分類子学習のためのクラス生成に関する認知科学的実験

3T－3

清水 健太郎　鈴木 英之進

横浜国立大学工学部電子情報工学科

1 はじめに

教師なし学習と教師つき学習の統合学習は，単なる教師なし学習とは異なり，優れた分類子によって説明されるクラスを生成できる利点がある．例えば，統合学習において，製造段階での自動車の部品構成などから故障を早期発見するシステムが提案されている[1]．このシステムは，いままで知られていなかった故障原因の発見に成功し，統合学習の有用性を示している．しかし，この研究は応用事例であり，一般的な統合学習に関する研究は，ほとんど報告されていない．

このような統合学習においては，教師なし学習におけるあまりに複雑なクラス分けは，分類子の可読性と正確性を低下させる．しかし，クラス分布が極端に偏り単純な場合は得られる情報が少ない．したがって，クラスの複雑さと分類子の良さを両立させる，教師なし学習における属性選択規準が重要となる．そこで今回，この属性選択規準を人間がどのように設定し，有用な知識を発見していくかを調べるために，認知科学的実験を行い，その結果を解析した．

2 統合学習

教師なし学習と教師つき学習を適用するデータ集合を，それぞれ目的データ集合 D_P，説明データ集合 D_E とする．また，教師なし学習と教師つき学習のアルゴリズムを，それぞれ A_P，A_E とする．D_E と D_P は，含む例は同じだが，属性は異なるものである．統合学習では，まず，A_P に $D_{P\alpha}=\{e_{P1}, e_{P2}, \cdots, e_{Pn}\}$ を入力し，，出力 $\{S_1, S_2, \cdots, S_m\}$ を得る．ただし，$D_{P\alpha}$ は，属性選択規準 α によって，D_P から特定の属性だけを抽出したデータ集合であり，e_{Pj} は，α によって抽出された属性集合をもつ例である．また，S_i は一つのクラスを表し，クラスは互いに排反であり，$\sum S_i = D_P$ であるとする．次に，この S_i をクラス統合規準 β に照らし合わせ，複数のクラス $S_i, S_j, \cdots$ を統合したり，$D_{P\alpha}$ の属性について追加や削除を行う．S_i が良いと判断されたら，これを D_E へクラス属性として付け加え，A_E を適用し分類子を得る．A_P により生成されたクラスは，この分類子により，D_E の属性を用いて説明されることになる．分類子の正答率，可読性が優れていれば終了，そうでなければ属性選択へもどり，これまでの手順を繰り返す．

3 認知科学的実験と比較実験

認知科学的実験実験では，A_P としてクラスタリングアルゴリズム *Autoclass*[2] を，A_E として決定木構築アルゴリズム *C4.5*[3] を用いた．被験者は20名の学生であり，彼らの学習結果について，クラスの複雑さを表す情報量，分類子の正確さを表す正答率及び分類子の可読性を表すノード数を記録した．この実験で被験者が行う操作は，最初の属性選択と，クラスタリング後のクラス統合の二つである．また，実験の最後に，被験者にこれらの操作において用いた規準 α, β を記述させた．対象とするデータは，1994年のJR東日本キヨスク東京地区212店舗における，炭酸飲料やガムなどの商品についての棚卸し量を記録したデータと，営業開始時間や社員数などの店舗構成を記録したデータである．これらはそれぞれ目的データ集合 D_P，説明データ集合 D_E に相当し，属性数はそれぞれ52, 47である．

上記の実験結果と比較するために，6個あるいは9個の属性をランダムに選択してクラスとし，そのクラスを説明する決定木を構築する実験を，1000回繰り返した．そして，認知科学的実験で得られる正答率などの数値について，その値を1000回中何回越えたかをパーセンテージで表す難易度を求めた．難易度は，小さいほど得られた数値が良いことを意味する．

4 実験結果

2節の統合学習では，属性選択規準 α とクラス統合規準 β が解析対象として存在するが，今回はより重要と判明した属性選択規準 α を主に調べた．

実験結果を表1に示す．全ての規準を通しての平均属性選択数は7.7個で，6個から9個の範囲に全体の62.5%が入っていた．この表に示すように，本実験では属性選択規準 α を，特定の消費者が購入しそうな商品を選択する共通消費者規準，類似商品を選択する共通商品特徴規準，クラスの情報量が低下しないようにする情報量規準，生成されたクラスごとの棚卸し量が類似している商品を選択する生成クラス規準，適当に選択する無作為規準，および優れた分類子が得られる属性だけを選択する分類子規準に分類した．最初の二つの規準が領域知識を用いる規準であり，次の四つ

Cognitive Science Experiments for Class Generation in Classification

Kentaro Shimizu and Einoshin Suzuki
Division of Electrical and Computer Engineering, Faculty of Engineering, Yokohama National University
79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan

表 1: 認知科学的実験結果 （かっこ内の数値は最小値ー最大値を示す）

	属性選択規準	度数	人数	平均情報量 (bit)	平均正答率 (%)	平均ノード数
本実験	共通消費者規準	8	7	1.30(1.15-1.60)	91.8(87.8-94.8)	13.3 (9-22)
	共通商品特徴規準	6	5	0.82(0.39-1.52)	90.1(85.4-96.2)	14.7(9-22)
	情報量規準	1	1	1.21	92.0	16.0
	生成クラス規準	1	1	1.36	87.8	18.0
	無作為規準	4	4	0.94(0.65-1.22)	90.0(88.2-91.1)	20.5(20-21)
	分類子規準	3	2	1.14(0.58-1.38)	90.5(89.4-91.6)	11.9(6-14)
追加実験	情報量規準	6	5	1.42(1.31-1.53)	88.6(86.8-89.9)	16.6(14-19)
	生成クラス規準	7	5	1.34(1.22-1.56)	88.1(86.3-89.1)	16.0(9-23)

が領域知識を用いない規準である．これらのうち情報量規準と生成クラス規準は共に度数が 1 と少ないため，追加実験を行った．追加実験では被験者は 10 名であり，情報量規準を採用する者 5 名と，生成クラス規準を採用する者 5 名に分けた．

以上の実験より，領域知識を用いる規準では，共通消費者規準が共通商品特徴規準に比べて情報量，正答率，ノード数全てで優れていた．属性選択数 6 個の比較実験における，正答率 91.8%と 90.1%の間，ノード数 13.3 と 14.7 の間の難易度の差は，それぞれ 1.6%，0.7%とほとんどない．しかし，情報量1.30bit と 0.82bit の間では23.8%もあり，クラスの複雑さという面で，共通消費者規準がより優れていると考えられる．これは，共通商品特徴規準が，特定の共通点を持った商品を選択するためと考えられる．例えば飲料という特徴をもつ商品だけを選択した場合，飲料を扱っている店舗とそうでない店舗という，クラス分布が極端に偏り単純なクラスが生成される可能性がある．

一方，領域知識を用いない規準では，情報量規準と生成クラス規準は，平均情報量が高く平均正答率が低い．また，無作為規準と分類子規準は，平均情報量が低く平均正答率が高い．前者と後者で，属性選択数 6 個の比較実験における正答率の難易度の差は 2.0%程度である．しかし，情報量では難易度の差が 28.8%もある．したがって，前者の方が，正答率と情報量を総合して考えると有効である．なお，ノード数については，分類子規準が非常に優れており，情報量規準と生成クラス規準が平均的，無作為規準は劣っている．しかし，分類子規準では，ノード数が小さくなるとクラスの情報量は小さくなる傾向がある．そのため，分類子規準における情報量の最低値は 0.58bit である．この値は，比較実験での難易度は，属性選択数 6 個の場合で 97.5%である．これより，情報量規準と生成クラス規準が，領域知識を用いない規準の中では総合的に最も優れた規準であるといえる．

情報量規準と生成クラス規準を共通消費者規準と比較すると，正答率とノード数では難易度は約 2%劣るが，情報量で約 8%優れているので，同程度に有効と考えられる．

クラス統合規準 β では，統合の対象となったクラスは，互いに属性値の分布が似たクラスであった．二つのクラスを統合すると，分類子の正答率は平均約 3%向上するが，クラスの情報量は平均約 0.2bit 低下してしまう．しかし，比較実験により，正答率 3%の差は難易度にして約 2%の差であるのに対し，クラスの情報量 0.2bit の差は 16%もの差に相当する．さらに，領域知識を用いる場合は必ずクラスが統合されたのに対し，用いない場合は約半数の場合で統合されなかった．したがってクラス統合規準 β は，正答率向上のための補助的な役割を果たすものであり，クラスが複雑過ぎて正答率が低い場合だけに有効である．

5 おわりに

本研究では，教師なし学習と教師つき学習の統合学習について，人間がどのような規準でクラスを生成し，有用な知識を得ていくのかを調べるために，商用データを用いた認知科学的実験を行った．その結果，購買者を想定した共通消費者規準が，属性選択規準として最も優れていた．また，生成クラスの情報量あるいは特徴に注目した規準も，生成クラスの複雑さ，分類子の正答率および可読性を総合すると，共通消費者規準と同様に優れていた．なお，クラス統合は，これら属性選択規準の補助的な役割を果たすことがわかった．

参考文献

[1] R. Wirth and T. P. Reinartz : "Detecting Early Indicator Cars in an Automotive Databases", *Proc. KDD-96*, pp.76-81 (1996).

[2] P. Cheeseman and J. Stutz: "Bayesian Classification(Autoclass)", *Advances in Knowledge Discovery and Data Mining*, pp.153-180, AAAI Press/MIT Press (1996).

[3] J. R. Quinlan: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993).

謝辞

本研究は，財団法人日産科学振興財団の援助を受けている．