特集論文

一般性と正確性に基づくルール発見の最悪解析

Worst-Case Analysis of Rule Discovery Based on Generality and Accuracy

鈴木 英之進
Einoshin Suzuki横浜国立大学大学院工学研究院
Faculty of Engineering, Graduate School, Yokohama National University.
suzuki@ynu.ac.jp, <http://www.slab.dnj.ynu.ac.jp/>**keywords:** rule discovery, worst-case analysis, PAGA discovery, Chernoff bounds.

Summary

In this paper, we perform a worst-case analysis of rule discovery based on generality and accuracy. A rule is defined as a probabilistic constraint of true assignment to the class attribute for corresponding examples. In data mining, a rule can be considered as representing an important class of discovered patterns. We accomplish the aforementioned objective by extending a preliminary version of PAC learning, which represents a worst-case analysis for classification. Our analysis consists of two cases: the case in which we try to avoid finding a bad rule, and the case in which we try to avoid overlooking a good rule. Discussions on related work are also provided for PAC learning, multiple comparison, analysis of association rule discovery, and simultaneous reliability evaluation of a discovered rule.

1. ま え が き

データマイニング [Fayyad 96] は大量データからの有用知識発見と定義され、情報化社会の進展に伴い多大な注目を集めるようになった。データマイニングは高度なデータ解析と見なせるが、典型的な解析過程は複数の段階から構成される。Fayyad らのデータマイニングプロセスモデル [Fayyad 96] や Wirth の CRISP-DM モデル [CRI 01] は、この見方を形式化したものである。パターン抽出はデータマイニングプロセスの重要な一段階と定義され、前処理されたデータからパターンと呼ばれる規則性を発見する過程を表す。ルールはデータ集合中に成立する確率的制約と定義され、データマイニングにおいて最も重要なパターンの一種類として広く認められている。

データマイニングにおいてルール発見は盛んに行われているが、その理論的解析はほとんど行われて来なかった。数少ない例外としては、Agrawal らによる相関ルール発見の解析 [Agrawal 96] や、鈴木による発見されたルールの信頼性同時評価 [Suzuki 98] などが挙げられるが、これらはルールの総数を考慮していない。このことは、これらの解析が発見問題の規模を必要例数に関連づけていないことを表し、データマイニングのより強固な基盤が望まれていることを示している。

この問題に対する第一歩として、われわれは分類学習に関する最悪解析である PAC 学習の初歩 [Russel 95] を拡張する [鈴木 01a, Suzuki 01b]。解析においては、悪

いルールの発見を回避する場合と良いルールの見落としを回避する場合を分けて考える。本論文では、PAC 学習 [Kearns 94, Russel 95]、多重比較 [Jensen 00]、相関ルール発見の解析 [Agrawal 96]、および発見されたルールの信頼性同時評価 [Suzuki 98] などの関連研究も議論する。

2. ルール発見

2.1 ルール

データ集合は m 個の例から構成され、各例は b 個の離散属性と 1 個のクラス属性によって記述されるとする。一般にルール発見ではクラス属性は限定されていないが、ここでは形式化のため、あるクラス属性が真となる場合を予測するルールを考える。

属性 A への属性値 V の割り当て $A = V$ をアトムと呼ぶ。本論文では与えられたデータ集合を、真のデータを復元抽出した結果と見なす。すなわち、真のデータをデータ集合 S_0 とし、データマイニング問題のデータをそこからの一様分布によるサンプル $S = \{s_1, s_2, \dots, s_m\}$ とする。ただし s_i は例を表す。 S_0 において、アトムを構成要素とする命題論理式 f が成立する例の割合を f の真の確率と呼び、 $\Pr(f)$ で表す。 S からの $\Pr(f)$ の推定値を f の推定確率と呼び、 $\widehat{\Pr}(f)$ で表す。 $\widehat{\Pr}(f)$ はラプラス推定や、 S において f が成立する例の数を全例数 m で割った値などで求められるが、本論文では後者の方法を用いる。

ルール r は、アトムを構成要素とする命題論理式である前提部 y と、クラス属性が真であることを表す結論部 x を用いて次のように表される。

$$r: y \rightarrow x$$

r の直観的な意味は、 y が成立する例が比較的多く存在し、それらの例について比較的高い割合で x が成立することである。後でルールの評価規準を定義するため、 $\Pr(y)$ と $\Pr(x|y)$ を r のそれぞれ一般性と正確性と呼ぶ。同様に、 $\widehat{\Pr}(y)$ と $\widehat{\Pr}(x|y)$ を r のそれぞれ推定された一般性と推定された正確性と呼ぶ。

2.2 関連するルールの種類

確率的 if-then ルール [Smyth 92] は、次のように定義される。

$$y_1 \wedge y_2 \wedge \cdots \wedge y_K \rightarrow x$$

ただし y_i は単一アトムである。論文 [Suzuki 98] では、確率的 if-then ルールを連言ルールと呼んでおり、本論文でもこれに従う。連言ルールは、前節のルールにおいて前提部が単一のアトムかアトムの連言に限定された場合であると見なすことができる。

連言ルールは前提部が (属性 = 値) の組み合わせで表されるため、その種類は通常きわめて多い。前節のルールとして表される連言ルールの種類数 $|R|$ は、データ集合に属性が b 個あり各属性がそれぞれ a 種類の値をとっているとすると、次の式で表される。

$$|R| = (a+1)^b - 1 \quad (1)$$

この式は、前提部に来る可能性がある b 通りの属性は属性が現れない場合を入れて $a+1$ 通りの状態をとることから説明できる。 $|R|$ の値は通常大きく、例えば 20 個の 2 値属性から構成されるデータ集合においては $|R| = 3,486,784,400$ となる。現実的な対策としては、前提部のアトムを高々 K 個に限定することが考えられる。この場合、種類数 $|R_K|$ は次の式で与えられる。

$$|R_K| = \sum_{i=1}^K a^i \binom{b}{i} \quad (2)$$

(1) は (2) に $K = b$ を代入して二項係数を考えることによって導出できる。

相関ルール発見 [Agrawal 96] においては、データ集合は 2 値属性から構成されるトランザクションデータに限定される。2 値属性が真となる条件はアイテムと呼ばれる。アイテム集合を単一のアイテムかアイテムの連言であると定義する。本来の相関ルールは、前提部と結論部が共にアイテム集合で表される。前節の枠組では、相関ルールは、連言ルールにおいて属性値が全て「真」に限定された場合と見なすことができる。前節のルールとして表される相関ルールの種類数と前提部のアイテムが高々 K 個である相関ルールの種類数は、それぞれ (1) と (2) に $a = 1$ を代入した場合として与えられる。

2.3 ルール発見問題

われわれは本論文におけるルール発見アルゴリズムが満たすべき条件として、一般性と正確性がそれぞれ $1-\zeta$ と $1-\epsilon$ 以上であるルールを、与えられたデータ集合から高い確率 $1-\delta$ で発見することを考える。すなわち、真のデータ S_0 からのサンプルを複数与えられた際、上記のルールを $1-\delta$ 以上の確率で発見する。発見されるルールは通常複数個あるが、本論文では解析のため単一とする。

アルゴリズムが満たす条件：発見ルール

$y \rightarrow x$ について

$$\Pr[\Pr(y) \geq 1-\zeta, \Pr(x|y) \geq 1-\epsilon] \geq 1-\delta \quad (3)$$

ただし $0 < \zeta, \epsilon, \delta < 1$

対象とする発見アルゴリズムは、推定された一般性と推定された正確性が、それぞれユーザが与えた閾値 θ_S と θ_F 以上であるルールを発見する。2.1 節で述べたように、与えられたデータは真のデータを復元抽出した結果であると考えられるため、ユーザはアルゴリズムの適用に当たって $\theta_S \neq 1-\zeta, \theta_F \neq 1-\epsilon$ となる閾値を用いる。

アルゴリズムの動作：次を満たすルール

$y \rightarrow x$ を発見

$$\widehat{\Pr}(y) \geq \theta_S, \widehat{\Pr}(x|y) \geq \theta_F \quad (4)$$

なおルール発見において、ユーザは必ずしも一般性が高いルールを求める訳ではなく、一般性の低いルールの方が興味深い場合が多いことが経験的に知られている。ただし、一般性が低いルールの個数は属性数に関してほぼ指数関数的に増えることも分かっており、通常のルール発見アルゴリズムでは何らかの制限を設け、一般性がある程度高いルールを発見する。われわれも、一般性は低いに興味深いルールを発見することが重要であるとの認識を持ち、例外ルール発見 [Suzuki 99, 鈴木 00, 鈴木 02] などの対策を提案して来た。ただし、ここでは解析のために、ユーザが一般性に関する最小閾値を (反復的に) 指定する場合を対象とする。

ここで興味深い問題として、目的 (3) をアルゴリズム (4) で達成するために必要な例数 m を、関連するパラメータに関して記述することが考えられる。この問題は有名な PAC (Probably Approximately Correct) 学習 [Kearns 94, Russel 95] にちなんで PAGA (Probably Approximately General and Accurate) 発見と名づけられるものであり、データマイニング、発見科学、およびアクティブマイニングの基礎研究として位置づけられる。

3. 悪いルールを発見したくない場合

2.3 節の問題について、悪いルールを発見したくない場合を仮定し、必要な例数を求める。この仮定は信頼性

を重視する応用分野において重要であると考えられる。

3.1 準備

最初に解析に必要な事項を定義しておく。確率変数 X の領域が $\{0, 1, \dots, \nu\}$ であり、その確率分布が次式を満たす場合、 X は二項分布に従うという [Feller 57]。これは、独立な m 回のコイン投げ (またはベルヌーイ試行、ベルヌーイの試み [Feller 57]) で、表が出る回数の総和が k 回になる確率に相当する。

$$\begin{aligned} \Pr(X = k) &= B(k; \nu, p) \\ &= \binom{\nu}{k} p^k (1-p)^{\nu-k} \end{aligned} \quad (5)$$

ただし p は $0 < p < 1$ を満たす定数、 $k = 0, 1, \dots, \nu$ である。Chernoff bounds によれば、任意の定数 $q > p$ について次の式が成立する [Agrawal 96]*1。

$$\Pr(X > \nu q) < \exp[-2\nu(q-p)^2] \quad (6)$$

3.2 理論的解析

(3) より悪いルール $r_b: y \rightarrow x$ は、次を満たす。

$$\Pr(y) < 1 - \zeta \text{ または } \Pr(x|y) < 1 - \epsilon \quad (7)$$

本節では悪いルールの発見を避ける場合を仮定するため、一般性に関する閾値 θ_S と正確性に関する閾値 θ_F は比較的大きい。この仮定、(3)、および (4) より、

$$\theta_S > 1 - \zeta \text{ かつ } \theta_F > 1 - \epsilon. \quad (8)$$

(7) と (8) より、

$$\theta_S > \Pr(y) \text{ または } \theta_F > \Pr(x|y). \quad (9)$$

$r_b: y \rightarrow x$ が発見されるためには、

$$\widehat{\Pr}(y) \geq \theta_S \text{ かつ } \widehat{\Pr}(x|y) \geq \theta_F. \quad (10)$$

与えられたデータ集合の例数を m とする。 y と xy がデータ集合においてそれぞれ $\lceil m\theta_S \rceil$ と $\lceil m\widehat{\Pr}(y)\theta_F \rceil$ 個以上の例で成立すれば、 r_b は発見されてしまう。 y と xy が成立する例数が二項分布に従うことより、

$$\begin{aligned} &\Pr(r_b \text{ が発見される}) \\ &\leq \text{MAX} \left[\sum_{k=\lceil m\theta_S \rceil}^m B(k; m, \Pr(y)), \right. \\ &\quad \left. \sum_{k=\lceil m\widehat{\Pr}(y)\theta_F \rceil}^{m\widehat{\Pr}(y)} B(k; m\widehat{\Pr}(y), \Pr(x|y)) \right] \\ &< \text{MAX} \left\{ \exp \left[-2m \left(\frac{\lceil m\theta_S \rceil}{m} - \Pr(y) \right)^2 \right], \right. \end{aligned} \quad (11)$$

*1 参考文献 [Agrawal 96] の p. 325, 最後の式において、 $h \rightarrow \nu$, $s \rightarrow p$, $s + \alpha \rightarrow q$ と置き換えることにより導出される。

$$\exp \left[-2m\widehat{\Pr}(y) \left(\frac{\lceil m\widehat{\Pr}(y)\theta_F \rceil}{m\widehat{\Pr}(y)} - \Pr(x|y) \right)^2 \right] \} \quad (12)$$

$$< \text{MAX} \left\{ \exp \left[-2m(\theta_S - 1 + \zeta)^2 \right], \exp \left[-2m\theta_S(\theta_F - 1 + \epsilon)^2 \right] \right\}. \quad (13)$$

(11) においては一般性が低いルール r_{b1} が発見される場合と正確性が低いルール r_{b2} が発見される場合を分けて考えている。第一、二項は (7) のそれぞれ前半、後半に該当する。 $\Pr(r_{b1})$ と $\Pr(r_{b2})$ が未知のため、 $\Pr(r_b \text{ が発見される})$ の上限を $\text{MAX}[\Pr(r_{b1} \text{ が発見される}), \Pr(r_{b2} \text{ が発見される})]$ とした。(12) では、(9) より (6) の Chernoff bounds を適用している。最後に (13) においては、(7) と (10) の前半を用いている。

全てのルールの集合と全ての悪いルールの集合をそれぞれ R , R_b とおき、集合 S の要素数を $|S|$ で表す。悪いルールが発見される確率は次式を満たす。

$\Pr(R_b \text{ が発見ルールを含む})$

$$< |R_b| \text{MAX} \left\{ \exp \left[-2m(\theta_S - 1 + \zeta)^2 \right], \exp \left[-2m\theta_S(\theta_F - 1 + \epsilon)^2 \right] \right\} \quad (14)$$

$$\leq |R| \text{MAX} \left\{ \exp \left[-2m(\theta_S - 1 + \zeta)^2 \right], \exp \left[-2m\theta_S(\theta_F - 1 + \epsilon)^2 \right] \right\} \quad (15)$$

(14) では複数個の悪いルールが発見条件を満たしてしまう場合を重複して数えることを許し、(15) では $|R| \geq |R_b|$ であることを用いた。悪いルールを発見しないためには、次式が十分小さい δ に関して成立する必要がある。

$$|R| \text{MAX} \left\{ \exp \left[-2m(\theta_S - 1 + \zeta)^2 \right], \exp \left[-2m\theta_S(\theta_F - 1 + \epsilon)^2 \right] \right\} \leq \delta \quad (16)$$

この式を m について解くと、悪いルールを発見しないために必要な例数が求まる。

$$m \geq \frac{\ln \left(\frac{|R|}{\delta} \right)}{2 \text{MIN} \left[(\theta_S - 1 + \zeta)^2, \theta_S(\theta_F - 1 + \epsilon)^2 \right]} \quad (17)$$

上式より、この場合のルール発見における各パラメータの必要例数への影響が定量的に分かる。2.2 節で述べたように $|R|$ は通常大きいため、対数関数的に効くのだが重要であることが分かる。次に重要な項は $\theta_S - 1 + \zeta$ と $\theta_F - 1 + \epsilon$ である。これらは 2 乗の逆数で影響するので小さい場合には問題となることが分かる。これらはそれぞれ、アルゴリズムで用いる閾値とユーザが要求する指標値の差を表し、それぞれ一般性と正確性に関するマージンと名付けられるものである。例えば典型的なルール発見では $\theta_S = 0.1$, $(\theta_S - 1 + \zeta) = 10^{-1}$ あるいは 10^{-2} と仮定できる。同様に $(\theta_F - 1 + \epsilon) = 10^{-1}$ あるいは 10^{-2} と仮定できる。この場合、分母は 2×10^{-3} あるいは 2×10^{-5}

となる．最後に，典型的な場合は $\delta = 0.01 - 0.05$ であるため， $|R|$ の分母としてだけ現れる δ の影響は通常小さい．

3.3 連言ルール発見への適用

(1) と (17) より，本研究で対象とする連言ルールを発見する場合に必要な例数は，次の式で与えられる．

$$m \geq \frac{\ln[(a+1)^b - 1] + \ln(\frac{1}{\delta})}{2\text{MIN}[(\theta_S - 1 + \zeta)^2, \theta_S(\theta_F - 1 + \epsilon)^2]} \quad (18)$$

相関ルールの場合は，上式に $a = 1$ を代入すれば良い．

まず $\delta = 0.01 - 0.05$ の場合 $\ln[(a+1)^b - 1] \gg \ln(1/\delta)$ なので $\ln(1/\delta)$ は通常無視できる．したがって，必要例数は b にほぼ比例する．次に属性値の種類数 a は右辺に対して対数関数的にしか効かないため，通常は b や一般性と正確性に関するマージンに比較して影響は小さい． $a = 2$ ， $\delta = 0.05$ と固定し， $b = 10^2, 10^3, 10^4$ の場合について，必要例数を $\text{MIN}[(\theta_S - 1 + \zeta)^2, \theta_S(\theta_F - 1 + \epsilon)^2]$ に関してプロットしたグラフを図 1 に示す．ただし縦軸も横軸も対数表示である．

図 1 悪い連言ルールを発見しないために必要な例数．図において MIN は $\text{MIN}[(\theta_S - 1 + \zeta)^2, \theta_S(\theta_F - 1 + \epsilon)^2]$ を表す．

図 1 を用いて典型的な連言ルール発見における必要例数を検討する．3.2 節の例では $\text{MIN}[(\theta_S - 1 + \zeta)^2, \theta_S(\theta_F - 1 + \epsilon)^2] = 10^{-3}$ あるいは 10^{-5} である．この時に必要例数は $b = 10^2 - 10^4$ に対して，約 $5.6 \times 10^4 - 5.6 \times 10^6$ あるいは約 $5.6 \times 10^6 - 5.6 \times 10^8$ である．したがって，マージンを小さくすると必要例数が問題となるほど増大してしまうことが分かる．ただしマージンが大きいことは閾値が大きいことを表し，ルールは通常発見されない*2．現実的で有効な対策としては，発見過程に応じて閾値を自動調節する手法 [Suzuki 99, 鈴木 00] が考えられる．もちろん，本論文は最悪解析を行っており，実際のルール発見における必要例数は上記の結果よりも少ない．

(2) と (17) より，前提部が高々 K 個のアトムから構成される連言ルールを発見する場合に必要な例数は，次の式で与えられる．

$$m \geq \frac{\ln \left[\sum_{i=1}^K a^i \binom{b}{i} \right] + \ln(\frac{1}{\delta})}{2\text{MIN}[(\theta_S - 1 + \zeta)^2, \theta_S(\theta_F - 1 + \epsilon)^2]} \quad (19)$$

相関ルールの場合は，上式に $a = 1$ を代入すれば良い．

この式において $a = 2$ ， $\delta = 0.05$ と固定し， $K = 2$ で $b = 10^2, 10^3, 10^4$ の場合と $b = 10^2$ で $K = 1, 2, 3, 4, 100 (= b)$ の場合を図 1 と同様にプロットしたグラフを，それぞれ図 2，図 3 に載せる．

図 2 悪い連言ルールを発見しないために必要な例数．ただし前提部のアトム数は高々 $K = 2$ ．

図 3 悪い連言ルールを発見しないために必要な例数．ただし属性数は $b = 100$ ．

図 2 より， $K = 2$ の場合 b の影響は比較的小さいことが分かる．一方図 3 は， $K \leq 4$ なら全ての連言ルールを求める場合 ($K = b = 100$) に比較して必要例数が約 $1/10$ となることが分かる．前提部が短いルールは可読性に優れるが，必要例数の観点から見ても発見対象として魅力的であることが，図より分かる．

*2 例えば，全てのルールについて推定された一般性が 0.2 以下である際に $\theta_S = 0.3$ と指定すると，ルールは発見されない．

4. 良いルールを見落としたくない場合

2・3 節の問題について、良いルールを見落としたくない場合を仮定し、必要な例数を求める。この仮定は可能性を重視する応用分野において重要であると考えられる。

(3) より良いルール $r_g: y \rightarrow x$ は、次を満たす。

$$\Pr(y) \geq 1 - \zeta \text{ かつ } \Pr(x|y) \geq 1 - \epsilon \quad (20)$$

本節では良いルールを見落としを避ける場合を仮定するため、一般性に関する閾値と正確性に関する閾値は比較的小さい。この仮定、(3)、および (4) より次が成立する。

$$\theta_S < 1 - \zeta \text{ かつ } \theta_F < 1 - \epsilon \quad (21)$$

(20) と (21) より、次の式が成立する。

$$\theta_S < \Pr(y) \text{ かつ } \theta_F < \Pr(x|y) \quad (22)$$

与えられたデータ集合の例数を m とする。データ集合において、 y が高々 $\lceil m\theta_S \rceil - 1$ 個の例で成立するか、 xy が高々 $\lceil m\widehat{\Pr}(y)\theta_F \rceil - 1$ 個の例で成立する条件のいずれかが成り立てば、また成り立つ時に限り、 r_g は発見されない。 y と xy が成立する例数が二項分布に従うことより、

$$\begin{aligned} & \Pr(r_g \text{ が発見されない}) \\ & \leq \text{MAX} \left[\sum_{k=0}^{\lceil m\theta_S \rceil - 1} B(k; m, \Pr(y)), \right. \\ & \quad \left. \sum_{k=0}^{\lceil m\widehat{\Pr}(y)\theta_F \rceil - 1} B(k; m\widehat{\Pr}(y), \Pr(x|y)) \right] \quad (23) \\ & = \text{MAX} \left[\sum_{k=\phi(m, \theta_S)}^m B(k; m, 1 - \Pr(y)), \right. \\ & \quad \left. \sum_{k=\phi(m\widehat{\Pr}(y), \theta_F)}^{m\widehat{\Pr}(y)} B(k; m\widehat{\Pr}(y), 1 - \Pr(x|y)) \right] \quad (24) \\ & < \text{MAX} \left\{ \exp \left[-2m \left(\frac{\phi(m, \theta_S)}{m} - 1 + \Pr(y) \right)^2 \right], \exp \left[-2m\widehat{\Pr}(y) \right. \right. \\ & \quad \left. \left. \left(\frac{\phi(m\widehat{\Pr}(y), \theta_F)}{m\widehat{\Pr}(y)} - 1 + \Pr(x|y) \right)^2 \right] \right\} \quad (25) \\ & < \text{MAX} \left\{ \exp \left[-2m(-\theta_S + 1 - \zeta)^2 \right], \right. \\ & \quad \left. \exp \left[-2m\theta_S(-\theta_F + 1 - \epsilon)^2 \right] \right\}. \quad (26) \end{aligned}$$

ただし $\phi(m, \theta) \equiv m - \lceil m\theta \rceil + 1$ 。

(23) においては良いルールについて推定された一般性が閾値よりも低い場合と推定された正確性が閾値よりも低い場合を分けて考えている。両者の確率が未知のため、

3・2 節と同様の方法を用いている。(24) は (5) において p を $1 - p$ に置き換えた場合に相当する。(25) では (22) より (6) の Chernoff bounds を適用している。最後に (26) では、(20) と $\widehat{\Pr}(y) \geq \theta_S$ を用いている。該当する項においては、 r_g が推定された正確性が低いために発見されないと仮定しており、後者の不等式が成立する。

3・2 節と同様にして、良いルールを高い確率で見落とさないために必要な例数は、次の式で与えられる。

$$m \geq \frac{\ln \left(\frac{|H|}{\delta} \right)}{2 \text{MIN} \left[(-\theta_S + 1 - \zeta)^2, \theta_S(-\theta_F + 1 - \epsilon)^2 \right]} \quad (27)$$

(27) は (17) に等しく、3・2 節、3・3 節とほぼ同様の議論が成立する。マージン (この場合は $1 - \zeta - \theta_S$ と $1 - \epsilon - \theta_F$) が大きいことは、閾値が小さいことに相当する。閾値が小さい場合は、通常は発見ルールの候補を多数調べなければならない。閾値の自動調節 [Suzuki 99, 鈴木 00] はこの問題に対しても現実的な対策法となりうる。

5. 関連研究に関する議論

5・1 PAC 学 習

PAC 学習 [Valiant 84] は、分類学習における必要例数の最悪解析であり、数多くの優れた成果が出ている。3 章の結果は、以下で述べるように PAC 学習の初歩 [Russel 95] の拡張と見なせる。まず、分類子は全ての例を対象とするので一般性を考えない。このため枠組 [Russel 95] では $\Pr(y)$, $\widehat{\Pr}(y)$, θ_S が考慮されていない。実際、枠組 [Russel 95] において学習アルゴリズムは、(3) を簡略化した条件を満たすと仮定される。

アルゴリズムが満たす条件：分類子 h について

$$\Pr[\Pr(h \text{ は } x \text{ を正しく予測}) \geq 1 - \epsilon] \geq 1 - \delta \quad (28)$$

ただし $\epsilon, \delta > 0$

次に、枠組 [Russel 95] では分類アルゴリズムは訓練例に無矛盾な分類子を返すと仮定している。これは、われわれの研究において $\theta_F = 1$ の場合だけを考慮することに相当する。

まとめると、枠組 [Russel 95] では一般性が低い分類子が学習されてしまう場合と、訓練例の一部に矛盾する分類子を学習してしまう場合を考えていない。この結果、Chernoff bounds を適用する必要がない上、悪い分類子を h_b で表すと、 $\Pr(h_b \text{ が学習される}) = (1 - \epsilon)^m$ と簡潔に表せる。枠組 [Russel 95] では、必要例数は (17) に類似する次式で与えられる。ただし H は全ての分類子の集合を表す。

$$m \geq \frac{\ln \left(\frac{|H|}{\delta} \right)}{\epsilon} \quad (29)$$

上式は, (17) において一般性を考えず ($\theta_S = 1$ かつ ζ なし), $\theta_F = 1$ とした場合に対して, ϵ^2 の 2 乗を削除し, 分母の 2 を省略したものとなっている. 最後の 2 点は, Chernoff bounds を適用しなくて済むためである.

なお正確性が低い場合の必要例数に関する研究 [Devroye 96] も存在するが, 本研究とは異なり一般性が低い場合を考慮していない. 関数従属性に関する PAC 学習可能性についての分析 [Akutsu 94, Kivinen 95] は本研究と関連するが, 異なる問題を扱っている.

5.2 Jensen と Cohen の多重比較

Jensen と Cohen の多重比較 [Jensen 00] は, 分類学習に対する慎重な見方を表している. 要約すると, 学習アルゴリズムにおいて考慮する分類子が多いほど, 最終的に得られる分類子の正答率が低くなってしまうことを確率論的に説明している. Jensen と Cohen の多重比較は, 過学習 [Schaffer 93] や過探索 [Quinlan 95] など多くの先行研究を統一的に理解する助けとなる上, いくつかの現実的な対策も提案している.

彼らの研究も分類学習を対象とするため, 前節と同様に一般性は考慮していない. したがって (11) において第 2 項だけを考慮する. ただし前節と異なり $\theta_F < 1$ の場合も考慮するため, 枠組 [Russel 95] に比較して設定がより現実的である. Jensen と Cohen はわれわれと異なり, Chernoff bounds 等の近似式を用いないで悪い分類子が $\lceil m\theta_F \rceil$ 個以上の例に関して正しいクラスを予測してしまう確率を, 二項分布を直接用いて計算している. さらに彼らは, (14) において複数個の悪いルールが発見条件を満たしてしまう場合を重複して数えることを許したわれわれとは異なり, 悪い分類子がまったく発見されない確率を正確に表している. 悪い分類子の集合を H_b で表すと, 彼らの研究では次のようになっている.

$$\Pr(H_b \text{ が学習される分類子を含む}) \\ = 1 - [1 - \Pr(h_b \text{ が学習される})]^{|H|} \quad (30)$$

計算において正確性を求めることは, いわゆる諸刃の剣にたとえられる. Jensen と Cohen は, 悪い分類子を学習しないための必要例数を解析的に求めている. われわれは, これが (30) を m に関して解くことが比較的困難であるためだと推測している. われわれは理論的解析においていくつかの近似を行ったが, これらは必要例数を解析的に求めるために必要であると考えられる. 思想的な相違点として, Jensen と Cohen が学習に関して悲観的であり, われわれがルール発見に関して現実的であることがあげられる. Jensen と Cohen は $|H|$ がきわめて大きいことを強調し, 悪い分類子を学習しないことが難しいことを種々の具体例で示している. われわれは $|R|$ がきわめて大きいことを認識しているが, その上で必要例数を $|R|$ に関して解析的に求めている.

5.3 関連ルール発見の理論的解析

関連ルール発見の解析 [Agrawal 96] は, データベースシステムを利用する枠組における問い合わせ数の下限値, ランダムなデータベースにおいて一般性条件を満たすアイテム集合の期待数, およびデータを復元抽出する場合におけるアイテム集合を満たす例についての解析から構成される. 本研究に深く関連するのは 3 番目の解析であり, 本研究と同様にルール発見において, 真のデータから m 個の例を復元抽出する場合を考慮している.

この解析においては, (6) の Chernoff bounds において, X をアイテム集合 f に関する $m\widehat{\Pr}(f)$ と考える. 最初に右辺 $\exp[-2m(q-p)^2]$ を $\widehat{\Pr}(f)$ が真のデータ集合における値 $p (= \Pr(f))$ から $q-p$ 以上ずれる確率の上限値と見なしている. 次に, $\exp[-2m(q-p)^2] = \delta$ において $q-p$ と δ について種々の数値例を与え, 対応する m の値を求めている.

対象とする発見手法では, まず $\widehat{\Pr}(f) \geq \theta_S$ を満たすアイテム集合 f 全てを Apriori と呼ばれるアルゴリズムで求め, 次にこれらのアイテム集合から関連ルールを生成する. 解析の動機は, 与えられたデータ集合をサンプリングすることで, Apriori の計算時間を短縮することである. したがって, この解析はわれわれのとは異なり, 正確性を考慮していない. さらに関連ルールの総数を考慮していないため, 必要例数をアルゴリズムを適用するデータ集合の属性数や発見する関連ルールの形式などに関連づけていない.

5.4 発見されたルールの信頼性同時評価

発見されたルールの信頼性同時評価 [Suzuki 98] も, 本論文の解析や前節の解析と同様に, ルール発見において真のデータから m 個の例を復元抽出する場合を対象としている. この研究では前節の解析とは異なり, 一般性と正確性を両方とも考慮している.

この研究では, 本研究と同様に (3) の目的を対象としている. ただしサンプリング例数 m を固定し, 閾値 θ_S と θ_F を用いない. x の否定を $\bar{x}$ で表すと, $(m\Pr(x, y), m\Pr(\bar{x}, y))$ が 2 次元正規分布に従うと仮定し, 目的が達成される必要十分条件を解析的に求めている. 本論文はこの枠組とは異なり, 式 (4) において閾値 θ_S と θ_F を固定し, サンプリング例数 m に関する必要数を求めている. 研究 [Suzuki 98] で扱われている問題は, 楕円と 2 直線が接する問題と等価であるためラグランジュの未定乗数法を用いることができ, 必要十分条件の解析解は次の式で与えられる.

$$\left(1 - \beta_\delta \sqrt{\frac{1 - \widehat{\Pr}(y)}{m\widehat{\Pr}(y)}}\right) \widehat{\Pr}(y) \geq 1 - \zeta \quad (31)$$

$$\left(1 - \beta_\delta \sqrt{\frac{\widehat{\Pr}(\bar{x}, y)}{\widehat{\Pr}(x, y)\psi(m, \delta, y)}}\right) \widehat{\Pr}(x|y) \geq 1 - \epsilon$$

(32)

$$\text{ただし } \psi(m, \delta, y) \equiv (m + \beta_\delta^2) \widehat{\text{Pr}}(y) - \beta_\delta^2.$$

β_δ は $1 - \delta$ 信頼範囲, すなわち $(m \text{Pr}(x, y), m \text{Pr}(\bar{x}, y))$ についての楕円の大きさを定める正数であり, 簡単な数値積分によって求めることができる [Suzuki 98]. (31) と (32) はそれぞれ一般性と正確性に関する条件を表しており, 各推定値に楕円として表される信頼性範囲の大きさを表す係数を乗じた値が, それぞれユーザが要求する値 $1 - \zeta$ と $1 - \epsilon$ 以上であることを表している.

この研究は, 確率同時変数に関して分布を仮定するため最悪解析ではない. さらに前節の解析と同様にルールの総数を考慮していないため, 必要例数を発見問題の規模に関連づけていない.

6. む す び

本論文の貢献は, 主に次の 3 点である. 1) ルール発見の最悪解析を定式化した. ルール発見における一般性と正確性を定義し, それらについてアルゴリズムで用いる閾値 θ_S, θ_F とユーザが要求する値 $1 - \zeta, 1 - \epsilon$ が異なる枠組を提案した. ルール発見は, 悪いルールを発見したくない場合と良いルールを見落とさない場合の 2 通りを考えた. 2) ルール発見の最悪解析における必要例数を導出した. 確率論による定式化と適切な近似により, 上記の 2 通りのルール発見について個別に求めた. 定量的分析により, ルール総数 $|R|$, 一般性と正確性に関するマージン $|\theta_S - 1 + \zeta|, |\theta_F - 1 + \epsilon|$ が重要であることが分かった. 3) 連言ルール発見において必要例数を分析した. 典型的な連言ルール発見において必要例数を主要パラメータに関して具体的に検討することにより, 種々の知見を得た.

1) の貢献は, 本研究がルール発見において PAC 学習に相当する枠組を提案したことを示す. この枠組は, PAGA (Probably Approximately General and Accurate) 発見と名づけられた. PAGA 発見は, 必要に応じてデータを要求するアクティブマイニングにおける基盤理論としても有望であると考えられる. 2) と 3) の貢献は, ルール発見アルゴリズムを用いる際における種々の有用な指針を示している. 例えば対象とするルールの種類に応じて用いるデータ集合をサンプリング / 拡張したり, 利用できるデータ集合に応じて発見対象とするルールの種類を拡張 / 限定したりすることに有用である. これらの貢献と 5 章の議論により, われわれのルール発見に関する理解が深まったと結論づけられる. 今後の課題としては, 複数のルールを発見する場合や結論部を限定しない場合など, より現実的な設定における解析などがあげられる.

なお最悪解析は, 導かれる結果が過度に悲観的であるために現実的ではないとの指摘があるが, この指摘に対するわれわれの考えは次の通りである. 最悪解析は, その

定義から最も悲観的な結果を導くことが目的である. 例えば本論文で扱った問題に関しては, 解析の目的は, 実際に必要な例数の上限値を求めることであり, 実際に必要な例数を予測することではない. 本研究で行った最悪解析は, この問題における各種パラメータの影響を記述するため, 実際のルール発見における指針や保証を示し, 有用であると考えられる. すなわち, 属性集合や各種パラメータを決定する際に, 最悪の場合に必要な例数が分かるため, 問題設定などに有用である. なお平均的解析は, 最悪解析に比較してより現実的であるが, より困難である場合が多い. 解析をより容易にするために仮定を導入し対象問題を過度に単純化した場合, 平均的解析は現実的であるとは言えない.

従来, データマイニングにおける代表的な問題であるルール発見においては, アルゴリズムの効率に関する研究が主に行われており, 発見の枠組自体の統計的な立場からの検討はあまり行われていない. 本論文は, ルール発見の理論的な解析を行い, データサイズと期待される最悪時の性能について結果を求めたものであり, 現実の発見問題設定などに有用であると考えられる.

謝 辞

有川節夫先生には, われわれの従来研究と PAC 学習の関係を調べることを提案して頂き, この研究を開始することができました. 匿名の査読者からは, 有益なコメントを頂きました. 記して感謝します. 本研究の一部は, 文部科学省科学研究費特定領域研究「アクティブマイニング」の援助を受けている.

◇ 参 考 文 献 ◇

- [Agrawal 96] Agrawal, R., et al.: Fast Discovery of Association Rules, in Fayyad, U. M., et al. eds., *Advances in Knowledge Discovery and Data Mining*, pp. 307–328, AAAI/MIT Press, Menlo Park, Calif. (1996).
- [Akutsu 94] Akutsu, T. and Takasu, A.: On PAC Learnability of Functional Dependencies, *New Generation Computing*, Vol. 12, No. 4, pp. 359 – 374 (1994).
- [CRI 01] CRISP: *CRISP Process model* (2001), <http://www.crisp-dm.org/>.
- [Devroye 96] Devroye, L., Györfi, L., and Lugosi, G.: Vapnik-Chervonenskis Theory, in *A Probabilistic Theory of Pattern Recognition*, pp. 187–213, Springer-Verlag, New York (1996).
- [Fayyad 96] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P.: From Data Mining to Knowledge Discovery: An Overview, in Fayyad, U. M., et al. eds., *Advances in Knowledge Discovery and Data Mining*, pp. 1–34, AAAI/MIT Press, Menlo Park, Calif. (1996).
- [Feller 57] Feller, W.: *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, New York (1957).
- [Jensen 00] Jensen, D. D. and Cohen, P. R.: Multiple Comparisons in Induction Algorithms, *Machine Learning*, Vol. 38, No. 3, pp. 309–338 (2000).
- [Kearns 94] Kearns, M. J. and Vazirani, U. V.: *An Introduction to Computational Learning Theory*, MIT Press, Cambridge, Mass. (1994).
- [Kivinen 95] Kivinen, J. and Mannila, H.: Approximate In-

- ference of Functional Dependencies from Relations, *Theoretical Computer Science*, Vol. 149, No. 1, pp. 129 – 149 (1995).
- [Quinlan 95] Quinlan, J. R. and Cameron-Jones, R.: Oversearching and Layered Search in Empirical Learning, in *Proc. Fourteenth Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, pp. 1019 – 1024 (1995).
- [Russel 95] Russel, S. and Norvig, P.: Why Learning Works: Computational Learning Theory, in *Artificial Intelligence, a Modern Approach*, pp. 552–558, Prentice Hall, Upper Saddle River, N. J. (1995).
- [Schaffer 93] Schaffer, C.: Overfitting Avoidance as Bias, *Machine Learning*, Vol. 10, No. 2, pp. 153 – 178 (1993).
- [Smyth 92] Smyth, P. and Goodman, R. M.: An Information Theoretic Approach to Rule Induction from Databases, *IEEE Trans. Knowledge and Data Engineering*, Vol. 4, No. 4, pp. 301 – 316 (1992).
- [Suzuki 98] Suzuki, E.: Simultaneous Reliability Evaluation of Generality and Accuracy for Rule Discovery in Databases, in *Proc. Fourth Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 339 – 343 (1998).
- [Suzuki 99] Suzuki, E.: Scheduled Discovery of Exception Rules, in *Discovery Science, LNAI 1721 (DS)*, pp. 184–195, Springer-Verlag (1999).
- [鈴木 00] 鈴木英之進：閾値スケジューリングに基づく仮説駆動型例外ルール発見, 人工知能学会誌, Vol. 15, No. 4, pp. 649 – 656 (2000).
- [鈴木 01a] 鈴木英之進：ルール発見の最悪解析, 第 46 回人工知能学会人工知能基礎論研究会 & 第 54 回人工知能学会知識ベースシステム研究会 合同研究会, pp. 189 – 194 (2001).
- [Suzuki 01b] Suzuki, E.: Worst-Case Analysis of Rule Discovery, in *Discovery Science, LNAI 2226 (DS)*, pp. 365 – 377, Springer-Verlag (2001), (Erratum: <http://www.slab.dnj.ynu.ac.jp/erratumds2001.pdf>).
- [鈴木 02] 鈴木英之進：例外ルールの発見, システム制御情報学会論文誌, Vol. 13, No. 4, pp. 197 – 202 (2002).
- [Valiant 84] Valiant, L.: A Theory of the Learnable, *Communication of the ACM*, Vol. 27, No. 11, pp. 1134 – 1142 (1984).

〔担当委員：鷲尾 隆〕

2002 年 3 月 8 日 受理

著 者 紹 介

鈴木 英之進 (正会員)

1988 年 3 月東京大学工学部航空学科卒業。1993 年 3 月同博士課程修了。博士 (工学)。同年 4 月東京工業大学工学部情報工学科助手。1996 年 4 月横浜国立大学工学部電子情報工学科講師を経て, 1997 年 4 月より同助教授。1997 年度本学会論文賞, 研究奨励賞, 全国大会優秀論文賞受賞。データマイニング, 機械学習, 知識工学に関する研究に従事。情報処理学会, AAAI, ACM, IEEE 各会員。