可変精度ラフ集合理論に基づく医療データベースからの知識獲得*

2C－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

前大会で我々はラフ集合理論に基づいて三種類の診断知識を医療データベースから獲得するシステムPRIMEROSE-REX について報告した [4]。このシステムはラフ集合理論を確率的に拡張したものを利用したが、理論的な基盤についての検討は不十分であった。今回、我々は本システムでの確率的な拡張を Ziarko によって提唱された可変精度ラフ集合理論 (Variable Precision Rough Set Model: VPRS)[5] の枠組みで形式化した。VPRS は accuracy にある precision を設定することで、ラフ集合理論における命題とそれを支持する標本の集合との関係を自然に拡張したものであるが、本システムで利用した形式はこの VPRS にさらに被覆度 (coverage) を導入することで拡張した形式であることが確認された。

2. 可変精度ラフ集合モデル

2.1 Pawlak のラフ集合モデル

ラフ集合における基本的な考え方は、ある命題記述の背景には集合論的な包含関係があるというところにある [2]。まず、集合間の関係と命題による関係記述には次のような三つの関係が成立することが明らかであろう。

$$\begin{array}{lll} (1) & p \rightarrow \neg q & : [x]_p \cap [x]_q = \phi \\ (2) & ? & : [x]_p \cap [x]_q \neq \phi \\ (3) & p \rightarrow q & : [x]_p \subset [x]_q \end{array}$$

以上のことから、分類に関する次の三つの形態が定義できる。

$$\begin{array}{lll} \text{(1) negative region} & : p \rightarrow \neg D & : [x]_p \cap D = \phi \\ \text{(2) boundary region} & : ? & : [x]_p \cap [x]_q \neq \phi \\ \text{(3) positive region} & : p \rightarrow q & : [x]_p \subset [x]_q \end{array}$$

ここで、$[x]_p$はある命題 p をみたす標本の集合、Dは分類したいクラスに所属する集合を示している。

ラフ集合論は上記のうち、boundary region と positive region についての性質を研究する領域であるが、命題をある集合 (標本) から導出するという立場からいえば、positive region を探索する手続きは実は命題記述を導出する手続きに同等である。

これに対して、Pawlak は positive region のみに着目した学習モデルをラフ集合論の枠組みで考察した [2]。しかしながら、Pawlak の理論は決定論的な命題を導出するものであり、boundary region の構造は明らかにしていない。

2.2 Ziarko の可変精度ラフ集合モデル

Ziarko は、Pawlak の方法が positive region に基づく事からの問題点を boundary region も使うことにより解決しようとした [5]。しかし、ただ boundary region を使うのみでは分類能力の低い決定アルゴリズムを導出する可能性がある。このため、まず誤判別率 (= 1.0−正確度) にいき値 (precision)βを設定し、いき値以下の誤判別率を示す決定アルゴリズムの導出を行うように拡張した。この場合、いき値以上の正確度を示す boundary region の領域が拡張された postive region $\underline{R}_\beta C$の中に含まれることになる。このような拡張は Pawlak の定義によって得られる positive region の性質を保存する事が証明されており、Pawlak の方法自体は、このモデルの中で、いき値を 1.0 にした場合に Pawlak の方法が包含されることになり、自然な拡張方法となっている。

まず、Ziarko の定式化では、正確度α に対して次のような誤判別度 $c(B,C)$ を定義する。

$$c(B,C) = \begin{cases} 1 - \frac{card(B \cap C)}{card(B)}, & if \quad card(B) > 0 \\ 0, & if \quad card(B) = 0 \end{cases}$$

この定義を利用すれば、c(B,C)=0 の時、B が C の部分集合、つまり positive region を表すことが示せる。

$$B \subseteq C \quad iff \quad c(B,C) = 0$$

---

*Inductive Learning of Medical Knowledge Based on Variable Precision Rough Set Model

†Shusaku Tsumoto and Hiroshi Tanaka

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

ここで、precision をβとすれば、上記の包含関係の拡張として、

$$B \overset{\beta}{\subseteq} C \quad iff \quad c(B,C) \leq \beta$$

を定義できる。したがって、β-positive region として、

$$\underline{R}_\beta C = \bigcup \{E \in R^* | E \overset{\beta}{\subseteq} X\}$$

が定義できる。ここで、R^*は問題空間 U を同値類で分割したものを要素としたの集合を示している。

したがって、上記の包含関係は、誤判別率 $c(B,C)$ にあるいき値βを付与した形の命題に対応していると考えられる。つまり、上記のβ-positive region に対応して、$R \overset{c([x]_R,C)<\beta}{\rightarrow} C$で定義される拡張された形の命題を考えることができる。

3. RHINOS のルールの定式化

3.1. SI と CI

前大会において導入したルールに付与する基本的な指標はルールの正確度 (accuracy)$\alpha_{R_i}(D)(= 1 - c([x]_{R_i}, D))$ とルールの被覆度 (coverage)$\kappa_{R_i}(D)$ である。前者はある同値関係 R_iをみたす標本が Dという集合に含まれる割合を示し、後者は同値関係 R_iをみたし、Dに属する標本が D全体の中に含まれる割合を示している。

これらはラフ集合の記法で定式化すれば、

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

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

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、正確度と被覆度とは$([x]_{R_i} \cap D)$ 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り、これらは Rough Membership Function の一種である [3]。

3.2. ルールの定式化

上記の定式化に基づけば、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: $R_i \quad s.t. \quad \kappa_{R_i}(D) = 1.0.$

(2)Inclusive Rule: $R_i \quad s.t. \quad \alpha_{R_i}(D) > 0.75,$
$\kappa_{R_i}(D) > 0.5.$

(3)Disease Image: $\vee[a_i = v_j] \quad s.t. \quad \kappa_{[a_i=v_j]}(D) > 0.$

と定義できる。

4. 可変精度ラフ集合モデルによるルールの再定式化

上記の第2,3 節で明らかなように、我々のモデルは可変精度ラフ集合モデルの枠組みで記述することが可能である。まず、一般的に我々の用いている確率的なルールは次のように定義できる。

Definition 1 (確率的ルール) R_iをある同値関係、Dをあるクラスdに所属する集合とする。この時、dに関する確率ルールは以下の条件をみたす三つ組$< R_i \overset{\alpha,\kappa}{\rightarrow} d, \alpha, \kappa >$で定義できる: *(1)* $[x]_{R_i} \bigcap D \neq \phi$, *(2)* $\alpha = \alpha_{R_i}(D)$, *(3)* $\kappa = \kappa_{R_i}(D)$. □

以上の議論に基づけば、上記3 種類のルールは次のように記述できる:

(1)Exclusive Rule: $R_i \overset{\kappa=1.0}{\rightarrow} d,$

(2)Inclusive Rule: $R_i \overset{\alpha>0.75,\kappa>0.5}{\rightarrow} d,$

(3)Disease Image: $\{[a_i = v_j] \overset{\alpha\geq 0}{\rightarrow} d\}$

したがって、上記のルールは可変精度ラフ集合モデルに被覆度をさらに付与したものと考えることができ、正確度, 被覆度及びそれらに対する条件を付与した拡張形の命題であるとみなせる。

本大会ではこれらの可変精度ラフ集合モデル及び我々のルールモデルの性質について論じる。

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] Pawlak, Z. and Skowron, A. Rough Membership Function. in: Yager, R. et al. (eds.) *Advances in the Dempster-Shafer Theory of Evidence*, pp. 251-271, John Wiley & Sons, 1994, New York.

[4] 津本周作, 田中博. 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. 第52 回情報処理学会全国大会抄録集.

[5] Ziarko, W. Variable Precision Rough Set Model, Journal of Computer and System Sciences, **46**: 39- 59, 1993.