記憶に基づく推論と多変量解析との比較

Comparison between Memory-Based Reasoning and Multivariate Analysis

毛利 隆夫　　田中 英彦
Takao Mohri　　Hidehiko Tanaka
{mohri,tanaka}@MTL.T.u-tokyo.ac.jp
東京大学 工学部
Faculty of Engineering, The University of Tokyo

We compare two classification methods, Memory-Based Reasoning(MBR) and the Quantification Method II(QM2) which is one of the statistical multivariate analysis methods. Since popular evaluation through benchmark data sets has some serious drawbacks, we propose the way to compose artificial data sets which have similar characteristics to the real data sets, and use them to evaluate these classification methods. Experimental results show that the characteristics of data that produce higher accuracy for MBR are different from that for QM2.

1　はじめに

クラスが既知の事例の参考にして，クラスが未知の事例をクラスに分類する問題は，機械学習の分野では概念学習問題とよばれており，基本的な学習の問題である．また物事を分類することは判断の基本であり，応用範囲も広く盛んに研究されている．このような分類問題には，決定木を用いる方法や人工ニューラルネットを用いる方法，統計的な多変量解析や記憶に基づく推論など様々な手法が適用されている．これらの手法は同種の問題を対象としているため，手法間の比較研究も盛んに行なわれている．

本研究は，このような分類手法のうち記憶に基づく推論と，多変量解析の一種である数量化 II 類を実験的に比較する．従来よく行なわれてきたベンチマークデータを用いた比較にはいくつかの大きな欠点があるため，本研究では，自然なデータを模擬した人工データに対する分類実験によって比較する．実験結果を通して，記憶に基づく推論と数量化 II 類とでは，高い正答率が得られるデータの特性が大きく異なっていることを明らかにする．

2　分類問題とその手法

2.1　分類問題

分類問題では，所属するクラスが既知である事例 (訓練事例) が学習用に，どのクラスに属するかが不明の事例 (テスト事例) がテスト用に，それぞれ用意される．事例は問題部と回答部からなり，問題部は幾つかの属性から構成される．回答部はその事例が属するクラスである．

例えば，本研究で用いた vote というベンチマークデータ (図 1) には，アメリカ合衆国の 1984 年の下院での 16 の議案に対する投票結果がおさめられている．ここでは一人の議員が一つの事例に相当し，事例の問題部の属性は 16 のそれぞれの議案に対する投票結果 (賛成，反対，棄権) を表し，回答部はその議員の属する政党 (共和党か民主党) を表している．ここで解くべき問題は，回答が未知の事例を二つの政党に分類すること，つまり所属政党が未知の議員の投票結果を見て，その議員の所属が共和党か民主党かを予測することである．

このような分類問題には様々な手法が適用できる．決定木を用いる方法では，事例を分類する方法を示す規則を訓練事例から作成する．人工ニューラルネットを用いる方法では，訓練事例がうまく分類できるようにネットワークの結合の重み値を変化させる．本研究で対象とする記憶に基づく推論や数量化 II 類も，同種の分類問題を対象としている．

2.2　記憶に基づく推論

記憶に基づく推論 (Memory-Based Reasoning: MBR) は，大量の事例の中から質問に類似した事例を探索し，類似している問題であれば答えは同じなるとの仮定のもとに推論を行なう．つまり，訓練事例の中から，質問の事例に最も類似した問題部分をもつ事例を探索し，その事例の所属するクラスを，質問事例のクラスとしてそのまま回答する．図 1に MBR の仕組みを vote ベンチマークデータの一部と併せて示す．

MBR はルールを用いないため知識獲得が容易で，システムの構築が短期間で行なえるなどの特徴をもち，こ

れまでに英単語発音問題 [SW86]，機械翻訳 [SOF+93] などに応用され，成果を挙げている．

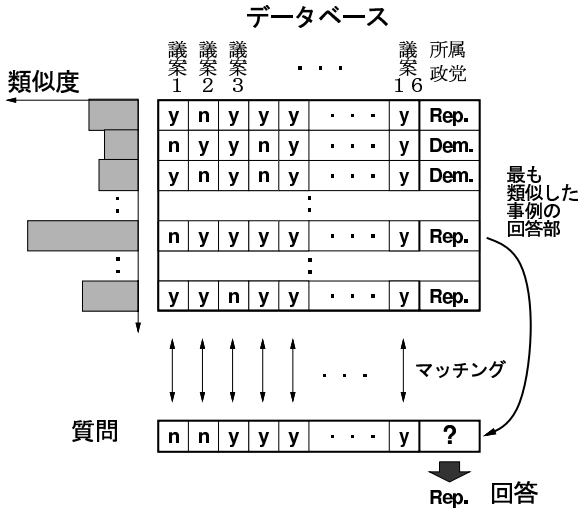

図 1: MBR の仕組みと vote ベンチマークデータ

MBR では事例間の類似度の計算方法，具体的には事例の属性の重み値の計算方法が，正答率に大きく影響する．これまで多数の属性重み付け手法が研究されているが，本研究では，MBR の属性重み付け方法として基本的な，VDM(Value Difference Metric)[SW86] を使用した．

2.3 数量化 II 類

数量化 II 類 (the Quantification Method II: QM2 と略) [林 93] は広く用いられている多変量解析手法の一種であり，やはり分類問題を対象としている．数量化 II 類では，事例の属性値を線形結合した一次式を作成して，もとの属性を複数の新しい属性に線形変換する．新しい属性は互いに独立であり，なおかつクラス毎の分散と全体の分散の比を最大にするように作成される．直観的に説明すると，事例の属するクラスが同じ場合に新しい属性値が近い値になるように，線形式が作成される．

本研究において数量化 II 類で回答を予測する際には，線形判別分析を用いることにする．すなわち，変換後の新しい属性値がそれぞれのクラス毎の平均と比較され，最も質問に近いクラスが回答される．なお数量化 II 類では新しい属性は元の属性と同じ数だけ生成されるが，必ずしもそのすべてを利用する必要はない．本研究では，累積寄与率が 99% に達するまでの新しい属性を使用している．

数量化 II 類は多変量解析の一種であるが，同じ多変量解析に属する主成分分析も良く用いられる手法であり，同種の分類問題に適用することができる．主成分分析も数量化 II 類と同様に，古い属性を互いに独立な新しい属性に変換するのであるが，その際の規準が異なる．主成分分析では新しい属性の分散が最大になるように線形式が計算されるが，その際には，事例の属するクラスの情報は用いられない．予備実験の結果 [毛利 95]，数量化 II 類の方が主成分分析よりもよい分類結果をもたらすことが多かったため，本研究では数量化 II 類を代表的な統計的手法として取り上げ，記憶に基づく推論と実験的に比較している．

3 分類手法の評価方法

3.1 ベンチマークデータによる評価の得失

一般に，MBR，数量化 II 類などの分類アルゴリズムには，それぞれ得意なデータと不得意なデータがあることが経験的に知られており，アルゴリズムのバイアスと呼ばれている．また Schaffer [Sch94] は，事例空間上での事例の分布を仮定した際に，その事例の属するクラスの値を全ての組合せを試した場合，任意の分類アルゴリズムの正答率の平均は，50% (クラス数が 2 の場合) になることを証明し，これを 「学習に関する保存則 (Conservation Law)」と名付けた．この法則は，どんなデータに対しても良い振舞いを示す汎用なアルゴリズムは存在しないこと意味している．このように，あらゆるデータに対して良い結果を残すアルゴリズムが存在しない以上，あるアルゴリズムがデータ空間のどの範囲で良い振舞いを示し，どの部分は不得手なのかを明らかにすることが本質的である．

これまでの分類手法の研究の多くでは，ベンチマークデータを用いた評価が行なわれており，その利点は，つぎのようにまとめられる．

- ○ 広く流通しており他の実験結果との比較が容易
- ○ 多くのベンチマークデータは，現実世界での事象からデータを採取しており，恣意性が少ない

ベンチマークテストによる学習手法の評価は広く行なわれているのだが，その反面，次のような問題点が指摘できる．

- × どの/いくつのデータを使えばいいのか不明
- × データの特性が明らかにされていない
- × アルゴリズムによる正答率の差が説明しづらい
- × 十分な数のベンチマークデータがそろっていない

これらの問題点を解決するため，人工的にデータを合成し，それらを利用してアルゴリズムの評価を行なう方法が考えられる．

3.2 人工データによる評価の得失

ベンチマークデータを用いずに，人工的に合成したデータによりアルゴリズムの評価を行なう場合の利点は，合成時のパラメータを操作することによって，特性が既知であるデータを，必要な数だけ合成できる点にある．したがって，どのパラメータがアルゴリズムの振舞いに影響しているかを，実験的に知ることができる．データの特性とアルゴリズムの性能との関係が明らかになれば，特性は既知のデータに対して，どのアルゴリズムが優位であるかを予測できるようになり，現実的にも有用であることが期待できる．

3.3 現実世界を対象としたデータ空間の限定

しかし，人工データによる評価を行なう場合には，どのようなデータを合成するかに注意を払う必要がある．

第1の注意点は，対象とするデータ空間中の領域から外れたデータを作成しないようすることである．Conservation Lawにより，全データ空間に対して有効であるアルゴリズムは存在しないことが証明されているため，対象とする事例集合の空間を限定し，その範囲内で特定のアルゴリズムの有効性を示す必要がある．

ところで我々は通常，現実世界から得られるような事例集合を対象にしている．現実世界から大きくかけ離れた性質をもつようなデータのみを合成し，その結果特定のアルゴリズムの優位性が言えたとしても，実際の問題にそのアルゴリズムを適用した場合の振舞いを予測する指標としては，なんら役に立たないかもしれない．そこで本研究では，「現実世界から得られるデータに類似したデータ」のみを対象とする．

ここで問題になるのが，「現実世界」をどう定義するかである．現実世界は，曖昧な概念であるため，現実世界の範囲を定義したり，そこから得られる事例集合の特性を厳密に特定することは非常に困難である．そこで本研究では，カリフォルニア大学の機械学習データベース [MA95] から取得した11種類のデータ (表1) を，現実世界の基準として扱う．すなわち，これらの現実世界から得られたベンチマークデータに類似しているデータを対象とする．

第2の注意点は，データ合成時に用いるパラメータの数についてである．指定すべきパラメータが多い場合には，細かい指定が可能である反面，個々のパラメータの意味が人間にとって理解しづらく，説明性に乏しい．また，パラメータが多数ある場合，パラメータの組合せによるデータの種類が膨大になるため，すべてのデータを試験するのは困難である．したがって，パラメータは人間に理解しやすく，少数であることが好ましい．

表 1: ベンチマークデータ

名称	クラス	属性	事例	問題領域
iris	3	4	150	アイリス
segment	7	16	210	風景画像
wine	3	13	154	ぶどうの種類
breast	2	9	699	乳癌
diabetes	2	8	768	糖尿病
liver	2	6	345	肝臓障害
vote	2	16	300	議員の所属政党
soybean	19	35	683	大豆のかかる病気
crx	2	15	490	クレジットカード
hypo	5	36	2514	甲状腺機能低下症
hepatitis	2	19	115	肝炎

4 関連研究

人工データによる分類手法間の比較としては，Ahaの研究 [Aha92] や Rendell らの研究 [RC90] がある．しかし，人工データを合成する際のパラメータの選択に関しては十分な議論がなされておらず，合成された人工データが対象領域に含まれているか，また偏ったデータになっていないかについては疑問が残る．

これまでに，データの特性とそのデータに適するアルゴリズムとの関係を解析する研究としては，その関係を人手によりルール化する研究 [Bro93] や，そのようなルールを約20種類のベンチマークデータのテストによる結果から学習する研究 [BGH94] などが行なわれている．これらの研究に共通して言えるのは，一般に流通しているベンチマークデータを利用しているためにデータ数が少なく，さまざまな特性のデータを試験できていない点である．

5 パラメータを決定するための実験

5.1 実験方針

パラメータの決定は，次のような2段階で行なう．まず第1段階として，現実世界のベンチマークデータを再現できる程度に詳細なパラメータの集合を特定する．これには，

1. ベンチマークデータからパラメータ値を取得し，

2. そのパラメータ値を用いてデータを合成し，

3. もとのデータと合成したデータが類似した特性を示すことを検証する

という手順でパラメータ集合の評価を行なう．ここで得られるパラメータは，現実世界のデータに類似した人工データを作成するために必要なパラメータである．

ただしこのようなパラメータは数が多すぎるために，そのままでは人工データ合成には不向きである．そこで第 2 段階として，次にパラメータに共通の傾向を見つけて抽象化を行ない，パラメータ数を減少させる．

5.2 十分詳細なパラメータの選定

まず我々は，表 2のようなパラメータ level 1 から出発した．このパラメータの特徴は，個々の属性 a の値 vの出現確率を，事例の属するクラス c ごとに，条件つき確率 $p(v|c,a)$ として指定している点である．これはかなり詳細なパラメータであるということができる．

表 2: 人工データ合成のためのパラメータ (level 1)

パラメータ名	個数	意味
N_a	1	属性数
N_c	1	クラス数
N_d	1	事例数
$N_v(a)$	N_a	属性の取り得る値の数
p_c	N_c	クラスの比
$p(v\|c,a)$	$\sum_a N_c \times N_v(a)$	属性 a, クラス c のもとでのその属性の値 v の条件付き確率

なお，$p(v|c,a)$ は属性ごとに与えられる値であり，複数の属性が同時に指定された値をとる確率は与えていない．例えば，事例のクラスが c である場合に，属性 $a1$ が値 $v1$ をとり，属性 $a2$ が値 $v2$ を同時にとる確率 $p((v1|a1,c)\wedge(v2|a2,c))$ は与えていない．これは，このような属性の組合せを考えると，指定すべきパラメータが膨大になり，あとで行なうようなパラメータの共通の性質を見つけるような作業が困難になるからである．

元の vote ベンチマークデータとパラメータ level 1 で合成したデータでの MBR, QM2 の正答率を図 2に示す．このように，パラメータ level 1 では，もとのデータでの場合に比べてはるかに高い正答率が得られていることが分かる．これは，$p(v|c,a)$ が属性毎に与えられるパラメータであるので，各属性が独立にデータ合成されるためであると思われる．このままでは人工データは元のデータと類似しているとは言い難い．実際には各属性は互いに密に依存しあっていることが予想されるため，属性を互いに依存させる手続きが必要になる．

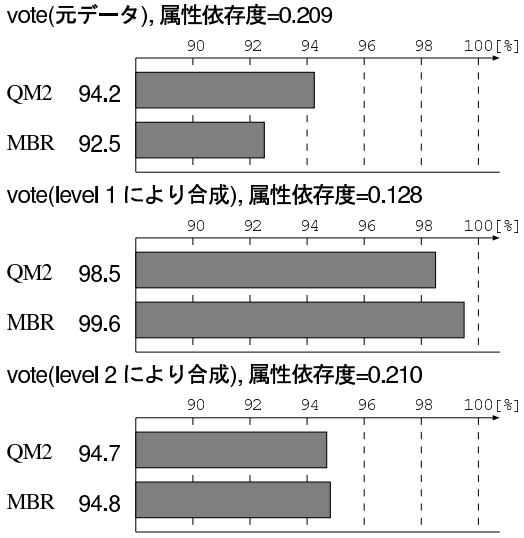

図 2: 元データと合成データ (level 1,2) の比較

5.3 書き換え操作

本研究では互いに依存した属性を持つデータを作成するために，各属性を一旦独立に作成しておき，その後同じ属性内で属性値の順番を入れ換える 書き換え」操作を新たに考案した．書き換え操作の基本的な考え方を図 3で説明する．まず各属性を独立に作成して，図 3左上のような事例ベースができたとする．ここで属性 a_1の取り得る値 p,q,r と属性 a_2の取り得る値 x,y,z との組合せの出現頻度は，図 3左下の表のようであったとしよう．

属性 a_1と a_2とを依存させることを考える．例えば，a_1が p であり，a_2が x である場合や (以下，$(a_1,a_2)=(p,x)$ と表記)，$(a_1,a_2)=(r,z)$ である場合の頻度は高いので，これを更に高くすれば両属性の依存度は高まる．そこで，同じクラスに属し，$(a_1,a_2)=(p,z)$ と $(a_1,a_2)=(r,x)$ とである事例をデータベース中から探し，2 事例の a_2での値を交換する．そうすると事例ベースは図 3右上，属性値の組合せの出現頻度は図 3右下のようになる．2 つの属性の間の依存度は，相互情報量をつかって計算できるが，書き換え前の属性 a_1と a_2との相互情報量が 0.101 bits であるのに対し，書き換え後には，0.197 bits に増加し，確かに両属性の依存度は深まったことがわかる．なお，このように 2 事例の値を交換しているのは，パラメータ level 1 での制約条件 $p(v|c,a)$ を変更しないためである．

パラメータ level 1 によって合成データが元データと大きく異なった理由は，属性間の依存性を考慮せずに

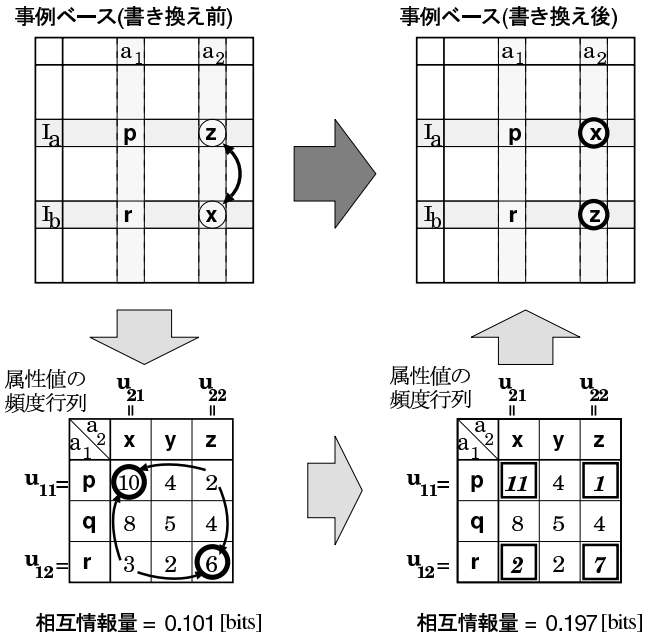

図 3: 書き換え例

データを作成したためであると考えられる．そこで，新たに 属性依存度」を新たにパラメータとして導入する．属性依存度は，個々の属性間の相互情報量の全属性の組での平均値を，取り得る最大の依存度で正規化したものである．0 から 1 までの値をとり，値が大きい程属性間が依存している．パラメータ level 1 に，属性依存度を追加したものを，パラメータ level 2 と定義する．

書き換えは，1 回だけでは，2 つの事例のそれぞれ一箇所の属性値が変更されるだけである．これだけでは高い属性依存度は得られないので，必要な属性依存度が得られるまで，書き換えは繰り返し行なわれる．

図 2には，vote ベンチマークから得られたパラメータをもとに，属性依存度も含めて元のデータと同じになるように合成したデータでの，MBR と QM2 での正答率が示されている．これから分かるとおり，パラメータ level2 によるデータの再現性は，level 1 の場合よりも向上していることが分かる．

5.4 パラメータの抽象化

level 2 のパラメータによって合成されたデータは，元データとほぼ同じ特性を示すことがわかった．しかし，level 2 での属性値の確率 $p(v|c,a)$ は詳細すぎて，そのままでは人工データを合成する際に指定づらいため，パラメータの抽象化を行なう必要がある．そこで level 2 の $p(v|c,a)$ を， attr_type と same_peak という 2 種類のパラメータに置き換え，パラメータ level 3 を作成した．

1. attr_type

   attr_type は 20 段階の属性の型であり，各属性の各クラス毎の取り得る値の頻度により決定される．attr_type は，属性値を 2 種類，すなわち最も高頻度 もしくは低頻度）である属性値とそれ以外に分類した後，20 種類の離型のうち最も近い型に決定される．図 4に $N_v = 3$, attr_type=0,5,10,15 の場合の attr_type の離型を示す．

2. same_peak

   ある属性で各クラスの attr_type が同じでも，最も高頻度 または低頻度）の属性値が，同じ属性内の各クラスで同じか異なっているかによって，その属性がクラスの決定に役立つかどうかが大きく異なってくる．ここでは，半分以上のクラスで同じ属性値が高頻度 もしくは低頻度）であった場合には same_peak=1 とし，それ以外の場合を same_peak=0 と定義する．

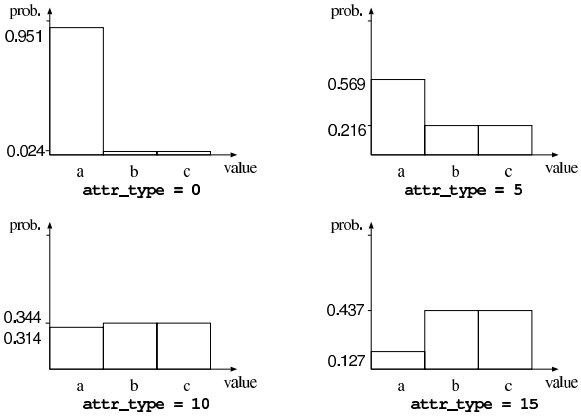

図 4: attr_type の離型

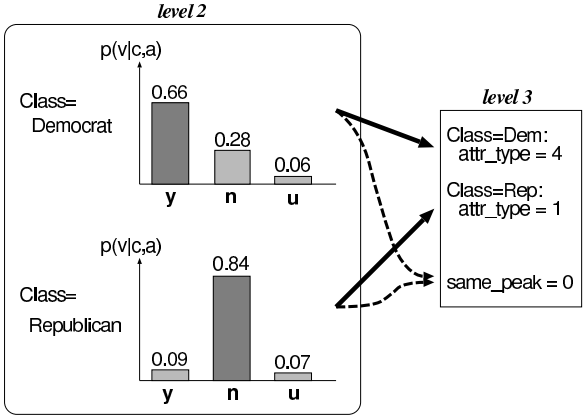

図 5: attr_type と same_peak の例

vote ベンチマークデータの ある属性を例にとり，attr_type と same_peak を決定する様子を図 5に示す．

Class が Democrat の場合には，属性値'y' が，Class が Republican の場合には属性値 'n' が高頻度な属性値になる．両者の高頻度な属性値は異なるため，same_peak= 0 である．Class が Republican の場合は，Democrat の場合よりも頻度差がはっきりしているため，attr_type は 1 に分類されている．

level 2 または level 3 で合成したデータと元データとの最高正答率の差を，11 種類全てのベンチマークデータについて図6に示す．多くのベンチマークデータでは level 3 のほうが level 2 よりも最高正答率の差が大きく，多少再現性が悪くなっていることがわかる．

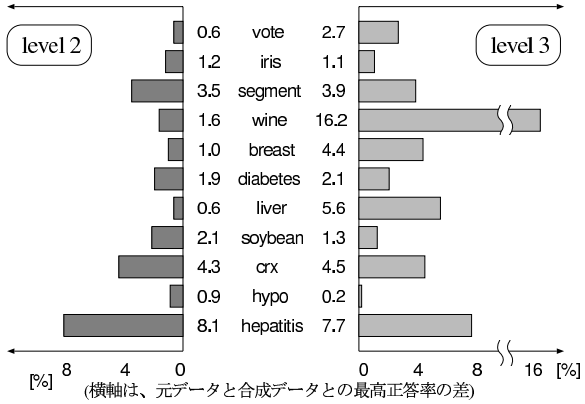

図 6: ベンチマークデータの再現性

5.5 データ空間探索のためのパラメータの抽象化

パラメータ level 3 は，データの再現性を大きく損なうことなく，level 2 のパラメータ数を減少させることに成功した．実際，vote での パラメータ level 2 の 自由度は 85 であったが level 3 では 69 自由度までに減少している．しかしまだ対象とする空間を広く探索するためには，69 次元のパラメータは多すぎて扱いづらく，さらにパラメータの抽象化を行う必要がある．

そこで，vote を含む 11 種類のベンチマークデータ (表 1) から抽出したパラメータ level 3 に対して，共通した傾向を解析することにした．解析の結果，attr_type に関する傾向を図 7 に示すような 3 種類に，same_peak の傾向を表 3に示すような 3 種類に，それぞれ分割した．

これらの解析結果から，パラメータ level 4 を定義する (表 4)．level 4 でのパラメータの自由度は 11 次元であり，level 2, 3 とくらべて大幅に簡略化されている．level 4 での個々のパラメータの意味は次の通りである: class_ratio は最頻クラスと他のクラスとの頻度の差である．一つのクラスをのぞき，ほかのクラスの出現頻度は同じとした．attr_edge1, attr_edge2, attr_noise は，attr_type がそれぞれ 0 付近，19 付近，10 付近である属性が多いことを示している．dependence は，属性依存度が書き換え (最高 5000 回) により高められたかどうかを示している．

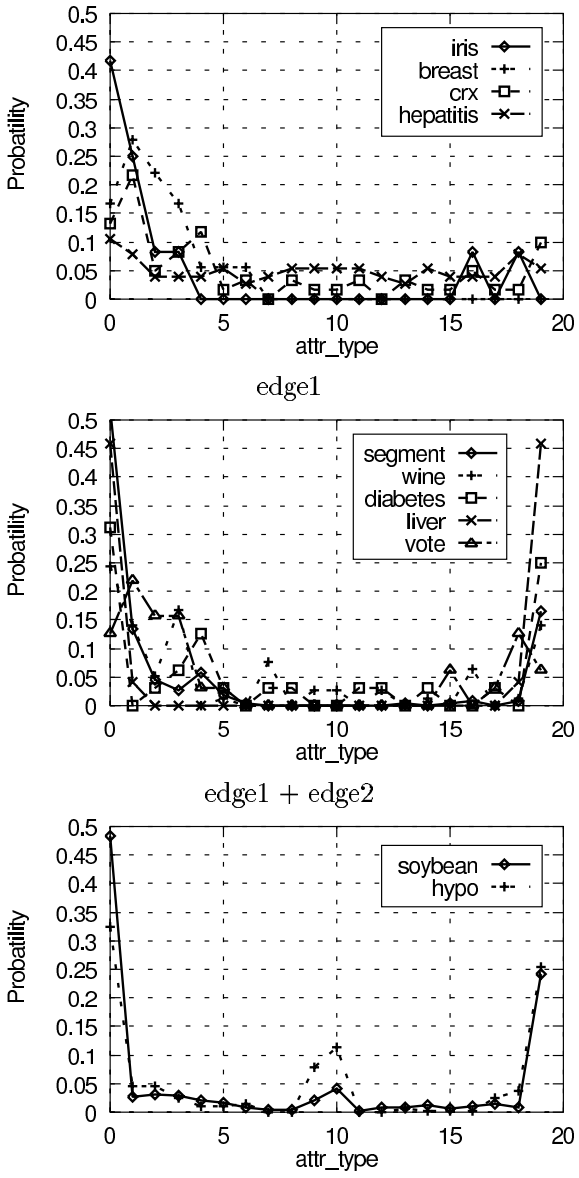

図 7: attr_type の傾向の分類

表 3: same_peak によるベンチマークデータの分類

データ	# same_peak / 属性数	比率	傾向
breast	1 / 9	0.111	0.0 付近
vote	3 / 16	0.188	
liver	2 / 6	0.333	
iris	2 / 4	0.500	0.5 付近
segment	8 / 16	0.500	
hepatitis	12 / 19	0.632	
crx	11 / 15	0.733	1.0 付近
soybean	30 / 35	0.857	
hypo	27 / 29	0.931	
wine	13 / 13	1.000	
diabetes	8 / 8	1.000	

表 4: 人工データ合成のためのパラメータ (level 4)

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

6 人工データによる分類実験

level 4 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多変量解析を比較する実験を行なった．乱数の種を 3 通りに変化させて，level 4 のパラメータの全ての組合せを試験したので，$(2^9 \times 3) \times 3 = 4608$ 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 [WK91] 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を 良い正答率」と呼び，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，信頼度 95%の平均値の同一性検定を用いた．

6.1 実験結果

この 4608 種類のデータ全体に対する分類実験の結果，良い正答率が得られた割合を図 8に示す．MBR の方が QM2 よりも多くの場合に良い正答率を挙げていることがわかる．また，MBR のみ，もしくは QM2 のみが高い正答率を挙げている場合があり，両手法の得意とするデータが異なっていることもわかる．

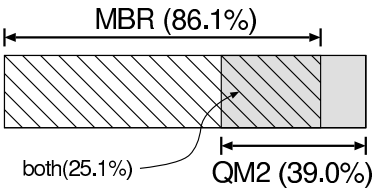

図 8: 分類実験全体で良い正答率が得られた割合

また，ある一つのパラメータだけに着目して集計した結果，パラメータ値の変化によって高い正答率を挙げる比率が大きく上下するものがみられ，特に属性数，事例数，取り得る属性値の数，クラス比，属性依存度で変化が激しかった．これら 5 種類のパラメータごとの集計結果を図 9に示す．この結果より，MBR, QM2 の両手法がどのようなデータを得意とするか，おおまかな傾向が理解できる．

次に，決定木を生成するプログラム C4.5[Qui93] を用いて，QM2 が MBR よりも高い正答率を挙げる可能性の高いデータの特性を解析した．実験結果より，QM2 が良い正答率を挙げるかどうかをクラスとし，データ合成のパラメータを属性とする事例集合を作成し，QM2 が高い正答率を挙げる条件をルール形式で求めた．おおまかな傾向をつかむために，決定木の枝の事例数 (-m option) を 100 以上として，生成される木を複雑にならないように限定している．

表 5に，QM2 が MBR よりも高い確率で良い正答率を得る場合の条件をいくつか示す．このように，決定木を生成することで，QM2 が MBR よりも良い場合の明示的な切り分けに成功した．

表 5: QM2 が MBR よりも良い正答率を多く挙げる場合

	QM2 が高い正答率を得る条件	実際に高い正答率を得た割合 [%]	
		MBR	QM2
R1	取り得る属性値の数=2, 属性数=8, クラス比=8:2, 属性依存度=低	46.5	85.4
R2	取り得る属性値の数=2, 属性数=8, 事例数=300, クラス比=5:5	56.2	83.0

7 考察

本研究では，まず 現実世界」のデータに類似したデータを作成することを試みた．そのために，パラメータ level 2, 3 では，元になったベンチマークデータからパラメータを抽出し，そのパラメータにより人工データを合成し，両データでの最高正答率の差によって，データの再現性を測定した．しかし，どの程度の再現性があれば十分であるかの議論は行なっておらず，今後の課題として残されている．

また，今回の実験では，パラメータ値の全ての組合せを試験したために少数のパラメータに限定する必要があり，パラメータの強引な抽象化を行なったが，パラメータが level 3 のように多数のままでも，モンテカルロ ・シミュレーションのようにパラメータ値をランダムに決定して，多数のデータをサンプルして試験するこ

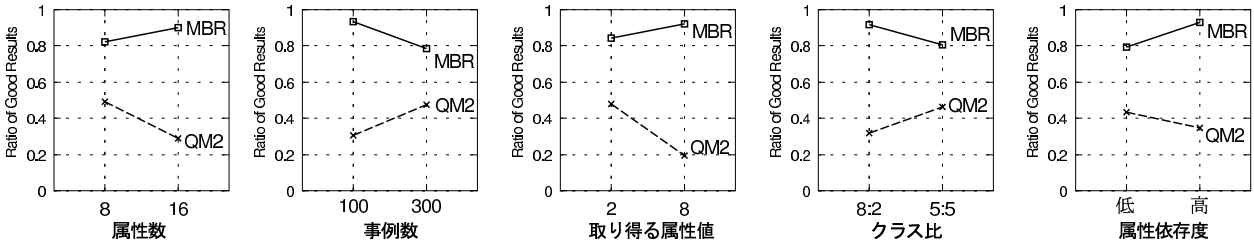

図 9: 特定のパラメータのみを変化させた場合の，高い正答率を挙げる割合の変化

とは可能である．また，他のパラメータを固定しつつ，一種類のパラメータのみを変化させた実験を行なう方法も考えられる．

なお，今回の実験では属性値は離散値のみに限定している点に注意すべきである．連続値を持つ属性は，あらかじめ離散化 [FI93] を行なった．MBR で用いた属性重み付け手法 VDM は離散値にしか使えないが，QM2 は連続値・離散値両方に適用可能であり，QM2 は離散化を行なわずに分類を行なった法が，離散化後よりも高い正答率が得られている [毛利 95]．本研究を連続値属性に拡張することも今後の課題である．

8 おわりに

本研究では，人工的に合成したデータを用いて記憶に基づく推論 (MBR) と，多変量解析の一種である数量化 II 類を比較した．人工データ合成に用いるパラメータに必要な性質を述べ，ベンチマークデータを用いて，現実世界のデータに類似した人工データを作成する方法を提案した．データ合成のため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，属性依存度を制御する書き換え操作を新たに考案した．合成した人工データを使用して MBR と数量化 II 類を比較した結果，属性数，事例数などのいくつかのパラメータでは，良い正答率が得られる割合の変化が， MBR と数量化 II 類では大きく異なることが判明した．また，数量化 II 類が MBR よりも有利なデータの特性をルールで表現することに成功した．

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

[Aha92] David W. Aha. Generalizing from case studies: A case study. In *Proceedings of the Ninth International Machine Learning Workshop(ML92)*, pp. 1–10, 1992.

[BGH94] Pavel Brazdil, João Gama, and Bob Henery. Characterizing the applicability of classification algorithms using meta-level learning. In *ECML-94*, pp. 83–102, 1994.

[Bro93] Carla E. Brodley. Addressing the selective superiority problem: Automatic algorithm/model class selection. In *Machine Learning: Proceedings of the Tenth Intl. Conf.*, pp. 17–24, 1993.

[FI93] Usama M. Fayyad and Keki B. Irani. Multi-inerval discretization of continuous-valued attributes for classification learning. In *13th International Joint Conference on Artificial Intelligence(IJCAI-93)*, pp. 1022–1027, 1993.

[MA95] 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.

[Qui93] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[RC90] Larry Rendell and Howard Cho. Empirical learning as a function of concept character. *Machine Learning*, Vol. 5, pp. 267–298, 1990.

[Sch94] Cullen Schaffer. A conservation law for generalization performance. In *Machine Learning: Proceedings of the Eleventh International Conference*, pp. 259–265, 1994.

[SOF+93] Eiichiro Sumita, Kozo Oi, Osamu Furuse, Hitoshi Iida, Tetsuya Higuchi, Naoto Takahashi, and Hiroaki Kitano. Example-based machine translation on massively parallel processors. In *13th International Joint Conference on Artificial Intelligence(IJCAI-93)*, pp. 1283–1288, 1993.

[SW86] Craig Stanfill and David Waltz. Toward memory-based reasoning. *Communications of the ACM*, Vol. 29, No. 12, pp. 1213–1228, December 1986.

[WK91] Sholom M. Weiss and Casimir A. Kulikowski. *Computer Systems That Learn*. Morgan Kaufmann, 1991.

[毛利 95] 毛利隆夫. 記憶に基づく推論に関する研究 — 属性重み付け手法の研究と天気予測への応用 —. 東京大学大学院工学系研究科情報工学専攻学位論文, 1995.

[林 93] 林知己夫. 数量化 –理論と方法–. 朝倉書店, 1993.