1F4-02

生物医学文献データベースを利用するデータマイニング

Data mining using biomedical literature databases

チャン・ナム・トアン[*1]　　沼尾 正行[*2]

Tuan-Nam TRAN　　Masayuki NUMAO

[*1]北陸先端科学技術大学院大学 知識科学研究科

School of Knowledge Science, Japan Advanced Institute of Science and Technology

[*2]大阪大学 産業科学研究所

The Institute of Scientific and Industrial Research, Osaka University

This paper presents a method for realizing active mining. We have constructed a mining system which combines active information gathering and user-centered mining together. Our method modifies C4.5rules by taking into consideration the external weight of each attribute, which can be calculated by means of the number of corresponding documents found in the literature. The experiments on medical data show that our proposed system is useful in terms of generating interesting rules as well as providing a solution for realizing active mining.

1. 研究の背景と目的

大量なデータの増加により, 巨大なデータベースからの知識発見は重要な分野として発展している. 医療データから重要な知識を抽出するのは, データマイニング手法の比較のためにしばしば用いられている. 医学からの知識発見においては, 生成された知識を専門医の知識から解釈する過程が必要なだけでなく, 前処理の段階でも, 専門医による専門知識が重要な役割を果たしている. しかし, 専門医からの知識をデータマイニングシステムに反映させるには, 大変な手間がかかる. そこで, 巨大な生物医学文献データベース MEDLINE から, 与えられたデータと関係のある情報を能動的に収集することにより, 専門家の負担を減らすことを試みた. データマイニング手法と組み合わせることにより, 新しい知識を発掘することも目指し、命題学習 C4.5 分類アルゴリズムにおいて, 各属性の MEDLINE における重みを考慮する手法を提案して, 医療データで評価を行なった.

2. 提案手法

実装したシステムは 2 段階を含んでいる. 第 1 段階では, MEDLINE における各属性の重みを計算し, 様々なヒューリスティック関数を使用しルールを生成する. 第 2 段階では, 予めユーザにより与えられた条件を満たすために, 得られたルールをフィルタリングし, 出力結果とする.

T を $A_1, A_2, \ldots, A_m$ からの決定木の訓練例とし、$fr(C_i, S)$ をクラス C_i に属する S の数とすると, T の エントロピー は次のようになる.

$$info(T) = -\sum_{j=1}^{k} \frac{fr(C_j,T)}{|T|} \times log_2\left(\frac{fr(C_j,T)}{|T|}\right) \quad (1)$$

ある属性 A_j の $gain$ と $gain\ ratio$ は以下のようになる [1].

$$info_j(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i) \quad (2)$$

$$gain_j = info(T) - info_j(T) \quad (3)$$

$$split\ info_j = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2(\frac{|T_i|}{|T|}) \quad (4)$$

$$gain\ ratio_j = gain_j / split\ info_j \quad (5)$$

各属性の「外部重み」を用いて上記のヒューリスティック関数を以下のように修正する.

$$gain_j^{'} = G(gain_j, \omega_j) \quad (6)$$

$$gain\ ratio_j^{'} = G(gain\ ratio_j, \omega_j) \quad (7)$$

ここでは, ω_j は属性 A_j の外部重みであり, A_j は与えられたデータと属性 A_j に関連する MEDLINE 文献数とする.

$$\omega_j = \frac{F(|A_j|)}{\sum_{i=1}^{m} F(|A_i|)} \quad (8)$$

３種類の関数 g_i $(i = 1,\ 2,\ 3)$ は,*costs of tests* に関する研究で用いられたものである.

$$g_1(x,w) = x \times w \quad (9)$$

$$g_2(x,w) = x^2 \times w \quad (10)$$

$$g_3(x,w) = (2^x - 1) \times \frac{w}{1+w} \quad (11)$$

関数 f に関しては, 以下のように線形関数と対数関数を用いている.

$$F_1(x) = x \quad (12)$$

$$F_2(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \lfloor (log_k(x)+1) \rfloor & \text{if } x > 0 \end{cases} \quad (13)$$

表 1 は, 現在に使用されているヒューリスティック関数のリストを示している.

2.1 ルールに関する指標

得られたルールを医学的な意味を判断する前に, データから何人の患者がそのルールを満たしているかを考慮することが重要である. その理由により, 本研究では, 従来のサポート, 確信度ではなく, 患者数ベースの指標を用いるとする. 例えば, ルー

表 1: 使用されているヒューリスティック関数

No.	ヒューリスティック関数	関数 f	関数 g
1	none		
2	log10	F_2 $(k=10)$	G_1
3	log	F_2 $(k=e)$	G_1
4	linear	F_1	G_1
5	log10 + CS-ID3	F_2 $(k=10)$	G_2
6	log + CS-ID3	F_2 $(k=e)$	G_2
7	linear + CS-ID3	F_1	G_2
8	log10 + EG2	F_2 $(k=10)$	G_3
9	log + EG2	F_2 $(k=e)$	G_3
10	linear + EG2	F_1	G_3

ル $A \to C$ のサポートとは, 条件 A と C の両方を満たした患者数とする. あるルールの確信度は, 以下のように定義される.

$$conf(A \to C) = \frac{sup(A \to C)}{sup(A)}$$

提案手法の概略を図 1 に示す.

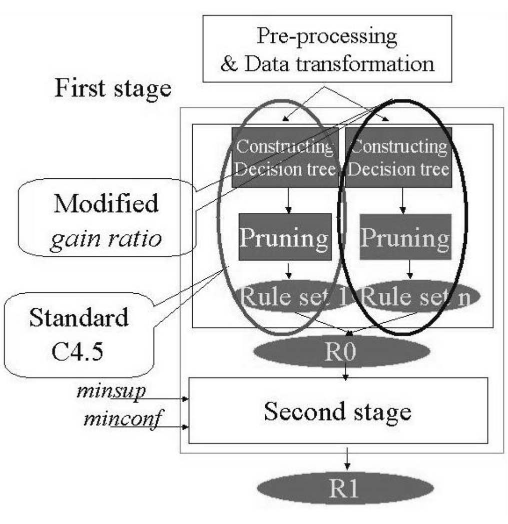

図 1: 提案手法の概略図

3. 実験

提案手法を用いて肝炎データで実験を行った. このデータは, 千葉大学付属病院第一内科の外来受診歴がある B 型, C 型の慢性肝炎患者のうち, １９８２年から２００１年の間に肝生検を受けている７７１名に関するものである. これまでに, 次のような課題についてのルール抽出を試みた.

- 肝炎の線維化と血液検査データとの関係
- インターフェロン治療の有効性
- 肝炎 B 型と C 型を分けるための特徴

1. IF $(CHE > 176)$
THEN 線維化 $=$ $F1$ [サポート : 124, 確信度 : 53.1%]

評価: CHE が高い場合, 線維化が進まないというのは, 専門医の知識と一致する.

2. IF ((性別 $=$ M) AND (年齢 $=$ ５０代) AND $(ALP = +)$ AND $(LAP = +)$)
THEN インターフェロン効果 = 著効 [サポート : 11, 確信度 : 61.1%]

評価: 左辺にある ALP と LAP の両方は異常値であることは専門医にとっては興味深い. 多くの場合, は ALP と LAP の正常値に関するマイニング結果が得られる.

3. IF $(CHE \le 12.58)$
THEN 肝炎型 = B [サポート : 117, 確信度 : 87.2%]

評価: 肝炎 B 型の方が,C 型より低い CHE を持つのが一般的であり, 肝炎 B 型の方は C 型より進行していると言える.

図 2: 提案手法により得られたいくつかのルール

前処理では, 与えられた６つの表から, データクリーニング, 属性選択および新しい属性追加を行ない, 各患者の ID と検査日に基づく一つの表を作成しておく [2].

3.1 提案手法により得られたルール

提案手法により得られたルールのうち, 専門医に高く評価されたものを図 2 に示す.

4. まとめ

本研究では, 医学文献データベースを利用するデータマイニングアプローチを提案し, 各属性の文献データベースにおける重みを考慮した様々なヒューリスティック関数を提案した. 文献データベースから情報収集を行なうことにより, 各属性の重要度を事前に把握することができ, 前処理過程にも貢献できると考えられる. 実装したシステムは, 医療データで確認した結果, 興味深いパターンが得られ, 提案手法の有効性を示している.

参考文献

[1] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

[2] T. N. Tran, R. Ichise, and M. Numao. Mining hepatitis data set using information gathered from biomedical literature. In *Proc. of International Workshop on Active Mining (AM-2002), the IEEE International Conference on Data Mining (ICDM)*, pages 136–141, 2002.