Rough 集合理論に基づく医療エキスパートシステムのルールの帰納学習*

4J－4

津本周作, 田中博†

東京医科歯科大学難治疾患研究所情報医学研究部門医薬情報‡

1. はじめに

従来の機械学習ではデータベースからの分類知識の獲得が中心であったが、このような分類知識は数千例といった多数のデータを利用しなければ、信頼度の高い知識は獲得できなかった。しかるに、医師をはじめとした専門家は数百例で妥当な知識を獲得する。このような機械学習の手法と専門家の学習手法との大きな違いは、専門家が分類規則のみならず、他の種類の診断規則も獲得しているところにあると考えられる。本研究では、専門家の三種類の診断知識をラフ集合理論に基づいて形式化し、これらの規則を導出するアルゴリズムを開発し、実際の医療データベースに適用、他の分類知識のみを獲得する機械学習システムと比較した。結果として、より医師の判断規則に近い規則が得られた。

2. RHINOS

RHINOS(Rule based Headache and facial pain INformation Organizing System) は、頭痛及び顔面痛の診断を行うエキスパートシステムである [3]。松村らは、医師の診断の過程には次の 3 つの診断論理が用いられているという診断モデルを立て、各々に対するルール:exclusive rule,inclusive rule,disease image の 3 種類のルールとその知識獲得の手法を提唱した。

1) Exclusive Rule: これは疾患を疑うのに必要な所見を集めたものであり、頭痛の中で基本項目を含んだ次の質問項目に従って獲得した専門家の知識に基づいて作成している: 1. 起り得る年齢,2. 痛みの部位,3. 痛みの性状, 4. 痛みの程度,5. 発症様式,6.Jolt Headache の有無。1 から 6 の解答を条件部として、対応する疾患名と組合わせ、Exclusive rule を作成している。

2)Inclusive Rule: このルールの条件部は、問診および身体所見で構成され、これを満足する患者が対応する患者が対応する疾患である確率が高くなるように設定されている。この inclusive rule は、それぞれの疾患につき、次の様な手順に従って獲得した専門家の知識に基づいて作成している: 1. この疾患であると強く疑う時の条件 2. その条件を満たす患者がこの疾患である確率: SI(Satisfactory Index) 3. この疾患の患者全体の中で、この疾患に関するこれまでに得られた条件のどれかを満足する患者の割合: CI(Covering Index) 4.CI=1 であれば、終了。しかし CI=1 でなければ、1 から 4 を繰り返し、同一疾患のセットをいくつか作る。ルール作成終了後、1. で得た症状セットと、2. で得た Satisfactory Index, 疾患名を組合わせ、Inclusive Rule とする。また、セット全体での疾患の cover の度合いとして、CI を定義する。

3)Disease Image: 2) のルールで肯定された疾患について、併発症の論理を用いて、併発の有無を検討する。このために用意されているのが、Disease Image であり、これは所見の中で対応する疾患が起る可能性のあるものを or で結合させ列挙したものである。

3. RHINOS のルールの定式化

3.1. SI と CI

RHINOS において基本的な指標は SI と CI である。SI は、ルールの正確度 (accuracy) に対応し、CI は、ルールの被覆度 (coverage) に対応しているが、これらはラフ集合の記法 [4] で定式化すれば、

$$\mathrm{SI}(R_i,D) = \frac{card\ ([x]_{R_i} \bigcap D)}{card\ [x]_{R_i}},$$

$$\mathrm{CI}(R_i,D) = \frac{card\ ([x]_{R_i} \bigcap D)}{card\ D}$$

ここで、R_iは同値関係であり、$[x]_{R_i}$ は R_iを満たす要素の集合、Dはあるクラスに所属する要素の集合を示す。この式からも明らかなように、SI と CI とは $([x]_{R_i} \cap D)$ を R_iの立場でみるか、Dの立場でみるかによって得られる指標である。

*Induction of Rules for Medical Expert Systems from Clinical Databases Based on Rough Set Theory

†Shusaku Tsumoto and Hiroshi Tanaka

‡Medical Research Institute, Tokyo Medical and Dental University 1-5-45 Yushima,Bunkyo-ku,Tokyo 113,Japan

3.2. ルールの定式化

SI と CI の自然な定式化に基づけば、RHINOS のルールの条件部はラフ集合の記法により、次のように定式化できる。

1)Exclusive Rule: $R_i \quad s.t. \quad CI(R_i, D) = 1.0.$

正確には、R_iの中に含まれる属性には上記のように制限がある。これらの属性を $a_1, a_2, \cdots, a_6$と表せば、$\wedge_{i=1}^{6}[a_i = v_j] \quad s.t. \quad CI([a_i = v_j], D) = 1.0$ と表せる。

2)Inclusive Rule:

$R_i \quad s.t. \quad SI(R_i, D) > 0.75, CI(R_i, D) > 0.5.$

3)Disease Image:

$\vee[a_i = v_j] \quad s.t. \quad CI([a_i = v_j], D) > 0.$

と定義できる。

4. PRIMEROSE-REX

4.1 アルゴリズム

以上のような定式化に基づけば、RHINOS のルールをデータベースから自動的に抽出する問題は SI と CI に関する拘束条件を満たす同値関係を探索する問題に帰着できる。PRIMEROSE-REX (Probabilistic Rule Induction MEthod based on ROugh SEts for Rules of EXpert Systems) は、このような同値関係を探索するプログラムであり、次のようなアルゴリズムで動作する[1]: **(1)** L を属性-値の対のリストとする。**(2)** L からある疾患に所属する標本 Dに関して、$SI(R, D) > 0$ となるような関係式 Rを取り出す。これを Rの Disease Image のリストに加える。**(3)** Rが $CI(R, D) = 1.0$ を満たしていれば、Exclusive Rule のリストに加える。**(4)** Rが $SI(R, D) > 0.75$, $CI(R, D) > 0.5$ を満たしていれば、Inclusive Rule のリストに加える。満たしていなければ、Rをリスト Mに加える。Rを Lから消去し、(5) へ。**(5)** L が空なら、(6) へ。そうでなければ、(2) へ。**(6)** Mが空なら、終了。そうでなければ、Mに含まれるすべての属性-値の対に関して連言を生成し、これを新たな L とし、(2) へ。

4.2 SI と CI の推定

訓練標本による結果は、すべての標本 (母集団) に比べて偏っている (biased) 可能性が高く、この偏りを是正する必要がある。我々はこの目的で、リサンプリング法の中で交叉検証法 (Cross-Validation method) とブートストラップ法 (the Bootstrap method) を導入した [2]。

Cross validation 法 [2] は、ランダムに定められた一定数のデータを抜き取り、抜き取った残りのデータで、判別方式を求め、抜き取った標本で判別方式を評価するという手法である。

Bootstrap 法 [1, 2] は訓練標本 (標本数:n) から経験分布関数 (一般には各標本に $1/n$ の確率を付与) を構成し、そこから重複を許し、n だけ random に sampling し、これによってルールを導出し、このルールをもとの訓練標本をテスト標本として評価する方法である。一般に、この試行が反復的に施行され、各試行の推定値の平均が推定量とされる。Efron[1] によれば、連続値データによる判別分析では cross-validation 法の推定値が unexpected data による推定値 $err_{pattern}$に漸近的に一致し、bootstrap 法の推定値は最尤推定量に漸近的に近づくとされており、カテゴリカルデータにおいても同様の傾向が存在すると考えられる。したがって、bootstrap 法の推定値では、例えば誤判別率の推定値が underestimate 側から最尤推定量に、cross-validation 法では overestimate 側から $err_{pattern}$に近づく [1] ということから、この二法による推定値は、それぞれ実際の真の値の上限と下限を与えるとみなせる。

そこで、我々はPRIMEROSE-REX において両者の併用することによって、SI と CI の不偏推定量を推定し、比較的良い推定値が得た。本大会では、医療データベースを用いた PRIMEROSE-REX の評価について供覧する。

[1] ここでは、exclusive rule に関する属性の制限は含まれていないが、(3) に条件式を加えるだけで、容易に可能である。

参考文献

[1] Efron, B. Estimating the error rate of a prediction rule: improvement on cross validation. *J.Amer.Statist.Assoc.* **78**, 316-331, 1983.

[2] 小西貞則, 本多正幸. 判別分析における誤判別率推定とブートストラップ法, 応用統計学, **21**, 67-101, 1992.

[3] 松村泰志, 松永隆, 木村道男, 前田祐輔, 津本周作, 松村浩. 診断過程のシミュレーション-頭痛 顔面痛診断支援システム RHINOS. 医療情報学 7(2),183-190,1987.

[4] Pawlak, Z. *Rough Sets*, Kluwer Academic Publishers, 1991, Dordrecht.