人工データを用いたMBRの属性重み付け手法の評価

1P－8

毛利 隆夫* 田中 英彦
{mohri,tanaka}@MTL.T.u-tokyo.ac.jp
東京大学　工学部

1 はじめに

一般に概念学習のアルゴリズムの評価には，現実世界の問題から得られたベンチマークデータがよく用いられるが，ベンチマークデータの選定に恣意性の介入する余地があるなどの問題があった．本研究では，MBR (Memory-Based Reasoning: 記憶に基づく推論)[2] の属性重み付け手法を例にとり，人工的に合成したデータによるアルゴリズムの評価を行なう．

2 MBRにおける属性重み付け手法

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．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つ．MBRはこれまでに英単語発音問題 [2] や機械翻訳などに応用され，成果を挙げている．

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おり，ベンチマークデータによる比較も行われている [3]．本研究では後で述べるベンチマークデータによる比較の欠点を避けるため，人工的に合成したデータによる属性重み付け手法の評価を行なった．具体的には，条件付き確率を元にした方法 (PCF,CCF,VDM)，相互情報量を元にした方法 (MIC)，等重み値法 (NN)，incremental な方法 (IB4)，および比較対象として主成分分析 (PCA)，数量化II類 (QM2) とその事例ベースへの拡張 (QM2y,QM2m) を試験した (各手法の詳細については [3] を参照されたい)．

3 ベンチマークデータによる評価の問題点

一般に，MBRに限らず概念学習のアルゴリズムの評価には，現実世界の問題から得られたベンチマークデータがよく用いられる．ベンチマークデータを用いた評価の利点は，次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

実際，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

これらの問題点を解決するため，本研究では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を提案する．人工データによる概念学習のアルゴリズムの比較の研究には [1] などがあるが，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いない．

4 人工データによる評価の利点

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，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．一方，人工データを作成する際には，指定するパラメータの選択に注意する必要がある．パラメータが十分細かくない場合には，生成するデータが偏り，特定のアルゴリズムのみに都合の良い結果ばかりが得られかねない．

そこで本研究では，ベンチマークデータの特性をパラメータとして抽出し，そのパラメータにより合成された人工データが元のデータと同等であるかどうかを調べる．もし同等であれば，パラメータはベンチマークデータを再現できる程度に十分細かいことになり，パラメータの細かさが検証できる．なおデータの同等性は，2つのデータでの良いアルゴリズム (最高の正答率か，95%の信頼度の等平均検定により最高正答率と同等とみなせる程高い正答率を得たアルゴリズム) の重なっている割合や，両データの最高の正答率の差で判断する．

5 属性間の依存した人工データの合成方法

一般に人工データを合成する際，各属性毎に値を合成した場合には属性間は独立に近くなる．しかし現実のデータでは，属性同士は互いに依存している場合が多い．本研究では互いに依存した属性で構成されるデータを生成するプログラム sdd(Symbolic Dependent Data generator) を新たに作成した．sdd では一旦各属性を個別に生成したあとで，属性間が依存するようにデータを書き換える．しかもこの書き換えは，同じクラス数，属性数，属性値数などの条件のもとで行なわれ，これらの制約を破壊しないという長所をもつ．なお sdd は 離散値 (シンボル) の属性値のみを対象にしている．

図 1では，属性 a_1 と a_2 に着目し，a_2 の属性値 x と z が入れ替わるようにデータの書き換えが行なわれている．この書き換えにより，$((a_1 = p) \wedge (a_2 = x))$，$((a_1 = r) \wedge (a_2 = z))$ である確率が増加している．その結果，属性 a_1 と a_2 との間の相互情報量も，書き換え以前の 0.101[bits] から 0.197[bits] に増加し，両属性の依存度が増加したことがわかる．sdd ではランダムに2つの属性を選択し，書き換えを何度も繰り返して属性間の依存度を高める．

---

*日本学術振興会特別研究員

[0]Evaluation of Attribute Weighting Methods in MBR using Aritificial Data

Takao MOHRI and Hidehiko TANAKA
Faculty of Engineering, The University of Tokyo

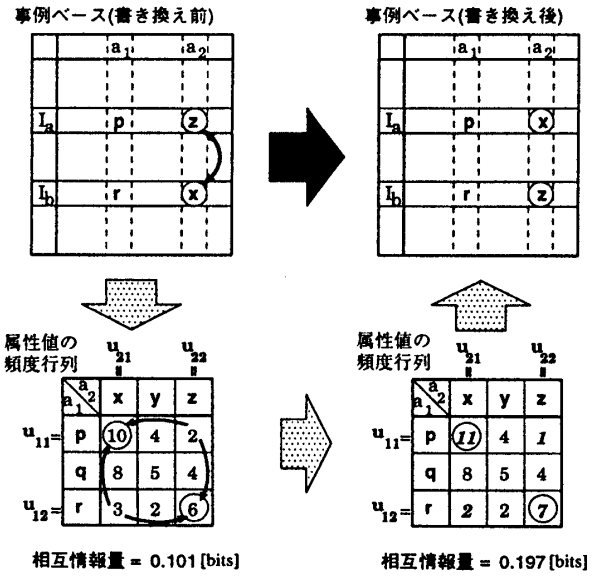

図 1: 属性間を依存させるデータの書き換え

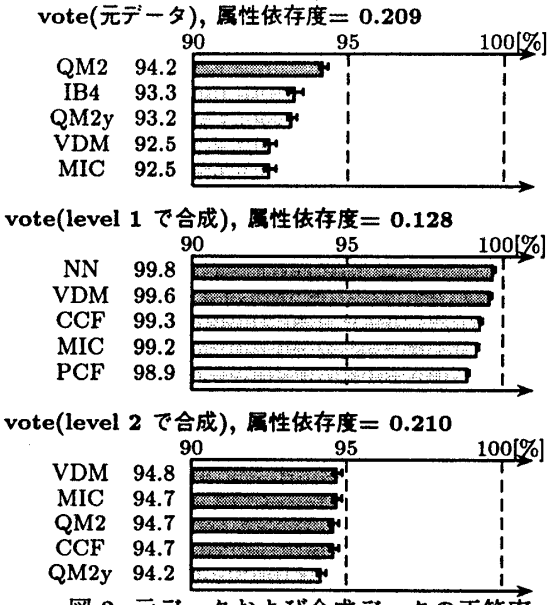

図 2: 元データおよび合成データの正答率

6 人工データによる評価実験

まず，表1にあるようなパラメータ level 1 を vote ベンチマークデータから抽出し，得られたパラメータにより人工データを合成し，元のデータと比較した．図2中の濃い色のグラフは最高もしくはそれと同等の正答率が得られたこと示す．属性依存度は，相互情報量をもとにした値で，0から1までの値をとり，大きい程属性間が依存していることを示している．level 1 は属性間の依存度を考慮していないが，2つのデータの最高正答率や上位のアルゴリズムの傾向は大幅に異なっており，人工データ合成用のパラメータとしては不十分であると言える．

そこで，パラメータ level 1 に属性依存度を加えたパラメータ level 2 により，vote ベンチマークの再現を試みた．属性依存度 0.210 を得るために，3000 回の書き換えを行なった．図 2からは，最高の正答率，および良い結果が得られたアルゴリズムも，元のデータに類似していることがわかる．

書き換えの回数を 0 回から 10000 回まで変化させ，さまざまな属性依存度において，各アルゴリズムの正答率を測定した (図 3)．図に示した全てのアルゴリズムは，属性依存度に応じて正答率が変化している．属性依存度が低い場合のほうが一般に高い正答率が得られており，この傾向は PCF に著しい．これらの結果から，属性間の依存度はアルゴリズムの振舞いに大きな影響を与えることが分かる．

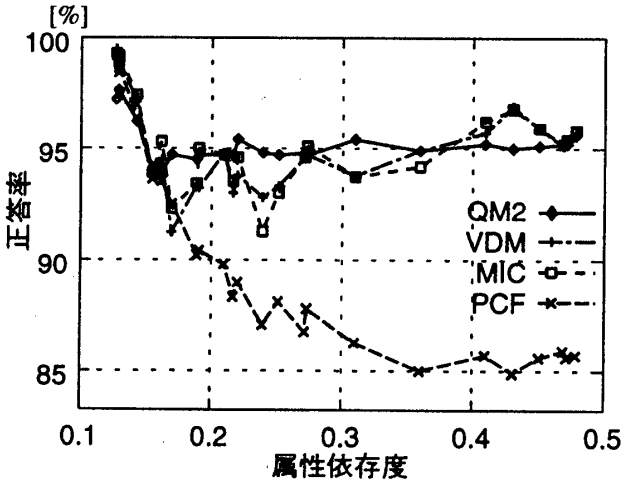

図 3: 属性依存度と正答率

表 1: 人工データ合成用のパラメータ (level 1)

パラメータ	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
$p(v\|c,a)$	$\sum_a N_c \cdot N_v(a)$	属性 a, クラス c のもとでの属性値 vの条件付き確率

7 おわりに

本研究では人工データによる MBR の属性重み付け手法の評価の第一歩として，合成時に指定すべきパラメータを求めた．データの属性間の依存度は従来あまり考慮されていなかったが，正答率に大きく影響するため，データ合成時の重要なパラメータであることを示した．今後はパラメータを変化させて様々なデータを作成し，データ特性とアルゴリズムの優位性との関係を解析する予定である．なお本研究の一部は文部省科学研究費の助成による．

参考文献

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

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

[3] 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, 1994.