

典型性に基づく概念学習アルゴリズム

谷澤 正幸 上原 邦昭 前川 穎男
神戸大学工学部情報知能工学科

本稿では、過去の事例集合から得られる統計的な情報を用いて未学習の事例を分類するアルゴリズムを提案する。従来の帰納的概念学習アプローチでは、事例集合からカテゴリー間の違いを最小限のルールとして導く手法に焦点があてられていた。本稿で提案するアルゴリズムは、概念（カテゴリー）の典型的な特徴に着目して、未学習の事例を各々のカテゴリーの典型的な特徴と比較しながら分類を行なう手法である。本アルゴリズムは、各カテゴリーに属する全ての特徴に対して典型性の度合いを求めて分類に用いているために、少ない訓練例からも十分に高い分類精度を実現することができる。また、事例を平均化した典型的なカテゴリーの概念が獲得されるために、事例の一部に特徴の欠落や誤りが含まれていたとしても影響を受けることはほとんどない。さらに、本アルゴリズムは複数のカテゴリーを分類候補として導出することにも可能であり、分類候補の絞り込みアルゴリズムとしても利用できるという特徴がある。

Prototype Based Concept Learning Algorithm

Masayuki Tanizawa Kuniaki Uehara Sadao Maekawa
Department of Computer Science and Systems Engineering
Kobe University
Nada, Kobe, 657 Japan

This paper describes a prototype based statistical concept learning algorithm from cases. Most of the existing inductive classification algorithms mainly concentrate on the extraction of minimum discrimination rules to separate each categories. On the other hand, our approach is to classify a new case into nearest categories by prototype based distance metric, where prototype theory was proposed E, Rosch. Prototype based classification approach uses more conceptual information from training cases than rule induction approaches and this approach achieves high accuracy to classify from small training cases. Furthermore this approach can derive some possible categories for a new case.

1 はじめに

従来の帰納学習のアプローチでは、過去の事例を一般化してルールを導く手法が中心的なテーマであった。これは、従来、人間は経験からルールを帰納して学習を行っていると考えられていたこと、計算機にとってはルール形式が取り扱いやすいことなどの理由による。しかしながら、人間は常に計算機が用いるような完全なルールを用いているわけではなく、概念（カテゴリー）の典型性（prototypicality、事例集合から得られる概念の中心的、理想的な代表型）[1] を用いた不完全なルールを組み合わせて問題解決をしている場合も多い。

本稿では、カテゴリーの典型性に基づいて事例の分類を行うアルゴリズムを提案する [11], [12]。分類問題は、機械学習の分野において代表的な問題の一つであり、診断、パターン認識、事例ベース推論における事例の検索など幅広い応用分野がある。しかしながら、従来の分類問題では、事例の属するカテゴリーを分離するための最低限度のルールを導くことに焦点を置いた研究が多く、典型性を利用した研究はほとんどなかった。これは、カテゴリーの典型性が完全な情報ではないためにルール化が困難であること、カテゴリーの典型性は分類操作と直接的には関係ないと考えられていたこと（典型的な特徴であっても分類作業に重要な特徴とはいえない）などの理由による。

本稿では、まず2章で分類問題について説明し、ルールに基づく手法と典型性に基づく手法の比較を行う。また、典型性に基づいて事例を分類する場合には、各特徴に重要度をどのようにして割り当てるかということが重要となる。このため3章では、特徴の重要度という概念を典型性に基づく分類手法に導入する。さらに4章では、2種類の異なる重要度を用いた二段階による典型性に基づく分類手法について述べる。一方、典型性に基づく分類手法には、ルールに基づく手法に比べて分類に用いるデータ量が膨大になるという問題がある。このため5章では、典型性に基づく分類手法におけるデータ量と分類精度の関係について、ルールに基づく手法との比較を行いながら本手法の有効性を示す。最後に6章では、他のアルゴリズムと比較しながら、本手法の問題点および今後の課題などについて検討する。

2 ルールに基づく手法と典型性に基づく手法の比較

2.1 分類問題

分類問題の一例として、医療における診断作業について検討する。人間の病気には解明されていないものも多く、場合によっては、体を解剖して初めて病気の原因や病名が分かることがある。このため、医学は医者（専門家）の勘や経験に基づいて問題解決が行われる「知識が不完全な領域」となりうることが多い。また、診断作業は、通常の場合、集団検査などの予備的検査や患者の体の異常の訴えに基づいて可能性のある病名を推定し、診断結果に応じてさらに精密検査を行い、詳細な病名や処置法を決定

づけるという手順を探る。この初期段階の検査や症状から正しい（詳細な）病名を推定することが事例の分類問題に相当している。すなわち、医学における分類問題とは、初期段階の検査結果や症状と精密検査や解剖後の正しい病名の組を事例として、症状と病気の関係を推定するという分類問題とみなすことができる。医療の診断においては、患者の症状（特徴集合）から病気（カテゴリー）の診断をした場合、両者の対を患者の事例と呼ぶ。同様に、本稿では、事例はカテゴリーと特徴集合の対によって与えられるものとする。また、事例集合が与えられたとき、事例集合のデータから未知の事例に対して適切な分類（診断）結果を導く手法を分類問題と定義している。

2.2 ルールに基づく分類手法

本節では、事例からルールを帰納する代表的なアルゴリズムとしてID3 [3]について説明する。ID3は、事例集合から新たな事例を分類するための適切な決定木を生成するアルゴリズムである。決定木とは、新たな事例を分類するための手続きを表したものであり、木の終端ノードには分類されるべきカテゴリーがラベル付けされている。終端ノード以外のノードには属性がラベル付けされており、新たな事例が入力されると、最上位ノードの属性から事例の属性値を調べ、その値によって下位のノードへとたどり、終端ノードに達するとそのカテゴリーが決定される。決定木はノードに割り当てる属性の選択によって構造や大きさが変化するために、各ノードに適切な属性を割当てることが重要となる。また、ID3では各属性の情報量を用いて各ノードの属性を選択し、適切な決定木を生成するようしている。

ID3のように帰納的にルールを抽出するアルゴリズムでは、カテゴリー間で異なる特徴を用いて最小限度に事例集合をルール化することが多い。したがって、ルールを導くアルゴリズムは、一般に、最適なルール集合を導く組み合わせ問題となり、計算コストはかかるが最小限のルール集合が得られるために、計算機によって利用しやすく、得られたルールによる分類も高速に行うことができるという利点がある。

2.3 典型性に基づく分類手法

ルールに基づく手法は、カテゴリー間の異なった特徴をルール化するアルゴリズムである。このため、ルールに基づく手法では、典型的な特徴（あるカテゴリーに属する多くの事例で出現する特徴ではあるが、そのカテゴリーに属するすべての事例に含まれているとは限らない特徴）や頻度の少ない特徴についてはほとんど考慮されることがない。本節では、典型的な特徴や出現頻度が少ない特徴が分類に与える影響について考察し、典型性に基づく分類手法について説明する。

まず、カテゴリーが典型的な特徴の分類に与える影響として「脊椎動物の分類」について考える。脊椎動物のうち、「空を飛ぶ」という特徴は鳥にとって典型的な特徴である。しかしながら、ペンギンのように飛べない鳥も存在するし、蝙蝠のように鳥類以外で空を飛ぶ脊椎動物もある。したがって、「空を飛ぶ」という特徴を

鳥への分類ルールに採用することはできない。すなわち、「空を飛ぶ」という特徴は鳥にとって典型的な特徴であるが、分類にとっては完全でないために、不適切な特徴とみなされる¹。しかしながら、人が分類を行う場合には、「空を飛ぶ」という特徴が与えられると、まず一番に鳥である可能性を考える。典型性に基づく分類手法では、このような人の行う分類方法に習って、カテゴリーに典型的な特徴を分類に反映させようとする手法である。

つぎに、頻度の少ない特徴が分類に与える影響として「自動車の分類」を考える。「クレーン」という特徴は「トラック」にとってあまり頻度の多い特徴ではないが、「乗用車」や「オートバイ」には「クレーン」という特徴がないために、「クレーン」だけで「トラック」と断定できるはずである。したがって、頻度の少ない特徴であっても分類に関係ないとは判断できない。典型性に基づく分類手法では、この頻度の少ない特徴の影響も考えて、ルールに基づく手法とは異なり、かなり冗長なデータを残して分類に利用している。

以上のような観点に基づいて、本稿で提案する典型性に基づく分類アルゴリズムは以下のようになる。まず、与えられた未学習の事例（特徴集合 F として与えられる）の各特徴 f_i が、それぞれのカテゴリー内の事例においてどの程度存在しているかという値（存在率）を算出し、すべての存在率について総和を取った値を求める（式 (1) 参照）。この値を特徴集合 F からカテゴリー C への連想度と呼ぶ。

$$\text{特徴集合 } F \text{ からカテゴリー } C \text{ への連想度} = \sum_{f_i \in F} f_i \text{ の } C \text{ での存在率} \quad (1)$$

連想度によって、特徴集合とカテゴリー内の事例の類似している度合が算出される。未学習の事例は連想度の最も高い値を示すカテゴリーに分類される。なお、本稿で提案する典型性に基づく分類手法は、(1) 式で定めた連想度により分類を行うため、以降では連想度手法と呼ぶ。

2.4 ルールに基づく手法と典型性に基づく手法の比較

帰納的ルール抽出アルゴリズム ID3 と連想度手法による soy bean data (特徴数 50、カテゴリー数 17 からなる大豆の病気に関するデータ) を適用して両者の比較を行う。soy bean data は、文献 [2]において専門家とルール帰納アルゴリズム AQ の比較に用いられたデータである。soy bean data は全部で 289 例からなり、比較には訓練例として 145 例、分類のテストに残りの 144 例を用いた。未学習データ 144 例に対する分類精度の結果を表 1 に、また学習に用いる事例の増加による分類精度の変化を図 1 に示す。

ID3 から得られるルールは決定木の形をしており、与えられた特徴集合は決定木にしたがって唯一のカテゴリーに分類されるアルゴリズムである。しかしながら、soy bean data に含まれる事例は特徴数が多いために、各カテゴリーに事例が 1 例のみ含まれる場合には、誤った特徴が決定木のノードとして選択されることも

¹他の特徴と組み合わせて完全なルールにできる場合もある。

表 1: 分類結果 1: 適用例 144 例 括弧内は第 2 候補まで含めた精度を示す

	連想度	ID3
各カテゴリーにつき 1 例 (計 17 例)	64.6 % (85.4 %)	34.0 %
各カテゴリーにつき数例 (計 145 例)	74.3 % (91.7 %)	79.2 %

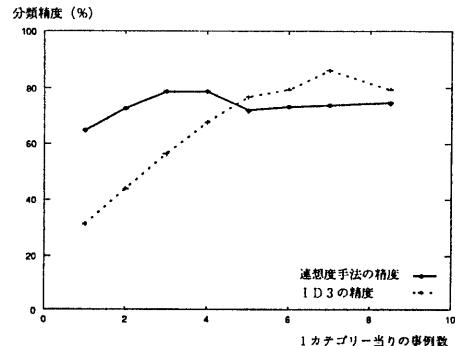


図 1: カテゴリー当たりの学習事例の増加による精度の変化

あり、分類精度が極端に悪くなっていることがわかる。また、ID3 は 1 カテゴリーあたりの事例数が多いほど分類精度が正確になるのに対し、連想度手法は事例が少ないのである程度正確に分類できるが、逆に事例数が増えても精度の改善は余りみられないことがわかる。しかしながら、2 番目の候補カテゴリーまでに正解を含む割合では、連想度手法が ID3 を上回っており、可能性のある候補カテゴリーを複数挙げる場合には連想度手法が有効であることがわかる。

3 特徴の重要度

3.1 重要度の導入

事例が増加しても連想度手法の分類性能が向上しない原因は、分類に関係のない不適切な特徴の影響を受けているためである。したがって、分類に関係ある特徴と関係ない特徴の重みを変えることができれば、事例の分類精度を改善することが可能になると考えられる。言い換えると、各カテゴリーに均等に含まれる特徴は、それが典型的な特徴であっても分類には意味がなく、各カテゴリーに偏在する特徴は頻度の少ない特徴であってもより重要なとなる。

このような考え方から、情報量の期待値（エントロピー）という概念を導入する。たとえば、多くのカテゴリーに含まれる特徴はエントロピーが大きいものである。また、一部のカテゴリーにのみ偏在する特徴はエントロピーが小さいものである。したがって、エントロピーの小さなものは特徴の重みを大きくすればよいということになる。なお、エントロピーを直接利用することは重要度として不適切であるために、本手法ではエントロピーの値

に対して式(2)に示す変換を施したもの用いている²。この変換によって、特徴の重要度の範囲は0から1の間をとり、重要度の値が1であるときにその特徴が最も重要（单一のカテゴリーのみに存在する特徴）となり、重要度の値が0に近づくほど重要な特徴となるようにしている。特徴の重要度によって拡張された連想度を式(3)に示す。以降では、重要度を用いた連想度手法を拡張連想度手法と呼ぶ。また、特徴 f_i のカテゴリー C_j での存在率と特徴 f_i の重要度の積を特徴 f_i からカテゴリー C_j への連想度と呼ぶ。

$$\text{特徴 } f_i \text{ の重要度 } I_i = 2^{-(\sum p_j \log_2 p_j)} = \prod p_j^{p_j} \quad (2)$$

ただし、 p_j はある特徴 f_i が観測されたときにその事例がカテゴリー C_j である可能性を示している。また、 $\sum p_j = 1$ である。

$$\text{特徴集合 } F \text{ からカテゴリー } C \text{ への拡張連想度} = \sum_{f_i \in F} f_i \text{ の } C \text{ での存在率} \times I_i \quad (3)$$

拡張連想度手法を用いて soy bean data を分類した結果を表2に、学習に用いる事例の増加による分類精度の変化を図2に示す。

表2: 分類結果2: 適用例 144 例 括弧内は第2候補を含めた精度を示す

	拡張連想度	連想度
各カテゴリーにつき1例(計17例)	64.6 % (86.8 %)	64.6 % (85.4 %)
各カテゴリーにつき数例(計145例)	94.4 % (99.3 %)	74.3 % (91.7 %)

bean data に限定した場合、候補絞り込み手法として有効に機能していることがわかる。

3.2 重要度の検証

本節では、前節で導入した重要度をさらに拡張して、重要度の値の最適性について検討する。この拡張は、前節で導入した重要度の式を式(4)のように指數乗したものである。なお、以下では(4)式で用いる乗数 x を重要度乗数と呼ぶ。

$$\text{特徴 } f_i \text{ に対する拡張した重要度 } EI_i = I_i^x \quad (4)$$

拡張した重要度 EI では、重要度乗数 x の値によって特徴のエントロピーによる重み付けの度合いを変化させることができるようになっている。また、3.1節で定義した重要度 I は0から1の間の値を取るために、重要度乗数が大きくなれば、 EI はエントロピーの値に敏感に反応する。また、重要度乗数が小さくなれば、エントロピーの値にあまり反応しなくなるという特性を持っている。言い換えると、重要度乗数を大きくしていくば、重要な特徴のみが強調されてエントロピーの大きな値は分類に影響しなくなり、重要度乗数が小さくなれば、あまり重要でない特徴も考慮した分類となる。重要度乗数による分類精度の変化を図3に示す。なお、当然のことながら、重要度乗数が1の場合には前節で導入した重要度と同一の結果を示すことになる。

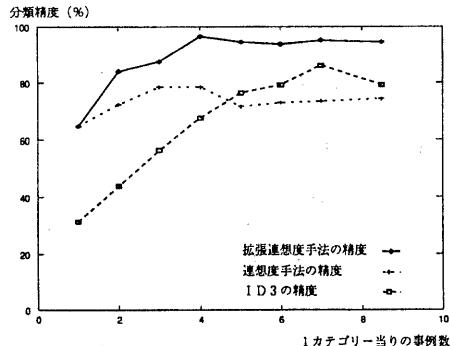


図2: カテゴリー当りの学習事例の増加による精度の変化

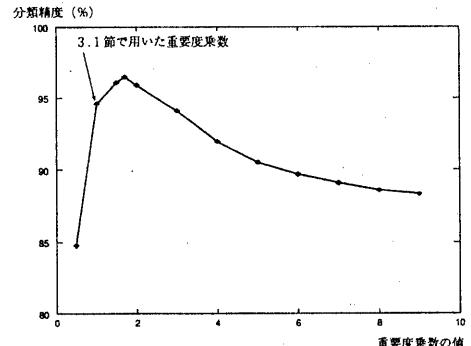


図3: 重要度乗数による分類精度の変化

以上の結果から、特徴の重要度の導入によって、本手法は重要度なしの連想度手法やID3と比較して、かなり高い分類精度が得られていることがわかる。また、本手法では、2番目の候補カテゴリーまでに正解カテゴリーをほぼ完全に含んでいるために、soy

²(2) 式と類似した式変換を用いる手法として、背景知識を用いた決定木生成アルゴリズム EG2[4]で用いられている、決定木のノード選択を行なう関数 ICFがある。

図3の結果から、重要度乗数を大きくしすぎると、重要度が小さな特徴の影響がなくなり分類精度が落ちることがわかる。逆に重要度乗数が小さすぎると、エントロピーの大きさの影響が小さくなり、重要度の差が小さくなるために分類精度が悪くなることがわかる。また、soy bean data に関しては重要度乗数の値が1.7前後で分類精度が最大になっていることがわかる。

4 特徴の重要度の状況依存性

3章では、重要度は各特徴ごとに1個の値を持つように定めている。しかしながら、特徴の重要度のなかには状況によって変化するものが存在する。たとえば、記憶に基づく推論 (memory-based reasoning) を扱った文献 [5] では、特徴の重要度の状況依存性について次のように指摘されている。腹部の痛みを訴える患者の痛みについて診断する場合を考える。痛みの原因が妊娠によるものか、あるいは潰瘍によるものかを考える場合、「性別」という特徴は妊娠に深く関係しているが、潰瘍の有無を判断する場合には意味のない特徴である。したがって、妊娠の場合と潰瘍の場合では特徴「性別」の重要度を別に用意しなければならない。一方、特徴の重要度はカテゴリーに依存しているだけでなく、比較するカテゴリー間の関係にも依存している。たとえば、「乗り物の分類」において、「タイヤが二本」という特徴は「自転車」と「オートバイ」を区別する場合には重要でないが、「オートバイ」と「乗用車」を区別する場合には重要な特徴となる。

このような状況依存した重要度に対応するために、本章では2段階の連想度手法（以降では単に2段階法と呼ぶ）を導入する。これは、2.1節で示した医療における診断のプロセスで用いられる2段階のステップ（まず可能性のある病名を推定し、その結果に応じてさらに詳細な検査を行う）と同様に、まず第1段階で連想度による候補の絞り込みを行い、第2段階でさらに詳細な分類を行うものである。なお、2段階法の1段階目では拡張連想度手法と同一の手法を探っているが、ここでは事例を連想度が最大のカテゴリーに分類するのではなく、カテゴリー候補の絞り込みのみが行われる。2段階目では、絞られた複数のカテゴリー候補の中で重要度を再計算し、この重要度を用いて事例の再分類を行い、最も高い連想度を示したカテゴリーに分類するようしている³。

2段階法と拡張連想度手法の分類精度の比較結果を表3に、学習に用いる事例の増加による分類精度の変化を図4に示す。なお、比較に用いた重要度乗数は1段階目が1.7、2段階目が4.0である。

表3: 分類結果3: 適用例 144例

	1段階の連想度	2段階の連想度
各カテゴリーにつき1例（計17例）	74.3 %	75.3 %
各カテゴリーにつき数例（計145例）	96.3 %	97.7 %

結果としては、わずかながら（1.5%程度）2段階法の分類精度が向上している。この向上は誤差程度ではあるが、データをランダムに入れ替えて100回のシミュレーション（標準偏差が1.3%）を行った結果であり、誤差によって生じた差ではない。また、この結果を誤分類率で考えると、訓練例145例の場合に2段階法が2.3%、1段階では3.7%であり、誤分類の可能性が3分の2に減少していることがわかる。なお、この例では1段階目と2段階目の重要度乗数を変えている。これは、1段階目と2段階目の分

³2段階目の連想度算出アルゴリズムは1段階目と全く同一のものであるが、分類候補のカテゴリーが絞られているために、絞られたカテゴリー間で重要な特徴の重要度が大きくなる。

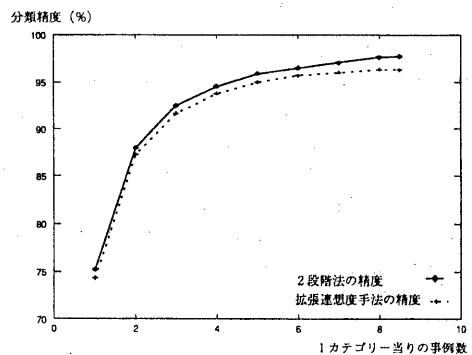


図4: カテゴリー当たりの学習事例の増加による精度の変化

類では性質が違うためである。この違いについて説明するために、3.2節で重要度の検証に用いた重要度乗数による精度変化の図に2段階法の結果を併せて図5に示す。

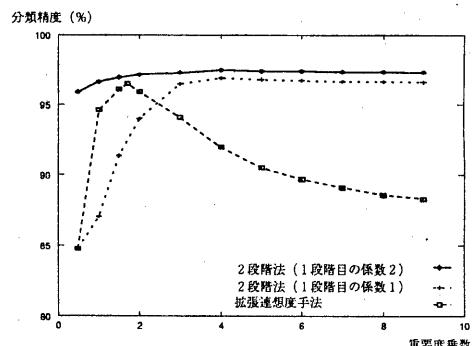


図5: 重要度乗数による分類精度の変化

図5の結果から、2段階法（1段階目の係数は1である）では、拡張連想度手法と比べて、2段階目の係数が2以下では分類精度が落ちていること、重要度乗数が大きくなても分類精度はほとんど低下していないことなどの特徴がある。これは、2段階法では重要な特徴を1段階目と比べてより強調しなければならないこと、また、2段階目ではあまり重要でない特徴は分類に影響していないことによる。すなわち、1段階目のおおまかな分類では、多くのカテゴリーを考慮するために重要度の小さな特徴も無視できないのに対し、2段階目の詳細分類では絞られた候補カテゴリー間でのみ分類が行なわれるために、重要度の低い特徴は逆に誤分類につながることを示している。したがって、2段階法では両段階とも同一の手法が用いられているが、分類の性質そのものは互いに異なったものである。

5 データ量と分類精度

分類性能の評価基準として、分類の精度とともにコストの問題が挙げられる。分類のコストには計算コストと記憶コストがあるが、典型性に基づく手法は冗長にデータを残すことによって高い分類精度を実現しているために、ルールに基づく手法に比べて、記憶コストがかかっていると考えられる。本章では、連想度のデータ量と分類精度の関係について検討し、さらにデータ量を削減した（削減する）連想度手法を提案する。なお以降では、簡単のために拡張連想度手法のみについて検討する。また誤解の生じない限り単に連想度手法と呼ぶ。

5.1 ルールに基づく手法との比較

連想度手法は、すべての特徴とカテゴリーの組について、特徴とカテゴリーの関連の度合い（連想度）を残している。このため、場合によってはデータ量が膨大となるという問題がある。たとえば、soy bean data はカテゴリー数 17、特徴数 203 からなるデータ集合であるために⁴、3,451 組の連想度が残されてしまう。仮に値が 0 の連想度を削除したとしても⁵、約 1,400 組の連想度の値を保持しなければならないことになる。連想度手法のデータ量がどれほど多いかについては直接的な判断ができないために、ルールに基づく代表的分類手法 ID3 および AQ との比較によってデータ量の問題を検討する。

連想度手法、ID3、AQ の 3 手法ともにデータ形式が異なるために直接的な比較は困難である。これは、連想度手法は連想度、ID3 は決定木、AQ はカテゴリー毎の分類ルールをデータとして利用しているためである。したがって、本節では各手法のデータ量を次のように定義して、それぞれのデータ量を算出している。連想度手法は一つの特徴とカテゴリーの組をデータ量 1 とする。決定木はノードの数を、AQ により導かれるルールはルールに現われるリテラルの数をそれぞれのデータ量とする。この基準にしたがって、各手法を図 2 で用いたデータセットに適用した結果、連想度手法が 1,389 データで分類精度 94.4 %、ID3 が 97 データで分類精度 79.2 %、AQ が 117 データで分類精度 83.1 % であった。この結果から連想度手法と他の 2 手法を比較すると、データ量が他の手法と比べて 10 倍強、誤分類率が 4 分の 1 から 3 分の 1 となっている。したがって、データ量と分類精度は直接的な比例関係にはないが、それぞれデータ量に応じた精度になっていると考えられる。

5.2 データ量の削減

連想度手法は、ルールに基づく手法に比べてかなり多くのデータを残す手法である。本節では、連想度手法のデータ量を削減することについて検討する。連想度手法は、特徴とカテゴリーの連想度の和によって分類の指標としているために、連想度の値が小さなデータは分類に与える影響が小さいとみなすと、それらのデータ

⁴ 実際のデータには 50 属性が含まれているが、本手法では属性の各属性値につき一個の特徴を持つように変形している。

⁵ 連想度 0 の特徴は分類に影響しないために保存しなくてもよい。

タを削除することができる。この削除によって得られた結果（連想度知識と呼ぶ）を図 6 に示す。なお、図 6 の連想度知識では重要度が 0.22 以下の属性を削除している。

属性	属性値	重要度
color of spot on reverse side	= none	0.23
leaf spot margins	= water soaked	0.42
leaf spot growth	= scattered with concentric rings	0.22
	= necrosis across veins	0.78
fruit spots	= colored spot	0.29

カテゴリー ALTERNARIA への連想度知識

属性	属性値	重要度
leaf spot color	= tan	0.37
leaf spot growth	= from edge of leaf inward	0.54

カテゴリー PHYLOSTICTA への連想度知識

図 6: 連想度知識の例

連想度知識は、各カテゴリー毎に典型的な特徴（属性と属性値の対）とその重要度の組からなる一種のルールである。soy bean data に対して図 6 の連想度知識を適用した結果、1,389 データから 72 データにデータ量が削減していること（約 20 分の 1）、分類精度は 94.4 % から 88.9 % に低下していることなどがわかった。さらに、図 6 の結果と 5.1 節の結果から、データ量を削減すると分類精度は若干低下するが、AQ や ID3 とはほぼ同一のデータ量で同等の分類精度を達成していることがわかる。

6 問題点と今後の課題

6.1 典型性の尺度を用いた他のアルゴリズムとの比較

拡張連想度手法と独立して開発された概念学習アルゴリズムとして PROTO-TO [6] がある。PROTO-TO と拡張連想度手法は、特徴の典型度と重要度の積を用いて分類を行なっている点が類似している。また、両者の相違点としては、PROTO-TO は属性ごとに、拡張連想度手法は属性値ごとに特徴の重要度を与えている点が挙げられる。このため、PROTO-TO の重要度では、属性値によって重要度が異なる特徴や、4 章で述べたように、比較するカテゴリーの組によって特徴の重要度が変化する領域では十分な分類精度が得られないと考えられる。逆に、拡張連想度手法では、属性値毎に重要度を与えてるために、直接的に実数値属性を扱うことができず、予めデータ集合の前処理を行なって適切に分割しておく必要があるという問題点がある。

拡張連想度手法と PROTO-TO を直接的に比較することはできないために、文献 [6] で PROTO-TO が用いたデータを用いて、拡張連想度手法と 2 段階法をほぼ同一の条件で実行した結果を表 5 に示す。また、適用したデータベース [9] と適用条件を表 4 に示す。

⁶ 分類を行なう場合に、拡張連想度手法は得られた連想度の値の和が最も大きくなるカテゴリーに、PROTO-TO では典型的なカテゴリーとの差の最も小さいカテゴリーに分類を行なうなど、分類に用いる式の細部は異なっている。

す。なお、PROTO-TO の分類結果と C4 (ID3 の改良アルゴリズム) の分類結果は文献 [6] から引用したものである。また、2段階法では、カテゴリー数が 3 個以上なければ絞り込むべきカテゴリーが存在しなくなるために、カテゴリー数が 2 個以下のデータベース (hepatitis, voting) については適用していない。

表 4: 適用したデータベースと適用条件

データベース名	訓練例	テスト例	属性	カテゴリー	特徴の欠落	実数値の属性
glass	107	107	9	6	あり	全ての属性
hepatitis	78	77	19	2	あり	一部あり
voting	218	217	16	2	あり	なし

表 5: 分類結果 4

データベース名	2段階法	拡張連想度	PROTO-TO	C4
glass	45 - 55 %	43 - 50 %	48.0 %	65.5 %
hepatitis	-	80 - 85 %	79.9 %	79.8 %
voting	-	90 - 94 %	90.4 %	95.3 %

表 5 の結果のうち、2段階法や拡張連想度手法の精度に幅があるのは、重要度乗数や実数値属性の分割の度合によって分類精度が変化しているためである。特に、glass のデータは全ての属性が実数値であるために、実数値の分割が分類精度に大きな影響を与えていている。したがって、適切な重要度定数や実数値の分割を決定するアルゴリズムの開発、あるいは拡張連想度手法や2段階法に実数を取り扱う能力を持たせることが必要であると考えられる。この二つの問題点に関しては後の節で詳細に検討する。

6.2 重要度基準

カテゴリーに対する特徴の重要度を示す基準は、記憶に基づく推論などでさまざまな基準が使われているが、基準は問題に依存するために、試行錯誤的に決定されている場合も多い。また、事例に基づく学習システム Bloom [7] の重要度のように、逐次的に重要度を少しづつ更新していくアルゴリズムや、専門家から知識獲得するシステム Protos [8] のように専門家の説明から発見的に重要度を獲得するアルゴリズムなどもある。

概念形成システム COBWEB [10] のカテゴリー分割基準に用いられている category utility は、特徴の評価の基準として属性予測度 (predictability) と帰属予測度 (predictiveness) の積を用いている。属性予測度は、事例のカテゴリーがわかっている場合にそれぞれの属性値がとりうる条件付き確率であり、帰属予測度は、ある属性値が観測された場合に事例が属しうるカテゴリーの条件付き確率である。したがって、属性予測度は本手法における特徴の典型性の度合と同じものであり、帰属予測度は拡張連想度手法の重要度に相当している。両者の相違点としては、拡張連想度手法の重要度が属性値ごとに算出されるのに対して、帰属予測度は属性値とカテゴリーの組に算出されるという点である。なお、両者を比較するために拡張連想度手法の重要度と帰属予測度をいれ

かえて分類を行なった結果を表 6 に示す。なお拡張連想度の重要度は、重要度乗数の調整を行なっていない場合 (重要度乗数 1.0) の分類精度を用いている。

表 6: 拡張連想度手法と帰属予測度の比較

データベース名	拡張連想度手法の重要度	帰属予測度
soy bean	94.6 %	88.1 %
hepatitis	83.5 %	78.5 %
voting	90.1 %	89.3 %

表 6 に示すように、拡張連想度手法の重要度は帰属予測度と比較して適切な値が与えられていると考えられる。

6.3 最適パラメータの獲得問題

拡張連想度手法は、重要度定数を変化させることによって、その領域に適切な感度に特徴の重要度を変化させることができある。しかしながら、この最適値は問題ごとに与える必要があり、その決定問題が逆に本手法の問題点ともなりうる。また、2段階法の適用についても、カテゴリーの絞りこみの度合などに適切な値を与えないといふいう問題がある。したがって、今後は、これらのパラメータの自動獲得を行なう機構の開発が必要である。現在検討している手法は、予め訓練例集合を二つに分けて一つを仮想的な訓練例集合、もう一つを仮想的なテストを行なう例の集合としておく。つぎに、仮想的な訓練例集合からパラメータを調整しながら試行錯誤的に連想度を求めて、仮想的な訓練例集合の分類を行ない、その分類精度の結果から適切なパラメータの値を発見的に獲得する手法である。この手法は、かなりの計算量を伴うと予想されるが、現在、パラメータの調整法などを検討中である。

6.4 実数値特徴の扱い

拡張連想度手法は、実数値による特徴をそのまま扱えないために、表 5 に示すように、実数の場合はその分割の基準⁷によって分類精度が大きく影響する。いくつかのヒューリスティックな基準を分割に用いて分類を行なったが、その基準は領域依存しており、適切な基準は得られない。前節で述べた最適パラメータを獲得する手法と同様なアルゴリズムで最適分割数を決定する手法もあるが、計算量的にも負担が大きいため、PROTO-TO が行なっているような直接的に実数値を扱える手法を検討中である。

6.5 学習に用いた例の再分類

ID3 や AQ のように、事例からルールを帰納するアルゴリズムは、訓練事例集合に矛盾がない限り、与えられた訓練事例は正しく分類されることを保証しているが、拡張連想度手法のアルゴリズムでは、必ずしも訓練事例を正しく分類することは保証されな

⁷ 本稿で用いている分割基準は分割数であり、ルールに基づく手法分類などで行なわれる最適分割点を求める基準ではない。

い。これは、拡張連想度手法は、訓練例平均的なカテゴリーの概念を生成するアルゴリズムであるため、例外的な訓練例は典型的な概念に吸収されていることによる。言い換えると、ルールを帰納するアルゴリズムは、全ての訓練例の分類を保証する代わりに、例外的な訓練例を分類可能とするために複雑なルールを生成してしまうという問題点がある。5.2 節の結果で、ID3 や AQ が比較的多くのデータ量を用いるにもかかわらず、削減された連想度知識と比べて分類精度が低いのは、この例外を説明するルールのためと考えられる。また、例外的な訓練例には、例外の場合とノイズの場合があり、例外的な事例の分類を保証することの適否は領域に依存している。すなわち、ノイズのない領域では、ID3 などのルール帰納のアルゴリズムが適切であり、ノイズのある領域では拡張連想度手法のような例外的な事例の影響を弱めた典型性を用いたアルゴリズムが適切である。また、拡張連想度手法を例外的な事例の分類も可能にすることは、次のようにアルゴリズムを拡張することによって実現できる。すなわち、拡張連想度手法において連想度を一旦学習し、その値を用いて訓練事例を全ての分類を行ない、分類を誤った場合に誤ったカテゴリーと正しいカテゴリーの違い (exemplar difference) を保存しておき、新たな例を分類する場合にはその違いの知識も利用する手法である。本稿で取り上げてている soy bean data では、誤分類が少ないためこの exemplar difference はあまり有効には動作しないが、今後は適用領域やさらに具体的な適用手法などを現在検討中である。

7 おわりに

本稿では、特徴の連想度を分類の指標とする典型性に基づく分類手法の提案した。連想度手法は、特徴とカテゴリーの関係の度合いを統計的に事例から獲得することによって、少ない訓練例から多くの情報を獲得し、かつ未学習のデータに対しても高い分類精度を実現している。また、複数カテゴリーを分類候補として導出するために、専門家の判断支援システムや事例ベース推論の検索アルゴリズムとして利用可能である。さらに、連想度手法を 2 段階に用いることにより、変化する重要度に対応した分類も可能となっている。

分類型問題の対象領域は広範囲に及ぶために、本手法は多くの領域で適用可能であると考えられるが、現状では検討した例題が少ないために、適用できる範囲やその条件などが完全に把握されているとは言い難い。また、実数値特徴の取り扱いや問題依存するパラメータの獲得などの問題点が残されている。今後は、これらの問題点を解決していくとともに、本手法を多くの例題に適用して適用条件や有効性を検証していくことが必要である。

謝辞

本研究の一部は文部省科学研究費重点領域研究（知識科学における概念形成と知識獲得）および平成 3 年度大川情報通信基金の援助による。本研究の初期段階では、東京工業大学 志村正道教授および東京電機大学 上野晴樹教授からの知的刺激が大きな励まし

となりました。また、慶應義塾大学の開一夫君にはいくつかの参考文献を教えて頂きました。あわせて感謝致します。

参考文献

- [1] Rosch, E.: Principles of Categorization, in E. Rosch and B.B.Lloyd (eds.) *Cognition and Categorization*, Erlbaum (1978).
- [2] Michalski, R.S. and Chilausky, R.L.: Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis, *International Journal of Policy Analysis and Information Systems*, Vol.4, No.2, pp.125-161 (1980).
- [3] Quinlan, J.R.: Induction of Decision Trees, *Machine Learning*, Vol.1, pp.81-106 (1986).
- [4] Nunez, M.: The Use of Background Knowledge in Decision Tree Induction, *Machine Learning*, Vol.6, pp.231-250 (1991).
- [5] Stanfill, C. and Waltz, D.: Toward Memory-Based Reasoning, *Comm. of ACM*, Vol.29, No.12, pp.1213-1228 (1986).
- [6] Maza, M.: A Prototype Based Symbolic Concept Learning System, *Proceedings of the Eighth International Workshop on Machine Learning*, pp.41-45 (1991).
- [7] Aha, D. W.: Incremental, Instance-based Learning of Independent and Graded Concept Descriptions, *Proceedings of the Sixth International Workshop on Machine Learning*, pp.387-391 (1989).
- [8] Porter, B. W., Bareiss, R. and Holte, R. C.: Knowledge Acquisition and heuristic classification in weak-theory domains. Technical Report AI-TR-88-96, Department of Computer Sciences, University of Texas at Austin (1989).
- [9] Murphy, P. M. and Aha, D. W.: UCI Repository of machine learning databases [Machine-readable data repository]. Irvine, CA: University of California, Department of Information and Computer Science (1992).
- [10] Fisher, D. H.: Knowledge Acquisition via Incremental Conceptual Clustering, *Machine Learning*, Vol.2, pp.139-172 (1987).
- [11] 谷澤正幸, 上原邦昭, 前川禎男: CBL と EBL を用いた体系的知識獲得, 第 43 回情報処理学会全国大会予稿集, 3D-8 (1991).
- [12] 谷澤正幸, 上原邦昭, 前川禎男: 事例に基づく学習とモデルベース推論に基づく知識獲得アルゴリズム, 第 36 回システム制御情報学会研究発表講演会, pp.617-618 (1992).