Rough 集合理論を用いた医療知識の帰納学習*

1P−7

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

従来の機械学習で扱われているデータと診断知識の一番大きな相違は、医学的な知識が不完全で確率的であるという点である。一見して、作用機序が全く異なり、相対するような疾患においても時にほぼ同じ症状を示すこともあり、機械学習を応用する場合もこのような矛盾を含んだ知識の学習も扱えなければならない。

このような矛盾を含んだ知識の獲得には次の二通りの考え方がある。一つは、領域の知識を含めることによって、新たな属性を構成、発見し、それによって分類能力を高めるという方法であり、もう一つは新たな構造を作りだすことなく、既存の属性による確信度を与えることによって信頼性の程度を計量化するという方法である。前者、後者共に領域特有な知識が必要であり、与えられた属性からのルールの導出そのものを「統語論的な知識」の導出と捉らえれば、いわゆる「意味論的な知識」の導出と考えることができる。

今回、我々はデータベースの位相的構造を抽出する rough 集合理論の形式を利用し、データベースから知識を獲得するシステム PRIMEROSE(Probabilistic Rule Induction based on ROugh SEts)[4] を開発した。この手法では、矛盾を含んだ知識に対して、訓練標本から確信度を推定することによってルールの信頼性を評価する方法(cross-validation 法) を導入した。結果として、専門家に近い診断知識と確信度が得られた。

2.Rough 集合理論と PRIMEROSE

Rough 集合理論は、端的に言えば、属性−値の型のデータベースにおいて、ある属性の値で、それぞれのデータの分類を行える強さはどれぐらいかを評価する手法である [3, 5]。例えば、5 例 {1,2,3,4,5} からなる簡単な頭痛の症例のデータベースについて考えてみよう。、nat(性状) という属性の値が"per"(全体) という同値関係を 1,2,5 が満たす時、この同値関係を満足する要素からなる集合を、次のように不可分集合 (indiscernbility set) として定義する: IND(nat="per")=U/(nat="per")={1,2,5} (U:universe の略で、この場合、訓練標本全体を示す)。この中で、1,2 は m.c.h.(筋収縮性頭痛) であり、5 は psycho(心因性頭痛) であると仮定すれば、nat="per" という項目を使えば、頭痛の診断を完全に行うことはできず、上のような曖昧さが残ることになる。しかし特異的な属性であればあるほど、この曖昧さが減じてくるに違いない。このような属性の値の組合わせを調べることで、その属性-値のセットでどこまで診断可能かを評価できる。例えば、IND((loc="who") and (nat="per"))={2} (M2:圧痛点 M2) が成立すれば、この属性-値のセットで、m.c.h. が分類できることがわかる。以上の基本的概念に基づき、導出すべきルールとして次の形式を設定した:

Definition 1 (ルールの定義) 関係 R_i を同値関係、X をある所属すべきクラスとし、$IND(X)$ を U の部分集合とする。次に、X^c を未観測の集合 U^c の部分集合であるとする。この時、次のような条件を満たす R_i を *probabilistic rule* の条件部と定義する:

$$IND(R_i) \cap IND(X) \neq \phi$$

次に、SI と CI は次のように定義する:

$$SI(R_i,X) = \frac{card\ \{(IND(R_i) \bigcup IND^c(R_i)) \bigcap (IND(X) \bigcup IND^c(X))\}}{card\ \{IND(R_i) \bigcup IND^c(R_i)\}}$$

$$CI(R_i,X) = \frac{card\ \{(IND(R_i) \bigcup IND^c(R_i)) \bigcap (IND(X) \bigcup IND^c(X))\}}{card\ \{IND(X) \bigcup IND^c(X)\}}$$

*Inductive Learning of Medical Knowledge based on Rough Sets

†Shusaku Tsumoto and Hiroshi Tanaka
‡Medical Research Institute, Tokyo Medical and Dental University 1-5-45 Yushima,Bunkyo-ku,Tokyo 113,Japan

以上により、*probabilistic rule:* $R \Rightarrow X\ (SI, CI)$ は$< X, R_i, SI(R_i, X), CI(R_i, X) >$なる組で表される。 □

ここで、SI,CI は、エキスパートシステム RHINOS[1] で使用したものを再定義したもので、SI はある症候を満たすもののなかで疾患である確率、CI はある疾患のなかでその症候をもつ確率を示す。また、$IND^c(X), IND^c(R_i)$ は、データベースには与えられていない未観測な事象 U^c の中で、それぞれ所属クラスが X、関係記述が R_i となる不可分集合を示す。このルールの導出は次の条件部を最小限の属性にまとめる **reduction** 及び SI,CI の推定を行う **cross-validation 法**の二種類の手続きからなる。

1)Reduction: PRIMEROSE では、reduction の対象として、全ての属性を使った不可分集合（素クラスター）を扱い、この要素数が変化しないような形で変数の削除を行う。つまり、全ての属性の数を p、その集合 Attr を $\{a_1, a_2, \cdots, a_p\}$ とする。indiscernbility 集合を表現するのに使われている属性の組合わせを R_i, 属性数を $|R_i|$ とする。$|R_i| = p$ となるような属性の組合わせによって得られる indiscernbility 集合を素クラスター (primitive cluster) と定義する。形式的には、$Prim(R_i) = IND_{|R_i|=p}(R_i) = \bigcap_i^p IND(a_i = v_i)$ の形で与えられる。ここで、ある関係 R_i の記述に使用する属性数を $|R_i|$、属性の集合を $A(R_i)$ で表す。この時、もし、$A(R_i) \subseteq A(R_j)$ をみたせば、$R_i \preceq R_j$ と表せば、上の定義を使い、素クラスターという不可分集合の要素を一定に保つ形で、reduction を次のように定義できる: もし、属性 a が等式 $Poss_{R_i-\{a\}}(X) = Poss_{R_i}(X) = Prim(R)(R_i \preceq R, |R| = p)$ を満たす時、属性 a は Rに *dispensable* といい、a は関係 R_i から削除可能である。

2) SI,CI の推定 − cross-validation 法: 上記の定義では、、実際にはこの未観測のデータを得ることは不可能で、何等かの手法でそれに相当するものを構成することが必要である。この構成手法として、PRIMEROSE では cross-validation 法 [2] を導入した [4]。

cross validation 法では、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。具体的には、training sample$\mathcal{L}$ を V個の部分集合 $\mathcal{L}_v (v = 1, \cdots, V)$ に random に分割し、V 個の訓練標本、$\mathcal{L}^v = \mathcal{L} - \mathcal{L}_v$を作り、それぞれに対して、$\mathcal{L}_v$ をテスト 標本として誤判別率を求める手法を、V-fold cross validation と呼ぶ。ここで、上記の未観測の集合としてこのテスト標本の値を利用し、SI,CI を推定する。

3. 結果

以上述べた手法を用いて、頭痛、髄膜炎、能血管障害それぞれ 100 症例のデータベースを訓練標本としてルールの導出を試みた。さらに、100 症例をテスト標本として利用した。この PRIMEROSE の手法と既存の AQ15,CART に対する正解率とルール当たりの属性数を比較した。結果として正解率は平均で CART 74.5%, AQ15 75.2% PRIMEROSE 87.5%と PRIMEROSE において良好であり、属性数の平均も、CART 9.4, AQ15 3.2 PRIMEROSE 5.9 であり、これらは専門家によるルールの場合の平均 6.5 に最も近い値が得られた。

参考文献

[1] 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.

[2] McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition,* John Wiley and Sons,1992.

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

[4] Tsumoto,S. and Tanaka,H. Induction of Probabilistic Rules based on Rough Set Theory. in:*ALT-93*,Springer Verlag,1993.

[5] Ziarko,W The Discovery,Analysis, and Representation of Data Dependencies in Databases, in:*Knowledge Discovery in Database*,Morgan Kaufmann, 1991.