2J-5 データ特性による記憶に基づく推論と数量化 II 類との比較

毛利 隆夫* 田中 英彦
東京大学 工学部

1 はじめに

概念学習問題には，決定木を作成する方法や，人工ニューラルネット，MBR，多変量解析など様々な手法が適用されている．しかし，どのようなデータに対してどの手法が優位であるかを明らかにする研究は十分には行なわれていない．本研究では，MBR と多変量解析の一種の数量化 II 類との比較を，データの特性が既知である人工データにより行ない，高い正答率をあげるデータの特性が大きく異なっていることを示す．

2 MBR と数量化 II 類

MBR(Memory-Based Reasoning: 記憶に基づく推論) [2] は大量の事例の中から質問に類似した事例を検索し，類似した事例であれば回答も同じになるとの仮定に基づいて推論を行なう．MBR では属性の分類への貢献度が考慮されていないため，属性に重み値をつける研究がなされている．本研究では属性重み付け手法として，VDM, CCF, MIC, NN, QM2y を使用した (各手法については，[5] を参照されたい)．一方，数量化 II 類 [4](Quantification Method II, QM2 と略) は広く用いられている多変量解析の一種で，今回 MBR の比較対象として取り上げた．

3 人工データによる概念学習手法の比較

概念学習手法の比較は，これまでベンチマークデータに手法を適用して行なわれるのが一般的であった．しかし多数のベンチマークデータの中からどのデータを用いて試験すればよいかの指針はなく，ベンチマークデータの特性も十分には解析されていない．

そこで本研究では，データ特性が既知であるような人工データを合成し，それを用いてアルゴリズムの比較を行なう．この場合，合成される人工データは，現実世界のデータに類似した特性をもつことが要求される．というのは，我々が興味を持っているのは，現実世界から得られるようなデータに対する応用であるからである．そこで現実世界から得られた 11 種類のベンチマークデータ[1]([1] から取得) が共通してもつデータの特性を示すような人工データを合成する．

3.1 属性の型

一般に，人工データを合成する際のパラメータ数を多くすれば，もとのベンチマークデータと同じ性質をデータを合成するのは容易になる．その反面，個々のパラメータの意味が理解しづらくなり，設定も面倒になる．そこで本実験では，属性の型を定義し，各データの属性の型の分布から共通する傾向を抽出して，人工データを合成する際のパラメータを決定する．

属性の型 (attr_type) には，$0, 1, \ldots, 19$ までの 20 種類の雛型が用意されている．この雛型は，あるクラスのもとでの，一つの属性値の出現確率 P_{top} と，それ以外の属性値の出現確率 P_{bottom} とで構成されている．P_{top} と P_{bottom} とは式 1,2のような関係にある．取り得る値が 3 種類で，attr_type=0 および 10 の場合の値の出現確率の様子を図 1に示す．各属性にはクラス毎に，各型の雛型との差分が最も小さな型が割り当てられる．

$$\frac{P_{top}}{P_{bottom}} = \frac{0.975 - 0.05 \times attr\_type}{0.025 + 0.05 \times attr\_type} \quad (1)$$

$$P_{top} + P_{bottom} \times (N_v - 1) = 1.0 \quad (2)$$

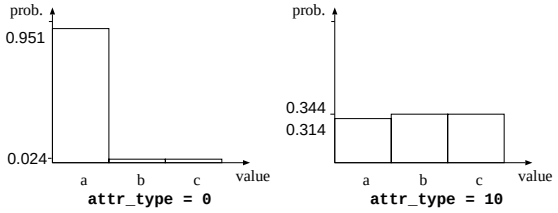

図 1: 属性の型 attr_type の雛型

属性の型が 0 に近い属性は，あるクラスを特定すると，一つの値の出現頻度が他と比べて非常に高い場合であり，このような属性は分類に大きく役立つといえる．また，属性の型が 10 に近い属性は，どの値の出現頻度も大差無い値であり，分類には役に立たない属性である．属性の型が 19 に近い属性は，一つの値の出現頻度が低く，他の値の出現頻度が高いため，これも分類にはあまり役に立たない．

3.2 属性の型の傾向の抽象化

11 種類のベンチマークデータの属性の型の傾向を調べると，おおまかにいって次のような 3 種類の傾向があることがわかった (図 2)．

edge1 attr_type が 0 の付近が多く，他の型は少ない

edge1+edge2 0 と 19 の両端の頻度が高い

edge1+edge2+noise 0, 10, 19 付近での頻度が高い

そこで，人工データを合成する際のパラメータを，表 1のように定義した．

ここで，一つの属性にはクラス毎に属性の型が割り振られているが，それぞれの P_{top} が同じ属性値である割合をピークが同じ (same_peak) であるとした．same_peak に関しても属性の型と同様にベンチマークデータでの傾向を調べたが，ピークが同じになる確率が 0%付近，50%付近，100%付近の 3 種類のデータに分類できたので，それをパラメータ値とした．

なお表 1中の属性依存度は，属性間の相互情報量をもとに計算される値で，一旦データを合成した後にデータを書き換えることで制御される．我々はすでに，人工データ合成の際には，属性間の依存度が重要なパラメータであることを明らかにしている [6]．

---

*日本学術振興会特別研究員

[0]Comparison between Memory-Based Reasoning and the Quantification Method II by Characteristic of Data, Takao MOHRI and Hidehiko TANAKA, Faculty of Engineering, The University of Tokyo, {mohri,tanaka}@MTL.T.u-tokyo.ac.jp

[1]iris, segment, wine, breast, diabetes, liver, vote, soybaen, crx, hypo, hepatitis の 11 種類

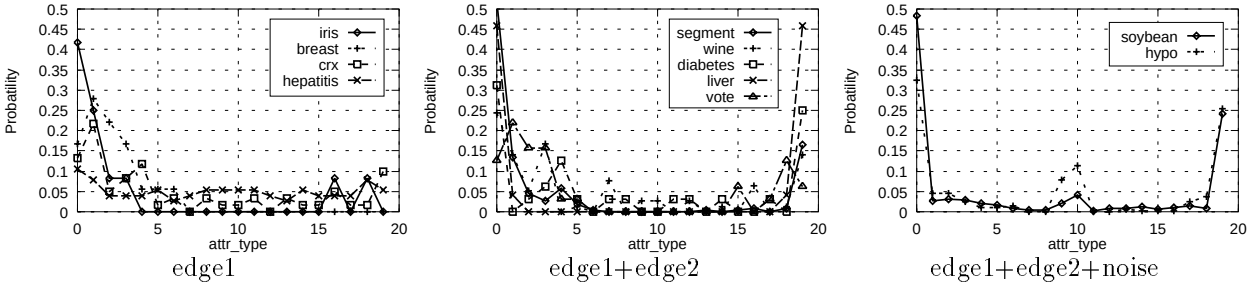

図 2: 属性の型の傾向

表 1: 人工データ合成のために指定するパラメータ

パラメータ名	個数	意味	値
N_a	1	属性数	8,16
N_c	1	クラス数	2,8
N_d	1	事例数	100,300
$N_v(a)$	1	属性の取り得る値の数	2,8
class_ratio	1	クラスの比	8:2, 5:5
attr_edge1	1	役に立つ属性が多い	無, 有
attr_edge2	1	ノイズ属性が多い	無, 有
attr_noise	1	ノイズ属性が多い	無, 有
same_peak	1	ピークが同じ確率 [%]	0,50,100
dependence	1	属性依存度	低, 高

4 実験

表 1のパラメータを変化させて人工的にデータを作成し，そのデータを用いて MBR の属性重み付け手法および数量化 II 類を比較する実験を行なった．

実験では，表 1のパラメータの全ての組合せを試験したので，$2^9 \times 3 = 1536$ 種類のデータが作成されテストされた．正答率のテストには，50 回繰り返しの e0 bootstrap 法 [3] を用いた．各データに対して，最高の正答率もしくは，それと同等とみなせる正答率が得られた場合に，そのアルゴリズムを良いアルゴリズムであるとした．正答率が同等であるかどうかの判断は，正答率の分布が正規分布であると仮定して，平均値の同一性検定を用いた．

表 2: 最高 (またはそれと同等) の正答率が得られた割合

順位	1	2	3	4	5	6
手法	VDM	MIC	NN	CCF	**QM2**	QM2y
割合	0.674	0.617	0.530	0.434	**0.328**	0.251

表 3: 手法が単独で最高 (またはそれと同等) の正答率を得る場合

手法	**QM2**	VDM	MIC	NN	CCF	QM2y
頻度	**154**	135	38	30	15	2
比率	**0.100**	0.088	0.025	0.020	0.010	0.001

表 2に，1536 種類のデータのうち，どれだけ割合で良い正答率が得られたかを示す．VDM, MIC などの属性重み付け方法を用いた MBR では，データ空間中の 6 割以上の点で良い正答率が得られているのに対し，QM2 は 30%程度の点でしか，良い結果が得られていない．

次に，手法間で良い正答率を挙げるデータの違い，つまり手法間の傾向の違いを調べてみた．表 3に，その手法のみが単独で良い正答率を得ていた場合の比率を示す．QM2 は良い正答率を挙げるデータは少ないものの，単独で高い正答率を得る場合が多く，他の MBR の手法と比較して異なる傾向のデータで良い正答率を得ていることが分かる．

MBR と数量化 II 類の得意とするデータの傾向は，特に，属性数，クラス比，データ数，属性依存度を変化させた場合に大きく変化した．そのうち 2 種類を図 3に示す．おおまかな傾向としては，MBR は数量化 II 類と比べて，分類しづらいようなデータの際により有効であるということができるだろう．

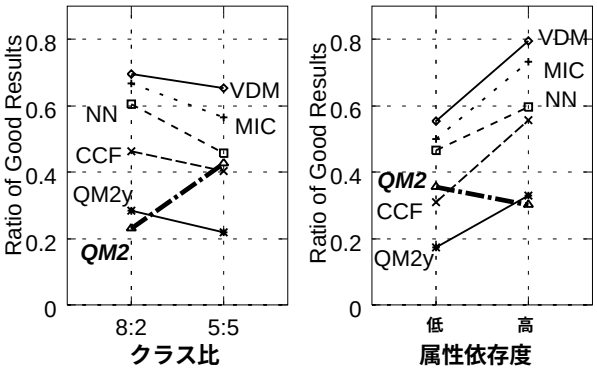

図 3: パラメータ毎のアルゴリズムの振舞いの差

5 おわりに

本研究では，人工的に合成したデータを用いて，MBR と数量化 II 類の特性を比較した．その結果，両者が得意とするデータの傾向は大きく異なることが分かった．

なお，本研究は文部省科学研究費補助金 (特別研究員奨励費, No.06004134) の援助を受けている．

参考文献

[1] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases. Irvine, CA: University of California, ftp://ics.uci.edu/pub/machine-learning-databases. 1995.

[2] Craig Stanfill and David Waltz. Toward memory-based reasoning. *Communications of the ACM*, Vol. 29, No. 12, pp. 1213–1228, December 1986.

[3] Sholom M. Weiss and Casimir A. Kulikowski. *Computer Systems That Learn*. Morgan Kaufmann, 1991.

[4] 林知己夫. 数量化 –理論と方法–. 朝倉書店, 1993.

[5] 毛利隆夫, 田中英彦. 最適性をもつ連続量・離散量両用の事例の属性の重み付け方法. 人工知能学会全国大会 (第 8 回) 予稿集, pp. 111–114, 1994.

[6] 毛利隆夫, 田中英彦. 人工データを用いた MBR の属性重み付け手法の評価. 第 50 回情報処理学会全国大会講演論文集 (2), pp. 141–142, 1995.