

医療診断を対象とした事例間の関係の構造化に基づく
事例ベース推論システムの作成

7N-2

小林祐治* 永田守男* 吉田勝美**

*慶應義塾大学理工学部

**慶應義塾大学医学部

1. はじめに

事例ベース推論(CBR)は、従来のルールベース推論によるエキスパートシステムの開発における知識獲得の困難さを大幅に軽減できると期待されている。ところが、ルール獲得の代わりに、事例の集合の構造化、与えられた問題と過去の事例との類似度の尺度、類似した事例を必要に応じて修正するための知識、などが必要となる。

本研究では、事例の集合の構造化に対しては、学習アルゴリズムを応用した手法を、問題と事例間の類似度の定義に対しては、事例から得られた確率を用いた手法をそれぞれ提案する。そして、甲状腺疾患の診断を例に、診断型エキスパートシステムを作成し、実際の医師の診断結果との比較により、典型的な事例を調べるだけでも全事例を調べた場合と同等の正答率が得られることを確認した。このことから、診断効率の向上が期待される。

2. 分類木を用いた事例集合の構造化

事例の集合の構造化は、事例の検索効率向上やコスト軽減に必要不可欠である⁽¹⁾。本研究では、システムによる自動的な構造化と、全事例を説明しうる典型的な事例の自動的な抽出を行う手法とを提案する。すなわち、機械学習システムID3⁽²⁾で考案された学習アルゴリズムを応用し、典型的な事例だけで構成される分類木を作成する。この手法では、事例を分類するための指標であるエントロピーを原因ベクトルとして定義される診断結果の各原因ごとに求める。これによって、2種以上の原因を診断結果とする複合診断を含む問題にも対応できる。

今、事例の中に含まれる属性全体を集合A(={A₁, A₂, ..., A_m})とし、これらの各属性A_iのとりう

る値の集合をRange(A_i)とする。一方、診断結果が持つ原因の集合をD(={D₁, D₂, ..., D_p})とし、各原因D_jは{1,0}の値を持つものとする。個々の事例は、この属性ベクトル(v₁, v₂, ..., v_m)と原因ベクトル(d₁, d₂, ..., d_p)で表現される。ここで、v_i ∈ Range(A_i) (i=1, ..., m)、d_j ∈ {1,0} (j=1, ..., p)とする。

また、事例を分類する属性を選ぶための指標に、属性A_iを知ることで得られる情報量をエントロピーの減少量としてとらえた相互情報量⁽³⁾Gを用いる。

$$G = I(S) - E(A_i, S)$$

ただし、

$$I(S) = \sum_{k=1}^p (P_{S, D_k} \log_2 P_{S, D_k} + (1 - P_{S, D_k}) \log_2 (1 - P_{S, D_k}))^2$$

$$E(A_i, S) = \sum_{v_j \in \text{Range}(A_i)} \frac{|S_{ij}|}{|S|} I(S_{ij})$$

なお上式中で、S_{ij}は事例集合S中で属性A_iの値がv_jの事例集合、P_{S, D_k}は事例集合S中で診断結果D_kを持つ事例の発生確率である。この手法の簡単なアルゴリズムを表1に示す。

本手法の特徴は、複合診断(2種以上の原因を持つ診断)を単項診断(1種の原因しか持たない診断)に含めることで診断結果を{(d₁, d₁+d₂, d₁+d₃, ...), (d₂, d₂+d₁, d₂+d₃, ...)}のようにグループ分けし、グループによる構造化を行うことである。これにより、診断結果の種類増加に伴う分類木の巨大化を防ぐことができる。

表1 構造化のアルゴリズム

Step 1	与えられた事例の集合S ₀ から、大きさWの部分集合を選び出す(Wはウインドサイズ、選ばれた部分集合はウインドウと呼ばれる)。
Step 2	現在のウインドウの分類木を以下の手順に従って作成する。 (1) ウインドウWをすべて根ノードに対応させる。(S=W) (2) 事例集合Sにおいて、全ての事例が同一のグループの診断結果を持てば、葉ノードとして、診断結果をラベルづけし、停止する。 (3) さもなければ、全ての属性の相互情報量を求め、最大の値を取る属性A _{max} を分岐ノードのテスト属性とし、Sを部分集合に分割する、その部分集合に対し(2)からの手続きを再帰的に繰り返す。
Step 3	事例集合S ₀ 中の全ての事例を調べて、葉ノードにつけられているラベル以外の診断結果を持つ(例外の)事例を見つける。
Step 4	例外の事例が存在しなければ、Step 5へ移る。さもなければ、Step 3で得られた例外の事例のうちいくつかをウインドウに追加、Step 2からの手順を繰り返す。
Step 5	最後にウインドウの中の事例をそれが属す葉ノードに格納する。

A Case-based Reasoning System Generating the Structure of Cases of Medical Diagnoses

Yuji Kobayashi, Morio Nagata, Katumi Yoshida
Keio University

3. 確率を用いた類似度の定義

最適な事例を選びだす尺度である問題と事例間の類似度は、これが適切に定義されるかどうかで、C B Rシステムの成否を大きく左右するほど重要な要素である⁽⁴⁾。本研究では、事例から得られた確率を用いて類似度を定義する。確率とは、原因 D_k のもとで、特定の性 A_i の値が v_j である条件つき確率 C_{ijk} である。その確率を用いて、原因 D_k における属性 A_i の重要度を次式で計算する。

$$E(A_i, D_k) = \sum_{v_j \in \text{Range}(A_i)} \frac{|S_{ijk}|}{|S_k|} C_{ijk} \log_2 C_{ijk} + \log_2 |\text{Range}(A_i)|$$

ここで、 S_k は原因 D_k を診断結果に持つ事例の集合、 S_{ijk} は事例集合 S_k 中で属性 A_i の値が v_j の事例の集合である。そして、問題や事例をこの重要度によって位置付けられた属性の多次元ベクトルで表現し、類似度をその問題と事例の距離によって定義する。

4. 実験システムの作成

甲状腺疾患の診断を例にした実験システムを作成した。このシステムは特徴(患者の兆候)を入力、診断結果を出力とする。またこの問題は、原因(病態)として、{バゼドウ病、慢性甲状腺炎、結節性甲状腺腫}の3種を持つ複合診断問題で、属性(兆候)には、医師の診断カルテから診断に必要な17の兆候を用いる。なお、定量的データは医師の経験的な基準により、すべて定性的データに変換して用いる。総事例数1433件のうち、718件を訓練事例、715件をテスト事例とし、システムを作成した。

5. 評価

本研究では、事例の検索範囲の問題と複合診断の扱いの問題について評価実験を行った。

5.1 検索範囲

以下の3つの検索方法について評価実験を行い、その結果を表2に示す。

1. 構造化された事例ベースの典型的な事例の検索 (ウィンドウサーチ)
2. 構造化された事例ベースの全事例の検索 (全サーチ)
3. 無構造の事例ベースの全事例の検索 (単純サーチ)

表2を見ると、正答率は3方法とも大差がないが、検索時間はウィンドウサーチが圧倒的に早い。このことから、検索範囲は典型的な事例のみの検索で十分であるといえる。

表2 検索範囲の問題の実験結果

	正答率	平均検索時間
ウィンドウサーチ	82.9%	0.053sec
全サーチ	84.2%	1.211sec
単純サーチ	83.6%	4.890sec

5.2 複合診断

複合診断をそのまま複合の病態として扱ったケース1とまったく別の病態として扱ったケース2の2通りの診断方法について評価実験を行った。

ケース1 {バゼドウ病, 慢性甲状腺炎, 結節性甲状腺腫}

ケース2 {バゼドウ病, 慢性甲状腺炎, 結節性甲状腺腫,

バゼドウ病+慢性甲状腺炎, ...}

その実験結果を表3に示す。

表3 複合診断問題の実験結果

	狭義の正答率	広義の正答率	平均の木の高さ	木に含まれる事例数	平均検索時間
ケース1	82.9%	95.8%	4.8	151	0.053sec
ケース2	85.3%	86.9%	6.5	224	0.014sec

※ 狭義の正答率とは最も類似度の高い事例の診断結果が医師の診断結果と完全に一致した場合、広義の正答率とは上位3つの事例の中で一致した事例がある場合を指す。

表3を見ると、どちらが有効かは一概には言えず、扱う問題によって有利、不利がある。結局、本研究の提案は、原因の個数が多く記憶コストを重視する問題で、主要原因だけわかればよいものに適している。

6. おわりに

本研究では、事例間の関係の構造化や類似度の定義にID3のアルゴリズムや事例から得られる確率を用いた手法を提案した。この提案に基づく実験システムを作成し、評価実験と考察により、有効性が確認された。より実用的なC B Rシステムを開発するには、本研究に含めなかった事例の修正法や修復法など、多くの研究課題が残されている。

参考文献

- (1) 小林重信: 事例ベース推論の現状と展望, 人工知能学会誌, 7, pp.559-566 (1992)
- (2) Quinlan, J. R.: Induction of Decision Tree, Machine Learnig 1, pp.81-106 (1989)
- (3) 宮川, 原島, 今井: 情報と符号の理論, 岩波講座 情報科学, 4, pp.35-59 (1982)
- (4) 松原仁: 推論技術の観点からみた事例に基づく推論, 人工知能学会誌, 7, pp.567-574 (1992)