, Y_\mu)$ 個の例をより短い記述長 $-\log \mathrm{p}(x|Y_\mu)$ の符号で表し，本来記述長 $-\log \mathrm{p}(\overline{x})$ の符号で記述される $n\mathrm{p}(\overline{x}, Y_\mu)$ 個の例を記述長 $-\log \mathrm{p}(\overline{x}|Y_\mu)$ の符号で表すと考えられる．ただし，特に断らない限り対数 log の底は 2 とする．このように記述長の減少すなわち圧縮情報量を用いることで，ルールを与えることによって圧縮された情報を定量的に測ることが可能となる．ルールを与えることによって圧縮された情報量をデータベース中にある例の数 n で割った値を，平均圧縮情報量と呼び $\mathrm{ACE}(x, Y_\mu)$ で表すと，

$$
\begin{aligned}
&\mathrm{ACE}(x, Y_\mu) \\
&\equiv [\{-n\mathrm{p}(x, Y_\mu)\log \mathrm{p}(x) - n\mathrm{p}(\overline{x}, Y_\mu)\log \mathrm{p}(\overline{x})\} \\
&\quad -\{-n\mathrm{p}(x, Y_\mu)\log \mathrm{p}(x|Y_\mu) - n\mathrm{p}(\overline{x}, Y_\mu) \\
&\quad \cdot \log \mathrm{p}(\overline{x}|Y_\mu)\}]/n \\
&= \mathrm{p}(x, Y_\mu)\log \frac{\mathrm{p}(x|Y_\mu)}{\mathrm{p}(x)} + \mathrm{p}(\overline{x}, Y_\mu)\log \frac{\mathrm{p}(\overline{x}|Y_\mu)}{\mathrm{p}(\overline{x})}
\end{aligned}
\tag{3}
$$

となる．平均圧縮情報量が大きいルールは，データベース中にある多くの情報を簡潔に表すという観点から有用であり，また興味深いと考えられる．したがって，発見されたルールの興味深さを平均圧縮情報量で評価することにする．実際，平均圧縮情報量は，ルールの結論部が低い確率で成立する意外性，ルールの前提部が真である時に結論部が高い確率で成立する安定性およびルールが多くの例について成立する一般性の三つを統一的に評価する規準ともなっている．また，Smyth [Smyth 91] は，データベースから $Y_\mu \to x$ 形式のルールを発見する際に，平均圧縮情報量が多くの望ましい性質を持つ規準であることを示している．

平均圧縮情報量が極めて小さい通常的ルール $Y_\mu \to x$ に関連する例外的ルール $Y_\mu \wedge Z_\nu \to x'$ は，その平均圧縮情報量が大きくてもあまり興味深いとは言えない．すなわち，例外的知識の興味深さは，例外的ルールの平均圧縮情報量だけではなく，関連する通常的ルールの平均圧縮情報量にも依存する．したがって，例外的知識の興味深さを，二つの平均圧縮情報量を考慮して定義するのが妥当である．例外的知識の興味深さは，それぞれの平均圧縮情報量が増加する時に増加し，減少する時に減少すべきである．そのような性質を持つ最も単純な式は，両者の和 $\mathrm{ACE}(x, Y_\mu) + \mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$ あるいは両者の積 $\mathrm{ACE}(x, Y_\mu) \cdot \mathrm{ACE}\ (x', Y_\mu \wedge Z_\nu)$ である．

以下，通常的ルールの結論部 x と例外的ルールの結論部 x' を固定して例外的ルールの平均圧縮情報量と通常的ルールの平均圧縮情報量を調べ，例外的知識の興味深さを適切に評価する規準を考える．まず，二つの平均圧縮情報量が共に最大となる場合を調べる．式 (2)，(3) より，次の式 (4)〜(6) を得る．

$$
\begin{aligned}
&\mathrm{ACE}(x, Y_\mu) \\
&= (a+b)\log\left(\frac{a+b}{a+b+c+d+e+f}\frac{1}{\mathrm{p}(x)}\right) \\
&\quad +(c+d+e+f) \\
&\quad \cdot \log\left(\frac{c+d+e+f}{a+b+c+d+e+f}\frac{1}{\mathrm{p}(\overline{x})}\right)
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
&\mathrm{ACE}(x', Y_\mu \wedge Z_\nu) \\
&= c\ \log\left(\frac{c}{a+c+e}\frac{1}{\mathrm{p}(x')}\right) \\
&\quad +(a+e)\log\left(\frac{a+e}{a+c+e}\frac{1}{\mathrm{p}(\overline{x'})}\right)
\end{aligned}
\tag{5}
$$

$$
\frac{a+b}{a+b+c+d+e+f} > \mathrm{p}(x), \qquad \frac{c}{a+c+e} > \mathrm{p}(x') \tag{6}
$$

ただし，a, b, c, d, e, f は，

$$
\begin{aligned}
&a = \mathrm{p}(x, Y_\mu, Z_\nu),\ b = \mathrm{p}(x, Y_\mu, \overline{Z_\nu}), \\
&c = \mathrm{p}(x', Y_\mu, Z_\nu),\ d = \mathrm{p}(x', Y_\mu, \overline{Z_\nu}), \\
&e = \mathrm{p}(\overline{x \vee x'}, Y_\mu, Z_\nu),\ f = \mathrm{p}(\overline{x \vee x'}, Y_\mu, \overline{Z_\nu})
\end{aligned}
\tag{7}
$$

を表す．これらについて次の式 (8) が成立することが分かる．

$$a \geq 0,\ b \geq 0,\ a+b \leq \mathrm{p}(x),\ c \geq 0,\ d \geq 0,\\ c+d \leq \mathrm{p}(x'),\ e \geq 0,\ f \geq 0,\\ e+f \leq \mathrm{p}(\overline{x \vee x'}) \quad (8)$$

平均圧縮情報量 $\mathrm{ACE}(x, Y_\mu)$, $\mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$ は，付録に示す補題より，$b = \mathrm{p}(x), a = d = e = f = 0$ の時にそれぞれ最大となることが分かる．したがって，$\mathrm{p}(x)$, $\mathrm{p}(x')$, c についての平均圧縮情報量 $\mathrm{ACE}(x, Y_\mu)$, $\mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$ の最大値をそれぞれ U, V とおくと，式 (4), (5) より，次の式 (9), (10) を得る．

$$U = \mathrm{p}(x) \log \frac{1}{\mathrm{p}(x)+c} + c \log \left(\frac{c}{\mathrm{p}(x)+c} \frac{1}{\mathrm{p}(\overline{x})} \right) \quad (9)$$

$$V = c \log \frac{1}{\mathrm{p}(x')} \quad (10)$$

また，式 (6), (8) より，

$$0 \leq c \leq \mathrm{p}(x'),\ c < \mathrm{p}(\overline{x}) \quad (11)$$

を得る．

平均圧縮情報量の大きいルールペアが発見対象であるので，以下，平均圧縮情報量 $\mathrm{ACE}(x, Y_\mu)$, $\mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$ が，それぞれ U, V の近傍にある場合について考える．簡単な計算により，$U+V$ が最大となる時，$V = 0$ あるいは $U \approx 0$ となることが分かる．すなわち，平均圧縮情報量の和は，その最大値が例外的ルールあるいは通常的ルールの平均圧縮情報量に支配されるので，例外的知識の興味深さを評価する規準としては不適切である．実際に，5 節で述べる投票データベースの実験において，平均圧縮情報量の和を評価規準として用いた場合，発見されたルールペアの中に $\mathrm{ACE}(x', Y_\mu \wedge Z_\nu) \approx 0$ である例外的ルールや，$\mathrm{ACE}(x, Y_\mu) \approx 0$ である通常的ルールを含むものがあり，有用でない知識が発見されている．これに対し，積 $U \cdot V$ について調べると，このような場合はないことが分かる．したがって，それぞれのルールについての平均圧縮情報量の積は，例外的知識の興味深さを評価する規準として適切である最も単純な式の一つであると考えられる．

以下，通常的ルールの平均圧縮情報量と例外的ルールの平均圧縮情報量の積を**平均圧縮情報量積**と呼び，例外的知識の興味深さについての評価関数とする．式 (1) で与えられるルールペアの平均圧縮情報量積を $\mathrm{ACEP}(x, Y_\mu, x', Z_\nu)$ で表すと，

$$\mathrm{ACEP}(x, Y_\mu, x', Z_\nu) \equiv \mathrm{ACE}(x, Y_\mu) \cdot \mathrm{ACE}(x', Y_\mu \wedge Z_\nu) \quad (12)$$

となる．

例外的知識の興味深さを評価する規準としては，平均圧縮情報量積の他にも，通常的ルールの平均圧縮情報量と例外的ルールの平均圧縮情報量を係数 $\beta\ (\geq 1)$ で重みづけした和 $\mathrm{ACE}(x, Y_\mu) + \beta \mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$ や，重みづけした積 $\mathrm{ACE}(x, Y_\mu) \cdot \mathrm{ACE}(x', Y_\mu \wedge Z_\nu)^\beta$ などが考えられる．前者 $\mathrm{ACE}(x, Y_\mu) + \beta \mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$ については，上述の解析と同様にして，例外的ルールの平均圧縮情報量あるいは通常的ルールの平均圧縮情報量がほとんど考慮されないので，不適切であることが分かる．一方，後者 $\mathrm{ACE}(x, Y_\mu) \cdot \mathrm{ACE}(x', Y_\mu \wedge Z_\nu)^\beta$ では，通常的ルールに対して例外的ルールを重視する割合 β を指定できるが，その適切な値を選択することは容易でないと考えられる．

4 発見アルゴリズム

平均圧縮情報量積が大きいルールペアを K 番目まで求める発見アルゴリズムについて述べる．アルゴリズムでは，データベースからの知識発見を，ルールペアを表すノードから構成される探索木についての探索問題として考える．探索法としては深さ優先探索法を用い，最大深さ D はユーザが指定することができるものとする．ただし，単純な深さ優先探索法では多数のノードを調べる必要があるので，後述する分岐限定法を用いて探索効率を改善している．

探索木において，各ノードは式 (1) で与えられるルールペア $r(\mu, \nu)$ を表している．ここで，$\mu = 0$, $\nu = 0$ はそれぞれ，前提部が y_i あるいは z_i を含まない場合である．$\mu = \nu = 0$ の場合を深さ 1 のノードとし，以降探索木において深さが 1 増す毎に，通常的ルールあるいは例外的ルール

の前提部にアトムを一つ加えることになる．深さ2のノードでは $\mu = 1$, $\nu = 0$, 深さ3のノードでは $\mu = \nu = 1$ であり，深さ l (≥ 4) では，$\mu + \nu = l - 1$ $(\mu, \nu \geq 1)$ である．

したがって，探索木において子孫ノードは，ルールペア $r(\mu', \nu')$ を表す．ただし，μ' と ν' は，$\mu' \geq \mu$, $\nu' \geq \nu$ となる整数である．ここで，付録に示す定理より，このルールペアについての平均圧縮情報量積には上限値が存在する．言い換えれば，子孫ノードが表すルールペアの平均圧縮情報量積は，この上限値よりも必ず小さい．探索中において，ルールペアを K 個以上発見した以降は K 番目に大きい平均圧縮情報量積を ACEP_K と表す時，現在探索中のノードについての上限値が ACEP_K よりも小さければ，子孫ノードを調べても平均圧縮情報量積が ACEP_K よりも大きいルールペアは得られない．すなわち，それ以下の子孫ノードを調べても無意味であり，ユーザに出力するルールペアは同じとなるので，枝を刈ることができる．このような子孫ノードを調べない分岐限定法を用いることにより，探索の効率化を実現している．もちろん，発見されるルールペアの集合は分岐限定法を用いない場合と同じになる．

5 データベースへの適用

前節で述べた手法に基づく知識発見システムMEPRO (database Miner based on the average compressed Entropy PROduct criterion) を構築し，種々の実験を行った．ここではMEPROを，機械学習における標準問題 [Murphy 94] である，投票データベースとマッシュルームデータベースに適用した結果について述べる．

投票データベースとは，1984年における合衆国国会議員435人が所属する政党を，16個の議題への投票結果とともに表したものである．各属性は2個ないし3個の属性値をとる．発見するルールペアの数 $K = 10$，探索の最大深さ $D = 8$ の場合におけるMEPROの出力結果を表1に示す．表中の xY, Y は，それぞれ前提部 (Y) と結論部 (x) をともに満たす例の数，前提部を満たす例の数を表す．ACE と $ACEP$ はそれぞれ3節で定義した平均圧縮情報量と平均圧縮情報量積であり，コンマは論理積，$COND$ は通常的ルールの前提部を表す．

表1より，得られたルールペアは興味深い例外性を示していることが分かる．例えば2位のルールペアによれば，議題 adoption に賛成を投じた議員は91%の高い確率で民主党員であり，その数は231人にのぼる．しかし，議題 adoption に賛成を投じていても，議題 physician と satellite に賛成を投じた17人の議員は，全員共和党員である．このルールペアから，議題 adoption に賛成を投じる珍しい共和党員が存在することが分かり，またその条件は例外的ルールの前提部で示されている．

2番目の標準問題であるマッシュルームデータベースとは，北アメリカに自生するキノコ8124種について，毒の有無を22個の属性で記述したデータベースである．各属性は2個から12個の属性値をとる．投票データベースへの適用例ではすべての属性が結論部の属性になりえたが，結論部の属性が限定されているルールペアを発見したい場合もある．ここでは結論部の属性として毒の有無を表すクラスだけを考え，$K = 10$, $D = 8$ の場合について実験を行った．その結果を表2に示す．

マッシュルームデータベースより得られたルールペアも興味深い例外性を示すことが表2より分かる．例えば3位のルールペアは，属性 g-size が b であるキノコの70%は毒がないが，属性 odor が f であると100 %毒があることを示している．

探索の最大深さ D は，前提部が多数のアトムから構成されるルールペアを調べられるように，十分大きくなければならない．しかし，ルールペアの数は，深さ優先探索においては深さに対して指数関数的に増加する．以下，このような非効率性を改善する上で，分岐限定法が有効であることを実験によって示す．

図1は，投票データベースにおいて，分岐限定法によって枝刈りされたノード数と最大深さが D である深さ優先探索によって調べられるノード数の比を示したものである．K が増加すると比は減少し，K が探索木における D 以内のノード数以上であれば比は0となる．図より，分岐限定法は深さが大きいほど有効であることが分かる．例え

表 1: 投票データベースにおける 10 位までのルールペア.

順位	ルールペア	p($x|Y$)	p(x)	xY	Y	ACE	$ACEP$
1	adoption=yes → physician=no	0.87	0.57	219	253	0.175	0.0115
	$COND$, party=rep → physician=yes	1.00	0.41	22	22	0.066	
2	adoption=yes → party=demo	0.91	0.61	231	253	0.195	0.0105
	$COND$, physician=yes, satellite=yes → party=rep	1.00	0.39	17	17	0.054	
3	satellite=yes → physician=no	0.82	0.57	197	239	0.118	0.0104
	$COND$, party=rep → physician=yes	0.95	0.41	37	39	0.088	
4	party=demo → salvador=no	0.75	0.48	200	267	0.135	0.0101
	$COND$,nicaraguan=no,crime=yes → salvador=yes	0.97	0.49	37	38	0.075	
5	crime=yes → party=rep	0.64	0.39	158	248	0.105	0.0101
	$COND$, physician=no → party=demo	0.97	0.61	74	76	0.095	
6	adoption=yes → party=demo	0.91	0.61	231	253	0.195	0.0099
	$COND$, physician=yes, synfuels=no, south-africa=yes → party=rep	1.00	0.39	16	16	0.050	
7	salvador=yes → party=rep	0.74	0.39	157	212	0.182	0.0098
	$COND$, physician=no → party=demo	0.98	0.61	41	42	0.054	
8	crime=yes → salvador=yes	0.78	0.49	194	248	0.151	0.0097
	$COND$,physician=no, satellite=yes, nicaraguan=yes → salvador=no	0.95	0.48	35	37	0.064	
9	nicaraguan=yes → party=demo	0.90	0.61	218	242	0.169	0.0096
	$COND$, physician=yes, synfuels=no → party=rep	1.00	0.39	18	18	0.057	
10	satellite=yes → party=demo	0.84	0.61	200	239	0.094	0.0095
	$COND$, physician=yes, salvador=yes → party=rep	1.00	0.39	32	32	0.100	

表 2: 結論部の属性を毒性の有無を表すクラスに限定した時の，マッシュルームデータベースにおける 10 位までのルールペア.

順位	ルールペア	p($x|Y$)	p(x)	xY	Y	ACE	$ACEP$
1	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	3256	4216	0.132	0.0141
	$COND$,odor=n, sc-bring=w → class=e	1.00	0.52	912	912	0.107	
2	bruises=f, veil-color=w, ring-number=o → class=p	0.77	0.48	3248	4208	0.132	0.0140
	$COND$,odor=n, sc-bring=w → class=e	1.00	0.52	912	912	0.107	
3	g-size=b → class=e	0.70	0.52	3920	5612	0.067	0.0138
	$COND$,odor=f → class=p	1.00	0.48	1584	1584	0.205	
3	g-size=b → class=e	0.70	0.52	3920	5612	0.067	0.0138
	$COND$,sp-color=h → class=p	1.00	0.48	1584	1584	0.205	
5	bruises=f, veil-color=w → class=p	0.72	0.48	3284	4548	0.096	0.0134
	$COND$,odor=n, sc-bring=w → class=e	1.00	0.52	1200	1200	0.140	
6	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	3256	4216	0.132	0.0134
	$COND$,stalk-root=e → class=e	1.00	0.52	864	864	0.101	
7	bruises=f, g-attachment=f → class=p	0.72	0.48	3274	4538	0.095	0.0133
	$COND$,odor=n, sc-bring=w → class=e	1.00	0.52	1200	1200	0.140	
8	bruises=f, veil-color=w, ring-number=o → class=p	0.77	0.48	3248	4208	0.132	0.0133
	$COND$,stalk-root=e → class=e	1.00	0.52	864	864	0.101	
9	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	3256	4216	0.132	0.0126
	$COND$,g-spacing=w, sc-aring=w, ring-type=e → class=e	1.00	0.52	816	816	0.095	
9	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	3256	4216	0.132	0.0126
	$COND$,g-spacing=w, veil-color=w, ring-type=e → class=e	1.00	0.52	816	816	0.095	
9	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	3256	4216	0.132	0.0126
	$COND$,odor=n, g-spacing=w, sc-aring=w → class=e	1.00	0.52	816	816	0.095	
9	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	3256	4216	0.132	0.0126
	$COND$,odor=n, g-spacing=w, veil-color=w → class=e	1.00	0.52	816	816	0.095	

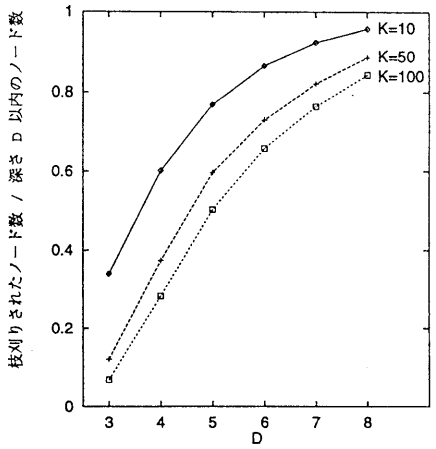

図 1: 探索の最大深さ D と発見するルールペアの数 K を変化した場合における分岐限定法の効果.

ば $D=8$ であれば，80 % 以上のノードが枝刈りされる．有用な例外的知識を発見するためには探索木を深く探索する必要があるので，この傾向は重要である．

6 おわりに

例外的知識は通常的知識に対する例外的なものとしてとらえられ，例外的知識の発見方法としては，通常的知識をあらかじめ与えておく方法が一般的である．例外的知識の興味深さを評価する方法として，あらかじめ与えた背景知識あるいは領域固有の規準で評価する方法は効率的であるが，一方でこの方法は，有用な知識を発見し損なう可能性がある．したがって，本論文ではあらかじめ通常的知識を準備することなく，例外的知識と通常的知識を一体として発見する枠組を考え，例外的ルールの平均圧縮情報量と対となる通常的ルールの平均圧縮情報量についての積で評価する方法を提案した．

また，発見アルゴリズムを効率化するために，平均圧縮情報量の積について解析し，その上限値を導き，上限値が小さいノードを探索しない分岐限定法を提案した．この分岐限定法は実装が容易であり，上限値も比較的簡単に計算できるという利点がある．

提案した手法に基づいて構築したシステム MEPRO を，機械学習の標準問題であるデータベースに適用した結果，有用で興味深い例外的知識を効率的に発見することに成功した．MEPRO は，このようにすぐれた性能を有しており，特にあらかじめ通常的知識が得にくいデータベースにおける知識発見に有効である．また，通常的知識を用いたことが原因で有用な例外的知識を発見できていない可能性があるデータベースに再適用することによっても，未知の有用な例外的知識を発見できると考えられる．

参考文献

[1] Almuallim, H., Akiba, Y. and Kaneda, S.: On Handling Tree-Structured Attributes in Decision Tree Learning, *Proc. of the 12th International Conference on Machine Learning*, Morgan Kaufmann, pp. 12–20 (1995).

[2] Dougherty, J., Kohavi, R. and Sahami, M.: Supervised and Unsupervised Discretization of Continuous Features, *Proc. of the 12th International Conference on Machine Learning*, Morgan Kaufmann, pp. 194–202 (1995).

[3] Frawley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J.: Knowledge Discovery in Databases: An Overview, *Knowledge Discovery in Databases*, Piatetsky-Shapiro, G. and Frawley, W. J. (eds), AAAI Press/ The MIT Press, pp. 1–27 (1991).

[4] Hoschka, P. and Klösgen, W.: A Support System For Interpreting Statistical Data, *Knowledge Discovery in Databases*, Piatetsky-Shapiro, G. and Frawley, W. J. (eds), AAAI Press/ The MIT Press, pp. 325–345 (1991).

[5] 河野浩之，西尾章治郎，Jiawei Han: データベースからの知識獲得技術，人工知能学会誌，Vol. 10, No. 1, pp. 38–44 (1995).

[6] Matheus, C. J., Chan, P. K. C. and Piatetsky-Shapiro, G.: Systems for Knowledge Discovery in Databases, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, No. 6, pp. 903–913 (1993).

[7] Matheus, C. J., Piatetsky-Shapiro, G. and McNeill, D.: An Application of KEFIR to the Analysis of Healthcare Information, *AAAI-94 Workshop on Knowledge Discovery in Databases*, Fayyad, U. M. and Uthurusamy, R. (eds), pp. 441–452 (1994).

[8] Murphy, P. M. and Aha, D. W.: UCI Repository of machine learning databases, *http://www.ics.uci.edu/~mlearn/MLRepository. html*, University of California, Department of Information and Computer Science (1994).

[9] Piatetsky-Shapiro, G. and Matheus, C. J.: The Interestingness of Deviations, *AAAI-94 Workshop on Knowledge Discovery in Databases*, Fayyad, U. M. and Uthurusamy, R. (eds), pp. 25–36 (1994).

[10] Smyth, P. and Goodman, R. M. : Rule Induction Using Information Theory, *Knowledge Discovery in Databases*, Piatetsky-Shapiro, G. and Frawley, W. J. (eds), AAAI Press/ The MIT Press, pp. 159–176 (1991).

付録

[補題]

$\sum_{i=1}^{n_1} a_i/\{\sum_{i=1}^{n_1} a_i + \sum_{i=1}^{n_2} b_i\} > \mathrm{p}(x)$ が成立する時,

$$G \equiv \sum_{i=1}^{n_1} a_i \log\left(\frac{\sum_{i=1}^{n_1} a_i}{\sum_{i=1}^{n_1} a_i + \sum_{i=1}^{n_2} b_i}\frac{1}{\mathrm{p}(x)}\right) + \sum_{i=1}^{n_2} b_i \log\left(\frac{\sum_{i=1}^{n_2} b_i}{\sum_{i=1}^{n_1} a_i + \sum_{i=1}^{n_2} b_i}\frac{1}{\mathrm{p}(\overline{x})}\right) \tag{13}$$

は，各 a_j に関して単調に増加し，各 b_j に関して単調に減少する．

証明) a_j, b_j について偏微分することにより，証明できる．□

[定理]

関数 $H(\alpha) \equiv \{\alpha/(1+\alpha)/\mathrm{p}(\overline{x})\}^{2\alpha}/(1+\alpha)/\mathrm{p}(x)$ を考える．また，α_1, α_2 が $H(\alpha_1) > 1 > H(\alpha_2)$ を満たすとする．この時，ルールペア $r(\mu', \nu')$ の平均圧縮情報量積 $\mathrm{ACEP}\,(x, Y_{\mu'}, x', Z_{\nu'}) = ACEP$ は，次の式 (14)，(15) を満たす．

$H\,(\mathrm{p}(x', Y_\mu, Z_\nu)/\mathrm{p}(x, Y_\mu)) \geq 1$ の時,

$$\begin{aligned} ACEP \leq & \left\{\mathrm{p}(x, Y_\mu)\log\left(\frac{\mathrm{p}(x, Y_\mu)}{\mathrm{p}(x, Y_\mu) + \mathrm{p}(x', Y_\mu, Z_\nu)} \cdot\frac{1}{\mathrm{p}(x)}\right) + \mathrm{p}(x', Y_\mu, Z_\nu) \cdot\log\left(\frac{\mathrm{p}(x', Y_\mu, Z_\nu)}{\mathrm{p}(x, Y_\mu) + \mathrm{p}(x', Y_\mu, Z_\nu)} \cdot\frac{1}{\mathrm{p}(\overline{x})}\right)\right\} \mathrm{p}(x', Y_\mu, Z_\nu)\log\frac{1}{\mathrm{p}(x')} \end{aligned} \tag{14}$$

その他の時,

$$ACEP < \alpha_2 \mathrm{p}(x, Y_\mu)^2 \left\{\log\left(\frac{1}{1+\alpha_1}\frac{1}{\mathrm{p}(x)}\right) + \alpha_1 \log\left(\frac{\alpha_1}{1+\alpha_1}\frac{1}{\mathrm{p}(\overline{x})}\right)\right\}\log\frac{1}{\mathrm{p}(x')} \tag{15}$$

証明) 平均圧縮情報量は非負 [Smyth 91] なので，$\mathrm{ACE}(x, Y_{\mu'})$ と $\mathrm{ACE}(x', Y_{\mu'} \wedge Z_{\nu'})$ が最大となる場合，平均圧縮情報量積も最大となる．補題より，これは，次の式 (16) が成立する場合であることが分かる．

$$\begin{aligned} \mathrm{p}(x, Y_{\mu'}, Z_\nu) &= \mathrm{p}(x, Y_\mu, Z_\nu), \\ \mathrm{p}(x, Y_{\mu'}, \overline{Z_\nu}) &= \mathrm{p}(x, Y_\mu, \overline{Z_\nu}), \\ \mathrm{p}(x', Y_{\mu'}, Z_{\nu'}) = q,\ \mathrm{p}(x', Y_{\mu'}, \overline{Z_\nu}) &= 0, \\ \mathrm{p}(\overline{x \vee x'}, Y_{\mu'}, Z_\nu) = \mathrm{p}(\overline{x \vee x'}, Y_{\mu'}, \overline{Z_\nu}) &= 0, \\ \mathrm{p}(x, Y_{\mu'}, Z_{\nu'}) = \mathrm{p}(\overline{x \vee x'}, Y_{\mu'}, Z_{\nu'}) &= 0. \end{aligned} \tag{16}$$

ただし，$q = \mathrm{p}(x', Y_{\mu'}, Z_\nu)$ である．$ACEP$ の q に関する最大値を，制約 $0 \leq q \leq \mathrm{p}(x', Y_\mu, Z_\nu)$, $q < \mathrm{p}(x, Y_\mu)\mathrm{p}(\overline{x})/\mathrm{p}(x)$ の元で求めることにより，証明できる．□