記憶に基づく推論における事例の典型度による選択

Selecting Instances by Typicality in Memory-Based Reasoning

毛利 隆夫 Takao MOHRI*
藍 天亮 Tien Liang NAH†
田中 英彦 Hidehiko TANAKA
東京大学 工学部
Faculty of Engineering, The University of Tokyo

In the basic framework of Memory-Based Reasoning (MBR), a large amount of storage is required and the cost of searching similar instances is high since all the training instances are saved in the storage. To overcome these limitations, instance selection is an active research topic in MBR. In this paper, we propose methods to select instances by using the typicality of instances. Giving a priority to instances who have low or middle typicality, the number of stored instances could be decreased without degrading the accuracies in some cases.

1 はじめに

記憶に基づく推論 (Memory-Based Reasoning: MBR)[SW86] は，どのクラスに属するかが未知の事例のクラスを予測する場合に，蓄積しておいた過去の大量の事例の中から類似している事例を検索し，それらを参考にして未知事例のクラスを決定する．同じ種類の分類問題を対象とする方法と比べて，MBR は高い正答率が得られる場合が数多くあるものの，大量の事例をそのまま蓄積するため，記憶量が膨大になったり，類似事例を検索する時間が多くかかるなどの問題があった．

本研究では，事例の典型度を定義し，典型度に基づいて事例を選択する方法を提案する．典型度の高い事例を優先する過去の関連研究を紹介した後，事例の典型度ごとに分類に果たす役割を考察し，典型度の低い事例を優先する方法，および中程度の典型度の事例を優先する方法を提案する．

2 記憶に基づく推論 (MBR)

2.1 MBR の基本的な枠組み

記憶に基づく推論 (Memory-Based Reasoning: MBR) は，ルールに基づく推論のように知識を抽象的な形では持たず，事例のまま保持するのが特徴である．

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，第5章の実験で用いる vote というデータには，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．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，個々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．

MBR は回答が未知の事例，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．予測の際には，投票結果が類似している議員の所属を参考にする．最も簡単には，投票結果がもっとも類似している議員の所属を，所属が未知の議員の所属政党だとして回答する．

このようなクラスが未知の事例を，どのクラスに属するかを予測する問題は，クラスタリングや概念学習と呼ばれており，活発な研究がなされている．最近ヨーロッパで行なわれた StatLog[MST94] というプロジェクトでは，約 20 種類の概念学習問題を解くアルゴリズムを，22 種類のベンチマークに対して適用するという大規模な比較実験が行なわれている．対象となったアルゴリズムには，MBR をはじめ，帰納推論を行なう C4.5 や，誤差逆伝搬法，線形判別関数による多変量解析などが含まれていた．その一連の比較実験実験で MBR は，22 種類中 4 種類のデータで 1 位，2 種類のデータで 2 位の正答率を得ている．1 位になった回数は，全アルゴリズム中最多である．これらの結果により，MBR は他のアルゴリズムと比較して，優位な面を持つ手法であるといえる．

2.2 MBR の改良点

しかし，基本的な MBR には幾つかの欠点が指摘できる．

1. 重要でない属性の影響を受けやすい

   基本的な MBR では，すべての属性が同じ重要度を持つと仮定して類似度が計算される．したがって，分類に貢献しないような属性や，ノイズを多く含む属性も，分類に必要な属性と同等に扱われてしまう．

2. 必要な記憶量が多く，回答生成時のコストも大きい

   MBR では，基本的にすべての事例を蓄えておくために，必要とする記憶容量は決定木による方法や多変量解析と比べてはるかに多い．また，回答生成時には質問と，蓄

連絡先: 毛利 隆夫　東京大学工学部 電気工学科 田中英彦研　〒 113 東京都文京区本郷 7-3-1
Phone: 03(3812)2111 内線 7413　Fax: 03(5800)6922　E-mail: mohri@MTL.T.u-tokyo.ac.jp
*日本学術振興会特別研究員,　†現在シンガポールに帰国

えている事例との類似度を個々に計算するため，回答に要するコストは，事例数に応じて線形に増加してしまう．

前者の欠点を克服するためには，属性を選択する方法 [AB94] や，属性に重み値を定める方法 [SW86, 毛利 94] が，活発に研究されている．また後者を解決するためには，事例に重みを付ける方法 [CS93] や，事例を選択する方法がある．本研究は後者の欠点の改良を目指すものである．

3 事例の典型度

事例の選択および重み付けを同時に行なう方法の一つとして，Zhang による研究 [Zha92] が挙げられる．Zhang は事例の典型度 (typicality) を定義し，それに基づいて事例の重み付けおよび選択を行なった．

事例の典型度は，事例間の類似度をもとに，式 (1) のように定義される．

$$\text{典型度} = \frac{\text{同じクラスの事例との類似度の平均}}{\text{異なるクラスの事例との類似度の平均}} \quad (1)$$

ここで，e_1, e_2 を m 個の属性をもつ事例，max_i, min_i を i 番目の属性の，データ全体での最大値，最小値とする．事例間の類似度 $sim(e_1, e_2)$ は，事例間の距離 $dis(e_1, e_2)$ をもとに次のように計算される．なお，W は第 4節で後述する事例の重み値である．

$$sim(e_1, e_2) = 1.0 - dis(e_1, e_2) \quad (2)$$

$$dis(e_1, e_2) = W \times \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{e_1 - e_2}{max_i - min_i}\right)^2} \quad (3)$$

3.1 典型度毎の事例の性質

事例の典型度のうち，高い典型度，1.0 に近い典型度，低い典型度には，それぞれ次のような意味がある．

まず，高い典型度を持つ場合には，その事例の近傍にある事例は，ほとんどが同じクラスに属していると考えられる．この事例は，まわりの事例を代表することができるような，いわゆる典型的な事例である．

典型度が 1.0 に近い事例の回りには，自分と同じクラスの事例とそうでないクラスの事例とが同様に分布している．したがってこの事例は，クラスの境界付近に存在すると考えられる．

最後に，典型度が 1.0 よりもずっと小さい場合には，その事例の回りには，自分のクラスとは異なるクラスに属する事例が多く，例外的な事例であるか，ノイズを含んだ事例であることが予想される．

4 典型度による事例の選択

Zhang の研究では，図 2 のようなアルゴリズム (TIBL: Typical Instance Based Learning) を用いている．3. では典型度

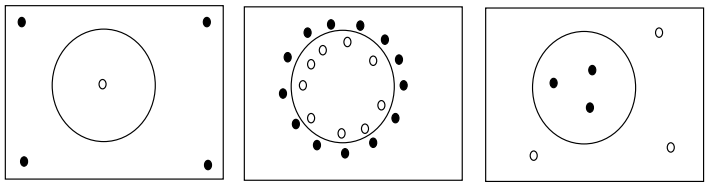

図 1: 典型度毎の事例の分布

1. すべての事例の典型度を計算
2. CD = null
   (CD: concept description: 概念記述)
3. X: 分類に失敗する最も典型的な事例 X,
   Y: X を正しく分類するのに役立つ事例のうち最も典型的な事例
4. 事例 Y に重みをつける
   $$weight(Y) = \frac{1}{typicality(Y)}$$
5. Y を CD に加える
6. すべて事例が正しく分類されるまで，3,4,5 を繰り返す

図 2: TIBL での処理の流れ

の高い順に CD(concept descrption) に蓄積するかどうかを判断しており，高い典型度をもつ事例を優先的に選択している．また，4. では典型度に基づいて事例に重み付けをしている．事例の重みは式 (3) によって距離に反映されるため，重みが小さいほど重要な事例である．

4.1 TIBL-L: 低い典型度を優先する方式

Zhang の TIBL では，高い典型度をもつ事例から順に，蓄積するかどうかを判断していた．しかし，分類の際に問題になるのは，境界に近い箇所に分布する事例の場合であって，境界から遠く離れた典型度の高い事例の場合は，分類を誤ったり，クラス間の得点が僅差になるようなことはないと考えられる．

そこで新たに，分類に実際に役に立つのは境界に近い事例，つまり典型度の低い事例であるとの仮説のもとに，低い典型度の事例を優先する方式 (TIBL-L (L は low の意味)) を考える．TIBL では事例を典型度の大きい順に蓄積するかどうかを判断していたが，TIBL-L は逆に，典型度の低い順に蓄積するかどうかをチェックする．

4.2 TIBL-M: 中程度の典型度を優先する方式

さらに，第 3.1節でみたように，典型度の低い事例は，例外やノイズを多く含む事例であることが予想されるため，典型度の低すぎる事例は除外しておいた方が正答率が高まることが予想される．

そこで，TIBL-L を改良し，まず全事例の典型度を計算し，典型度がある閾値よりも小さい事例は最初に除外するようにした．その後は TIBL-L と同様に，残った事例のうちから典型度の低い順に，蓄積するかどうかをチェックしていく．この方式は中程度の典型度をもつ事例を優先するため，TIBL-M (M は middle の意味) と呼ぶことにする．データ毎に典型度の平均値が異なるため，典型度を閾値として直接指定するのではなく，式 (4) のように，全事例の典型度の平均に係数をかけたものを閾値とする．以下ではこの係数を 「足切り係数」と呼ぶ．

$$\text{閾値} \quad = \quad \text{足切り係数} \times \text{平均典型度} \qquad (4)$$

5 実験

TIBL，TIBL-L, TIBL-M の 3 つの方法を，UCI Machine Learning Repository[MA95] にある 5 種類のベンチマークデータを用いて比較した．用いたデータの諸元を表 1に示す．表中には，全事例の典型度の平均値も同時に示した．実験での比較項目には，正答率および事例量 (もとの訓練事例の中から選択された事例の割合) の 2 つの項目を用いた．正答率および事例量は，10 fold cross-validation[WK91] により求めた．

表 1: 実験に用いたベンチマークデータ

データ名	属性	クラス数	事例数	平均典型度
wine	13C	3	154	1.17
vote	16B	2	435	2.34
heart	13C	2	303	1.14
breast	9C	2	699	1.86
diabetes	8C	2	768	1.03

C ... 連続値属性，B ... 離散値属性

5.1 TIBL と TIBL-L との比較

まず，高い典型度を優先する TIBL と 低い典型度を優先する TIBL-L との比較を行なった．実験結果を図 3，図 4に示す．正答率を見た場合，TIBL-L が少し劣る場合が多いが，diabetes のように TIBL-L が TIBL を上回る場合もある．その一方で，蓄積した事例量は，すべての場合で TIBL-L の方が少ない．特に heart や，diabetes では半分以下にまで減少している．したがって，TIBL-L は TIBL と比べてほぼ同程度の正答率を，より少ない事例量で得られる場合があるということができる．

5.2 足切り係数と正答率 -事例量

つぎに，TIBL-M の結果は足切り係数に大きく依存するため，TIBL-L と TIBL-M を比較する前に，まず TIBL-M の足切り係数を 0.0 ～ 1.0 の間を変化させ，正答率および事例量を測定してみた．足切り係数と正答率との関係を図 5に，事例量との関係を図 6に示す．

正答率は，足切り係数が 1.0 に近付くにつれて diabetes を除き下降する傾向にある．特に，データ wine は下降する幅が大きい．一方事例量は，足切り係数が 0.4 付近の小さい段階で大きく下降するもの (vote, breast) と，0.8 付近の大きな値で下降するものとに分けられる．このように，足切り係数に対するデータの振舞いはデータに強く依存している．

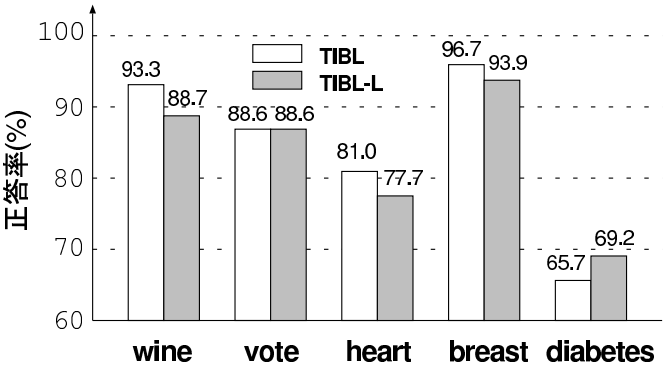

図 3: 低い典型度を優先した場合の正答率

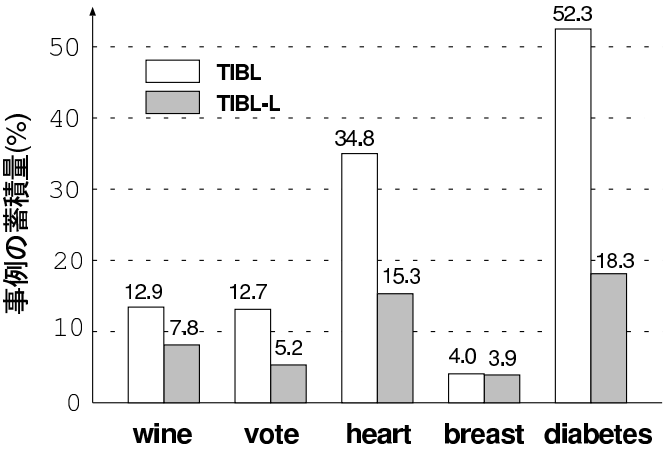

図 4: 低い典型度を優先した場合の事例量

5.3 TIBL-L と TIBL-M との比較

最後に，低い典型度の事例を優先した場合で，足切りを行なわない TIBL-L と，行なう TIBL-M とを比較した．TIBL-M での足切り係数は，正答率が同程度で事例量が少なくなる値に設定すべきであるが，足切り係数に対する正答率や事例量の変化がデータに依存するため，人間が適当な値を選んで TIBL-L と比較した．結果を図 7および図 8に示す．実験の結果，正答率は同程度で，さらに事例量を削減できていることが分かる．これは，適切な足切り係数を求める方法が考案できれば，正答率をほぼ同じに保ちつつ，さらに自動で事例量を削減できる余地があることを示している．

6 考察

TIBL と TIBL-L との比較では，正答率が下がってしまうデータが多く見られたが，この原因はまだ十分に解析できていな

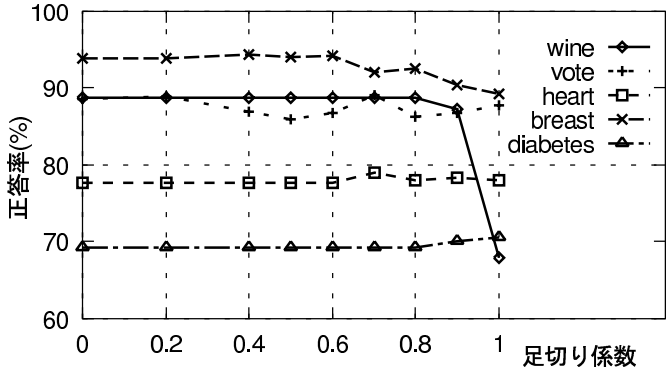

図 5: 足切り係数と正答率との関係

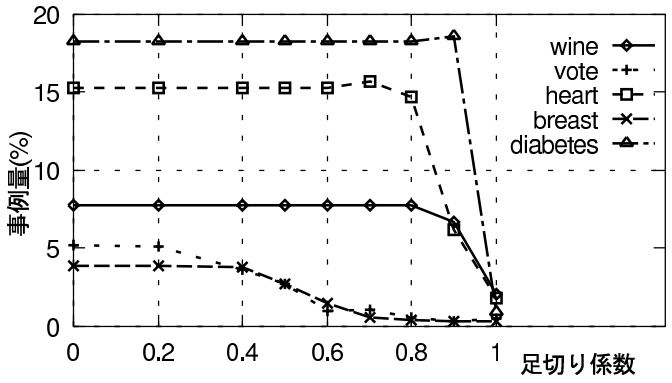

図 6: 足切り係数と事例量との関係

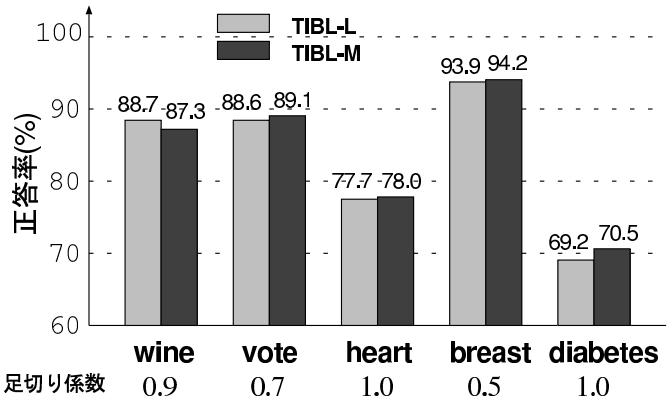

図 7: 足切り係数をデータ毎に定めた場合の正答率

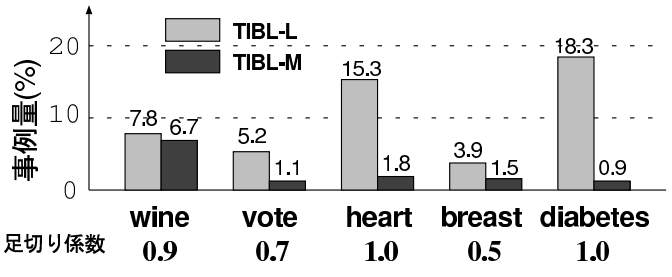

図 8: 足切り係数をデータ毎に定めた場合の事例量

い．しかし，一般に事例の蓄積量と正答率は，トレードオフの関係にあることが予想されるため，多少の正答率の低下よりも事例の蓄積量を少なく抑えたいような場合には，本手法は有効であると考えられる．

TIBL-L と TIBL-M との比較では，典型度の低い事例の足切りによって，正答率を同程度に保ちつつ，事例量を大幅に削減することに成功した．しかし，実験ではデータに依存しない最適な足切り係数を求めることができず，うまく足切り係数が決まった場合の正答率 •事例量への効果を示すに留まった．これは，足切りの閾値を決める際に，全事例の典型度の平均値のみを用いており，典型度の分布を全く考慮していないのが原因である．典型度の分布などを考慮して，自動的に足切りの際の閾値を定めるのは，今後の課題である．

7　結論および今後の課題

事例の典型度に注目し，典型度の低い事例を優先的に蓄積することで，ほぼ同程度の正答率で事例量を削減することができる場合があることがわかった．また，典型度の低い事例を足切りすることで事例量は更に削減可能であることを示した．

今後の課題としては，データに依存しない足切り係数の設定方法や，データの特性と本手法の有効性との関係の解析などが挙げられる．

謝辞

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

[AB94] David W. Aha and Richard L. Bankert. Feature selection for case-based classificaiton of cloud types: An empirical comparison. In *AAAI-94 Case-Based Reasoning workshop*, pp. 106–112, 1994.

[CS93] Scott Cost and Steven Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, Vol. 10, pp. 57–78, 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.

[MST94] D. Michie, D.J. Spiegelhalter, and C.C.(ed) Taylor. *Machine Learning, Neural and Statistical Classification.* Prentice Hall, 1994.

[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.

[Zha92] Jianping Zhang. Selecting typical instances in instance-based learning. In *Proceedings of the Ninth International Machine Learning Workshop(ML92)*, pp. 470–479, 1992.

[毛利 94] 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.