日産科学振興財団研究報告書，21(1998)

データマイニング技術を用いた人間の情報抽出特性に関する構成的研究

Constructive Analysis of Information Extraction Properties by Humans with Data Mining Techniques

研究代表者 横浜国立大学工学部電子情報工学科 助教授 鈴木 英之進
Associate Professor, Electrical and Computer Engineering,
Yokohama National University,
Einoshin SUZUKI

This paper presents a constructive analysis of information extraction properties by humans with data mining techniques. First, we have conducted cognitive science experiments for an integration of supervised and unsupervised learning. Then, an information extraction model has been built through a careful analysis of the results. Finally, we obtained a novel data mining technique for the integrated learning based on the information extraction properties by humans.

1 研究目的

高度情報化社会と呼ばれる今日，電子化された大規模データが急速に増加している．これらの大規模データは，そのままでは有用ではなく，人間にとって意味がある情報を抽出することが重要となっている．このような状況を背景に，データマイニングと呼ばれる，大規模なデータから有用な情報を発見することを目標とする研究分野が急速に発展している．データマイニングにおいてはこれまでに種々の技術が提案され，さまざまな分野において一定の成果をあげているが，人間の情報抽出特性に着目した研究はなかった．人間はさまざまな情報抽出戦略を融合して用いていると考えられ，その特性を解明することは，データマイニングに対しての貢献が大きいと考えられる．また，学習研究という名のもとにこれまで科学的な興味からしか調べられていなかった情報抽出特性を，工学的に新しい技術を確立するという観点から調べることは，斬新かつ興味深いと考えられる．

以上の点を考慮し，本研究では，1) 認知科学的実験を行い人間の情報抽出特性を解明すること，2) 同実験結果を分析し，既存のデータマイニング技術を融合する構成的接近に基づいて人間の情報抽出特性をモデル化すること，および 3) 同モデルを構成することにより，新しいデータマイニング技術を確立することの 3 点を目標とした．

2 研究経過

2.1 問題設定：統合学習

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．

統合学習の形式的な定義は，次のようになる．

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 $D_{Pa} = \{e_{P1}, e_{P2}, \cdots, e_{Pn}\}$ を入力し，出力 $\{S_1, S_2, \cdots, S_m\}$ を得る．ただし，D_{Pa}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，このクラスをクラス統合規準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_{Pa}の属性について追加や削除を行う．クラス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．

このような統合学習においては，教師なし学習における過度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．そこで主に，この属性選択規準αに着目して研究を進めた．

2.2 認知科学的実験

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およ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 6 個の実験で 1000 回中何回越えたかを百分率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．

2.3 実験結果の分析

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，正答率および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

表 1: 認知科学的実験結果 （かっこ内の数値は最小値－最大値を示す）

	属性選択規準	度数	人数	平均情報量	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
	Autoclass 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)
	結果特徴抽出規準	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	5	5	1.42	88.6%	16.6
	Autoclass 規準	5	5	1.34	88.1%	16

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．

2.4 人間の情報抽出モデルの構成

人間の情報抽出モデルは，前節で述べた属性選択規準αおよびクラス統合規準βを統合学習において用いることにより構成できる．ただし，本研究では，統合学習に有効なモデルを構成した．

属性選択規準αとして有効であるのは，共通消費者規準，情報量規準および Autoclass 規準である．ここで，共通消費者規準は，領域知識を用いるものであるために対象データに依存し，さらに従来の知識ベースシステムの範疇に属する．一方，情報量規準および Autoclass 規準は，実装にあいまい性があるものの，領域知識を用いないためにすべてのデータに適用可能である．本モデルでは，よりあいまい性が少ない情報量規準を属性選択規準αとして用いた．モデルは，難易度を参考にしながら情報量，正答率およびノード数を改善する山登り探索法を採用している．構成したモデルは，助成金で購入した計算機上にプロトタイ

プシステムとして実装した．なお，クラス統合規準βは，属性選択規準αの補助的なものであるため，モデルには含めなかった．

3 研究成果

3.1 人間の情報抽出特性

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．

人間の情報抽出特性を解明することは，認知科学の分野において学習研究という名のもとに盛んに行われてきた．対象領域としては，数年前までは小規模な，架空の問題が主に扱われてきたが，近年より現実的な問題が選択されている．しかし，これらの研究は主に科学的な観点から行われており，大規模なデータから有用な情報を発見するという，工学的な観点から行われた研究は極めて珍しいと考えられる．

3.2 人間の情報抽出モデルと統合学習アルゴリズム

統合学習において，情報量規準を用いた人間の情報抽出モデルを構成し，計算機上にプロトタイプシステムとして実装した．このシステムは，人間の情報抽出に着目した統合学習アルゴリズムともなっている．

本アルゴリズムは，領域知識を用いない一般的なものであり，種々の問題に適用可能である．また，小売業データを対象として，その有効性を確認している．難易度を参考にしながら情報量，正答率およびノード数を改善するアイデアは，従来なかったものであり，人間の情報抽出モデルを解析してはじめて生まれたものである．

4 今後の課題と発展

まず，今回は統合学習に限定したが，今後は人間の学習と発見過程一般について，工学的な観点から検討したい．例えば，なぜある種の学習や発見は他に優先するのかという問題を，構造的な側面と効用的な側面から調べることなどを考えている．その際には，本研究で得られた人間の情報抽出特性に関する実験結果と人間の情報抽出モデルが有用な知見となる．

また，本研究では人間の情報抽出特性を対象としたが，今後は逆に，人間が抽出しにくい情報を考えてみたい．この問題を考えることにより，人間の情報抽出能力を補うような対話型データマイニングシステムを開発できると考えている．

最後に，本研究で開発した統合学習アルゴリズムは，小売業以外の種々の問題に適用して起こりうる問題点を詳細に分析することにより，より頑健で有効な手法に改良していきたい．

5 発表論文リスト

1. 清水健太郎，鈴木英之進：「クラス生成を含む例からの帰納学習に関する認知科学的実験」，第34回人工知能学会人工知能基礎論研究会，1998年9月（印刷中）．

2. 清水健太郎，鈴木英之進：「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」，第57回情報処理学会全国大会講演論文集，1998年10月（印刷中）．

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).