Prime test を用いた知識獲得手法
A Method of Knowledge Discovery with Prime Test

†久本　賢　（岡山理科大学大学院）　††ラシキア・ジョージ　(岡山理科大学)

1．はじめに

知識獲得やデータマイニングでは、人間にとって理解しやすい結果が得られる手法が良いとされている。ニューラルネットワークのような数値計算的手法は、その結果の理解が非常に困難であり、あまり使用されない。近年では、結果の形のわかりやすさから論理ルールや決定木が用いられることが多い。これらの手法ではデータベース中から「if…then…」という形で結果を得ることができる。このような推論モデルは、人間の思考に良く似ていることからもよく使われている。しかし、従来の手法の多くでは、データベース中の属性全てを用いて知識の獲得を行っている。そのため、属性が多くなったり、データ中にノイズがあったりすると、正確な知識獲得ができないだけでなく得られる結果が複雑になる。

そこで本研究では、データベースから無駄のない、必要最低限の属性を選ぶ prime test[1]と言う概念を用いた新しい知識獲得アルゴリズム RGT（Rule Growing by Test）を提案する。そして、従来の知識獲得手法である ID3[3]及び CN2[4]との比較を行い、RGT の有効性を示す。

2．　Prime test について

2.1　Prime test

データベースに多数存在する属性の中からいくつかの属性を選び、元のデータベースの写像を表わすものを test と呼ぶ。この test は属性の数により多数存在するが、この中から無駄のないように属性の組み合わせを選ぶような test のことを prime test と呼ぶ。この prime test に含まれる属性のみを用いたデータベースは、元のデータベースよりもサイズが圧縮されながらも、情報を損なわないという特性を持つ。

本研究ではこの prime test をデータベースから求め、その prime test に含まれる属性についてのみ知識獲得に用いるようにする。Prime test の探索には PTD2.2 [2]を使用する

2.2　最適な prime test の探索

PTD を用いると、データベースによっては、複数の prime test が検出される場合がある。このような場合、知識獲得を行うための適切な prime test を選択しなければならない。

本研究では、検出された prime test に含まれる属性のみを使ったデータベースを符号化した際の最小符号長を、prime test の選択の基準に用いる。符号長が短いものは、データベースの持つ情報がより単純であるということになる。複数の prime test が検出された場合、どれを用いても元のデータベースと同じ情報を表わすことができるため、できるだけ圧縮データベースが単純なものを使うことが目的である。最小符号長を求めるにはエントロピーを用いて求める。[5]

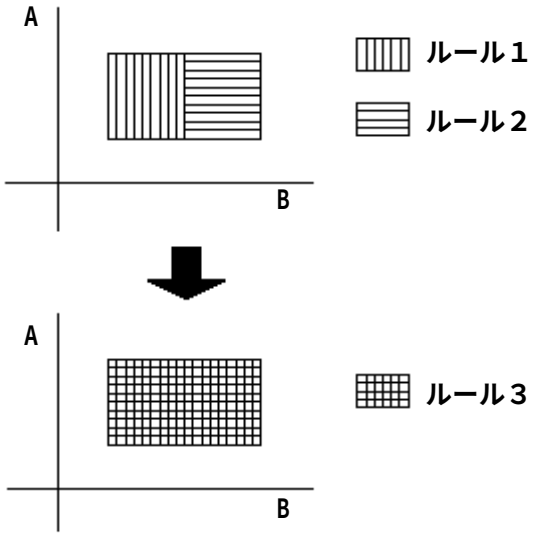

図１．ルールの拡張

3．RGT

従来の手法では、ルール中の条件は（属性=値）の形で表され、離散的な値しか扱うことができなかった。また、この形では値が限定されるので、一つのルールに対するデータのカバー数が少なく、データの個数が増えるとルールの個数も増え、より複雑になってくる。そこで、本研究では、この条件部分を（値－閾値≦属性≦値＋閾値）の形で表わす。この構成にすることで実数値を取り扱うことができるようになっている。また、ルール間で、重なっている範囲をまとめることで、一つのルールでカバーするデータの数が増えルールの数が減ることになる。結果、ルール全体が単純化されることになる。このときの閾値は PTD を用いて prime test を検出する際に同時に得ることができる。

本研究で提案する知識獲得手法 RGT では、まずデータベースから prime test に含まれる属性のみを用いて、一つ一つのデータにのみ対応するルールを

†Satoshi Hisamoto
Graduate School of Okayama University of Science, Okayama University of Science, Okayama, Japan

††George V. LASHKIA
Department of Information and Computer Engineering, Okayama University of Science, Okayama, Japan

	Monk1		Promoters		Breast Cancer	
	圧縮率	正確性	圧縮率	正確性	圧縮率	正確性
RGT	1.8%	100%	1.5%	100%	2.4%	100%
CN2	0.9%	79.0%	0.5%	80.0%	0.5%	91.6%
ID3	27.5%	98.4%	1.9%	100%	4.2%	100%

図２．獲得知識のデータベース再現の正確さ、及び圧縮率

	Monk1	Promoters	Breast Cancer
RGT	100%	72.2%	96.9%
CN2	75.0%	41.6%	91.2%
ID3	78.4%	66.6%	90.7%

図３．未知データに対する正答率

作成する。そこからルール間での範囲の重なりを考慮したルール拡張を行っていく。

図１ではルールの拡張について示している。ルール１とルール２について、属性 A では同じ範囲を示している。しかし、属性 B では違う範囲を示しているが、範囲が重なっている。そこで、この２つのルールを一つにまとめたルール３を作成する。このように、重なり合った部分を統合しルールを拡張していく。

これらの作業を可能な限り繰り返し、得られたものが結果となる。

４．実験及び比較結果

本研究で提案する知識獲得アルゴリズム RGT と、従来から用いられている手法である ID3、及び CN2 について比較実験を行う。これらの手法を用いて、いくつかのデータベースで知識獲得を行い、その結果について比較検討を行う。実験に用いるデータベースは、California 大学の機械学習データベースから Monk1、Promoters(Cleveland)、Breast Cancer（Wisconsin）の 3 種類を用いた。これらのデータベースの 2/3 のデータを知識獲得作業に用い、残りを未知データとして実験を行った。

図２は元のデータベースに対する獲得知識のデータベース再現の正確さとデータベースの圧縮率の結果を示した表である。圧縮率は以下の式で求めたものである。

$$(\text{圧縮率}) = \frac{(\text{ルールの総属性数})}{(\text{データベースの総属性数})} \times 100$$

データベースの圧縮率だけでは、RGT は CN2 ほどデータベースを圧縮できない結果になる。しかし、元のデータベース再現の正確さを合わせて考えると、CN2 は小さく圧縮できるがあまり正確ではなく、ID3 は正確ではあるが小さくできない場合がある。しかし、RGT では元のデータベースの情報をまったく損なわず、非常に小さく圧縮できることを示した。これは、RGT の特徴であり、RGT がデータベース圧縮の有効な手段であると言える。

次に、図３は知識獲得に用いたデータベースに無いデータ、いわゆる未知データに対するクラス分類の正答率を示している。ここでも RGT が従来手法 CN2、ID3 に比べ良い結果を示している。このことから、未知データに対する予測に関しても RGT が従来の手法よりも有効なものであると言える。

５．まとめ

本研究では、データベースからの知識獲得手法について、必要最低限の属性の組み合わせである prime test を求め、その属性を用いて知識獲得を行う新しい手法 RGT を提案した。

従来手法では全ての属性を用いて知識獲得を行うため、正確さに欠け、データベースの圧縮に向かないものもある。しかし RGT では、あらかじめ最適な属性の組み合わせを求めることで、データベースを小さく圧縮でき、さらに元のデータベースの情報を損なわない手法となっている。

また、データベースによって prime test が複数存在する場合がある。この場合には、prime test を用いて属性を限定したデータベースについて、最小符号長が最小の prime test を採用することにした。それにより、データベースがより簡単になり、得られる知識も簡単になる。

実験では、提案した手法 RGT と、従来手法 CN2、ID3 の比較を行った。知識獲得用データベースについては、RGT は他の手法に比べ、データベース情報を 100%保ちつつデータベースをある程度小さくできることを示した。また、未知データベースのクラス分類について、RGT はどのデータベースについても最高の結果を示し、最大では 20%近くの差を示したものもあった。総合的に見て、従来手法よりも RGT が有効な手法であることが示された。

参考文献

[1] G. Lashkia and S. Aleshin, “Test Feature Classifiers: Performance and Applications,” IEEE Trans. Syst., Man, Cybern. B, vol. 31, pp. 643-650, Aug. 2001.

[2] G. Lashkia and L. Anthony, “An Inductive Learning Method for Medical Diagnosis,” Pattern Recognition Letters, Elsevier, 24/1-3, 281-290, 2003.

[3] J. R. Quinlan, “Induction of decision trees”, Machine Learning, 1, 81-106, 1986.

[4] P. Clark and T. Niblett, “The CN2 induction algorithm”, Machine Learning, 3, 261-283, 1989.

[5] T. Mitchell, Machine learning, McGraw-Hill, 1997.