ラフ集合理論に基づく医療データベースからの漸増型学習システムの開発*

4M－8

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

1980年代後半より、医療データベースからの知識獲得のため、種々な機械学習のシステムが開発されてきた。しかしながら、これらのシステムは与えられたデータベースから一括して、ルールを静的に導出するものであり、動的な学習のためには、再度データベースを与えて、プログラムを再起動しなおす方法しか適用できなかった。今回、我々はラフ集合理論[2]に基づいて、時系列的に症例が与えられた場合に、漸増学習を進め、ルールを改訂していくシステムを開発した。本論文では、このシステムにおける手法、また前システムPR IMEROSEとの動作の比較、及びこの手法の問題点について論ずる。

表 1: 頭痛の簡単なデータベース1

U	部位	性状	様式	吐き気	所属クラス
1	全体	持続性	持続性	なし	筋収縮性頭痛
2	全体	持続性	持続性	なし	筋収縮性頭痛
3	側頭部	拍動性	持続性	なし	偏頭痛
4	全体	持続性	持続性	あり	偏頭痛
5	全体	持続性	持続性	なし	心因性性頭痛

2. 確率的規則 (Rule) の構造

ルールとしては、以前PRIMEROSEで報告したエキスパートシステムRHINOS[1]で定義したInclusive Ruleの形式を利用する[3]。このルールは命題とSI、CIという二つの指標を持ったものである。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言葉で定式化すれば、

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

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

ここで、R_iは同値関係であり、$[x]_{R_i}$はR_iをみたす要素の集合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。

ルールの定式化

SIとCIの自然な定式化に基づけば、ルールの条件部は次のように定式化できる。

$$R_i \quad s.t. \quad SI(R_i, D) > \delta_{SI}, CI(R_i, D) > \delta_{CI}$$

ここで、δはそれぞれのいき値を示しているが、本研究では、$SI \geq 0.5$、$CI \geq 0.3$とした。つまり、ルールの中で、50%以上の正確度を持ち、30%以上の被覆度をもつルールをInclusive Ruleと定義している。例えば、表1では、筋収縮性頭痛をクラスの例にとれば、

$$[部位 = 全体]\&[吐き気 = なし] \rightarrow 筋収縮性頭痛$$
$$SI: 2/3 = 0.67, CI = 1.0$$
$$[性状 = 持続性]\&[吐き気 = なし] \rightarrow 筋収縮性頭痛$$
$$SI: 2/3 = 0.67, CI = 1.0$$
$$[吐き気 = なし] \rightarrow 筋収縮性頭痛$$
$$SI: 1/2 = 0.50, CI = 1.0$$

がinclusive ruleとして求められる。

3. 漸増学習のアルゴリズム

ここでは、まず属性=値の対をelementary relationと呼ぶことにする。我々のアルゴリズムでは新たなサンプルが加わった時、elementary relationのSIとCIとを改訂し、その改訂した値によって、新たなルールが導出可能かを探索していく。この場合、サンプルあるいは関係記述に対するSIとCIのsensitivityが問題になるが、CIの方が両因子に対するsensitivityが低い。特に、後者に関しては、ルール中の属性の数が増えれば、一般にCIの値は減少する。したがって、elementary relationの中でCIがいき値δ_{CI}以上であ

*Incremental Learning 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

るものを選び、それらの連言を生成し、SI がいき値 δ_{SI} 以上の関係記述を探索すればよい。

つまり、新たなサンプルが加わった時、まず elementary relation の CI を改訂する。CI がいき値より増加した elementary relation は関係記述生成の候補とし、CI がいき値より減少したものに関しては、それを含んだルールを削除する。ただ、将来のサンプルによって再び削除されたルールがまたルールとして採択される可能性があるので、削除されたものは次候補としてプールしておく。

次に関係記述生成の候補に関して、すでに関係記述が存在しているものに関しては、SI と CI を改訂し、新たに加わったものに関してはこれを含んだ関係記述を生成する。

以上のようにして、漸増学習アルゴリズムは次のように定義できる。

1. 新たなサンプルが加わった際、elementary relation に関して SI と CI を改訂する。
2. 改訂した CI に関して、新たに $CI \geq \delta_{CI}$ となった elementary relation を $List1$ に $CI < \delta_{CI}$ となった elementary relation を $List2$ に store する。
3. $List1$ に関して、その要素を含んだ新たなルール ($SI > \delta_{SI}$, $CI > \delta_{CI}$) を $List_a$ から探索し、ルールをリスト $List_r$ に加える。
4. $List2$ に関して、その要素を含んだルールを $List_r$ から削除し、次候補のリスト $List_a$ に加える。
5. $List1$ と $List2$ に含まれない elementary relation からなるルールに関しては、SI と CI とを改訂し、$SI > \delta_{SI}$, $CI > \delta_{CI}$ をみたさないものは $List_a$ に加える。

例えば、表1の例では、1-5 までの例では筋収縮性頭痛に関して、第2節で示したルールが得られる。ここで、次のような標本が追加されたとする。

U	部位	性状	様式	吐き気	所属クラス
6	側頭部	拍動性	持続性	なし	心因性頭痛

ここで、新たに $CI > 0.3$ あるいは $CI \leq 0.3$ となる elemetary relation は存在しないが、$CI > 0.3$ である属性からなるルールで、第2節で述べたルール

$$[\text{吐き気} = \text{なし}] \rightarrow \text{筋収縮性頭痛}$$

は SI: 2/5=0.4 , CI=1.0 となるから、$List_a$ に、

$$[\text{性状} = \text{持続性}]\&[\text{吐き気} = \text{なし}] \rightarrow \text{筋収縮性頭痛}$$
$$[\text{性状} = \text{持続性}]\&[\text{吐き気} = \text{なし}] \rightarrow \text{筋収縮性頭痛}$$

は SI: 2/3=0.33 , CI=1.0 のままで登録される。

4. 評価

本研究では、髄膜炎のデータベース (99 例) を使用した。計算を簡単にするため、今回は診断に関わるルールの導出のみを考え、属性数を 20 に制限した。このデータベースに関して、repeated 10-fold cross-validation(反復回数:100 回) で分類の正確度、ルールの生成数、領域使用量、計算時間とを一括型のルール導出システム PRIMEROSE 及び Shan と Ziarko の決定行列による方法と比較した。なお、計算は 486Dx-75MHz の PS-2 compatible 機を用いた。表 2 に上記の評価の結果を示す。上記のように、漸増学習システム PRIMEROSE-INC は学習システム PRIMEROSE と同等のルールを導出できるが、そのために領域と時間計算量を犠牲にしなければならないことがわかった。

表 2: 実験結果 1

手法	正確度	ルール生成数
PRIMEROSE-INC	$81.5 \pm 3.2\%$	52.3 ± 1.4
PRIMEROSE	$81.5 \pm 3.2\%$	52.3 ± 1.4
決定行列	$72.1 \pm 2.7\%$	12.9 ± 2.1

表 3: 実験結果 2

手法	領域使用量	計算時間
PRIMEROSE-INC	1241 ± 34	1027 ± 71 sec
PRIMEROSE	210 ± 14	521 ± 11 sec
決定行列	119 ± 12	124 ± 15 sec

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

[3] Tsumoto, S. and Tanaka, H. PRIMEROSE: Probabilistic Rule Induction Method based on Rough Sets and Resampling Methods. *Computational Intelligence*, **11**, 389-405, 1995.