ファジィ帰納学習法の発展と今後の展開[†]

櫻井 茂明*

コンピュータ及びネットワーク環境が普及し、データの電子化が加速する一方、これらのデータに内在する知識を発見する帰納学習法の重要性が高まっている。本論文では、決定木形式の判断規則の集合を学習する代表的な帰納学習法 ID3 を中心として、提案されている他の帰納学習法及びファジィ帰納学習法を紹介し、各手法間の関連を明らかにする。また、関連技術の研究動向を紹介し、ファジィ帰納学習法の今後の研究の方向性を明らかにする。

1 はじめに

コンピュータ及びネットワーク環境の普及にともなう、大量のデータを電子的に蓄えることが可能になった。これらのデータの中には、重要な情報が埋もれていると考えられるものの、対象となるデータ量が多く、人手による分析は困難である。このため、蓄積されたデータは、必ずしも有効に活用されていない。

このような背景の下、コンピュータを利用することにより、データの中から有効な情報を引き出す技術として、大量のデータの中に内在する規則性を発見する帰納学習法の研究が活発化している。本論文では、帰納学習法、特に決定木形式の規則性を学習するタイプの帰納学習法を取り上げ、現在までの研究成果及びファジィ集合理論と組み合わされた帰納学習法の研究成果を振り返る。また、これらの概観及び関連技術の研究動向を踏まえた上で、ファジィ帰納学習法の今後の研究動向を述べる。すなわち、2 節において提案されている帰納学習法を説明し、3 節においてファジィ集合理論と関連する帰納学習法を説明し、4 節において現在の研究動向を述べるとともに、今後発展すると考えられる研究分野について述べる。最後に、5 節において本論文のまとめを述べる。

2 帰納学習法

1960年代に登場したエキスパートシステムは、分野の専門家の知識をコンピュータ上に実装することにより、コンピュータに専門家の代わりをさせるシステムである。エキスパートシステムが人間の専門家に匹敵する判断を行うには、専門家の知識を漏れなく、正し

く実装する必要がある。このような知識獲得は、問題ごとに行う必要がある一方、エキスパートシステムが構築されるにつれて、その獲得には、多大なる時間が必要になることが明らかとなった。このため、知識をいかにして専門家から獲得するかが問題となり、知識獲得ボトルネックと呼ばれるようになった。この原因は、コンピュータに記述すべき重要な知識を、専門家は通常無意識のうちに活用しており、そういった知識を書き下そうとしても、矛盾なく完全に記述することが難しいことに起因している。この問題に対して、状況が与えられれば、その状況を専門家がうまく処理できることに注目し、状況に関連するデータを多数集めることにより、その背後にある規則性を、機械が自動的に発見する帰納学習法[19]が研究されるようになった。

一般に帰納学習法は、入力として与えられるデータに、そのデータを判別する情報(分類クラス)が与えられる場合と与えられない場合で、観察からの学習(教師無し学習)と例からの学習(教師あり学習)とに分類することができる。前者の代表的な手法としては、概念クラスタリング[44]と呼ばれる手法が知られており、データの性質を表す複数の属性から構成される事例の集合を、その属性の値に基づいて、いくつかの部分集合(クラスター)に分割し、クラスターに共通する属性値に基づいた意味付けを行って、事例集合に内在する規則性を発見する。

一方、後者の学習法は、データの性質を表す複数の属性とデータを判別する分類クラスから構成される事例の集合を対象とする。これらの学習法は、学習される規則性の表現形式の違いにより、宣言的な判断規則の集合を学習する手法と木構造の判断規則の集合(決定木)を学習する手法に分類することができる。前者の学習法の代表的な手法としては、AQ アルゴリズム[23]が知られており、注目する分類クラスが与えられている事例を正事例、その他の分類クラスが与えられてい

[†] The past, the present, and the future of fuzzy inductive learning methods
Shigeaki SAKURAI

* (株)東芝 研究開発センター 知識メディアラボラトリー
Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation

る事例を負事例とし、正事例に基づいた条件の緩和と負事例に基づいた条件の制限を繰り返して、注目する分類クラスに関する規則性を発見する。この AQ アルゴリズムは、条件の緩和方法や条件の制限方法等に関して種々の改良が行われており、AQ11[24]、AQ14-NT[31]、AQ15[26]、AQ17-HCI[55]といった手法が提案されている。

本論文で主に取り上げる後者のタイプの学習法は、文献[14]に記載されている CLS をその起源としている。この CLS は指定した深さの木を多数生成し、その中から最も妥当な木を選択することにより、決定木を成長させる。このため、決定木の学習に膨大な計算量が必要であった。この計算量の問題に対して、指定した深さの木を多数生成する代わりに、情報量に基づいた属性選択を行いながら決定木を学習する ID3[32]が提案された。ID3は、決定木の最上位に位置付けられるルートノードを最初に生成し、与えられている事例集合の割り当てを行う。この事例集合に対して、事例の個数に基づいて各種確率を推定し、各属性が与えられる場合の相互情報量を計算する。すなわち、事例集合を S 、分類クラスが c_k となる S の部分事例集合を S_{c_k} 、属性 A_i の値が v_{ij} となる S の部分事例集合を $S_{v_{ij}}$ とすれば、 S において c_k となる確率 $p_{S_{c_k}}$ 、 S において v_{ij} となる確率 $p_{S_{v_{ij}}}$ 、 $S_{v_{ij}}$ において c_k となる確率 $p_{S_{v_{ij}}, c_k}$ は、式(1)～式(3)で計算される。ただし、記号 $|\cdot|$ は集合に含まれる事例の個数を求める演算を表す。

$$p_{S_{c_k}} = \frac{|S_{c_k}|}{|S|} \quad (1)$$

$$p_{S_{v_{ij}}} = \frac{|S_{v_{ij}}|}{|S|} \quad (2)$$

$$p_{S_{v_{ij}}, c_k} = \frac{|S_{v_{ij}} \cap S_{c_k}|}{|S_{v_{ij}}|} \quad (3)$$

この確率により、 S における情報量 $I(S)$ 、 S において属性 A_i が与えられる場合の条件付き情報量 $E(A_i, S)$ は、それぞれ、式(4)、式(5)で計算される。ただし、本式において、 C は取り得る分類クラスの集合、 $R(A_i)$ は属性 A_i が取り得る属性値の集合を表す。

$$I(S) = - \sum_{c_k \in C} p_{S_{c_k}} \log_2(p_{S_{c_k}}) \quad (4)$$

$$E(A_i, S) = - \sum_{v_{ij} \in R(A_i)} p_{S_{v_{ij}}} \sum_{c_k \in C} p_{S_{v_{ij}}, c_k} \log_2(p_{S_{v_{ij}}, c_k}) \quad (5)$$

また、相互情報量 $gain(A_i, S)$ は、式(6)で計算される。

$$gain(A_i, S) = I(S) - E(A_i, S) \quad (6)$$

次に、ID3は、相互情報量が最も大きくなる属性を選択し、その属性値が等しい事例ごとにまとめた部分事例集合を生成する。この部分事例集合ごとに、ノードを生成するとともに、元のノードと生成したノードとの間に、対応する属性値を割り当てた枝を生成する。このような事例集合の分割を、分割された事例集合に含まれる事例の個数が一定数以下になる場合、特定の分類クラスの出現する割合が一定値以上になる場合、等の条件を満たすまで、再帰的に繰り返すことにより決定木の生成を行う。

以上の ID3 には、実世界に存在する種々のデータを扱うには、いくつかの問題点があったものの、計算量が比較的少なく、高速に決定木を学習する実用的な手法であり、本手法の流れを汲む多数の手法が研究されている。

ID3の問題点を改良するという観点で、提案されている手法を分類してみると、INDECTS[3]、C4.5[33]、LMDT[7]は、属性値が離散的な値に限定されるという問題点を改良する手法となっている。INDECTS は、数値で与えられる属性(連続属性)に対して、分類クラスごとの平均値と分散に基づいた区間分割を行うことにより、数値を離散的な属性値に変換し、連続属性を扱うことを可能としている。これに対して、C4.5は、連続属性に対して、属性値の分布を2分割した場合に、その情報量が最も高くなる位置を探索し、その位置を境界とする区間分割を行うことにより、数値を離散的な属性値に変換し、連続属性を扱うことを可能としている。また、C4.5においては、属性値の数が多い属性が選択されやすいという問題点を改良するために、属性値の持つ情報量に基づいて相互情報量を補正した *gain_ratio* を属性選択の際に利用するといった改良も行われている。一方、LMDT は、複数の連続属性に対して、線形関数を構成し、線形関数を境界とする分割を行うことにより、数値を離散的な属性値に変換し、連続属性を扱うことを可能としている。LMDT は、INDECTS や C4.5 に比べて、連続属性に対して、適切な境界を設定できるものの、より大きな探索空間を探索する必要があるため、境界を発見するのに、多くの時間を必要とする。

これらの手法においては、必ずいずれかひとつの領域に事例を割り当てるクリスプな判定が行われており、境界付近において、判断誤りを起こしやすいという欠点があった。また、連続属性を扱えるように拡張がなされたものの、人間が主観的に与えた属性値を扱うことはできなかった。これらの問題に対して、ファジィ集合理論と帰納学習法を融合した手法も多数提案されているが、これらの手法の詳細に関しては、3節にお

いて説明する。

以上に述べた ID3 の拡張は、離散的な属性値しか扱えないという問題に対する拡張であった。これに対して、学習データが一括して与えられることを前提とし、バッチ的にしか学習できないという ID3 の問題点を克服する手法として、ID5R、ID3-IV-R [4] が提案されている。ID5R は、事例が入力されるごとに、今までに与えられている事例集合に当該事例を加えて、相互情報量の再計算を行う。この値が改良される場合には、決定木の当該部分の構造を変化させることにより、漸進的な決定木の更新を可能としている。この ID5R では、相互情報量の再計算を行うために、学習に利用した事例をすべて保存しておく必要がある。一方、ID3-IV-R は、学習に利用した事例を保存しない代わりに、学習済みの決定木から生成される可能性のある属性値を決定し、現在の決定木だけでは決定できない属性値を、乱数を用いて決定することにより、事例集合を再構成する。この再構成した事例集合と新たに追加された事例を用いることにより、決定木を再学習する。このため、ID3-IV-R は、決定木の再構成に必要な事例を少なくできる反面、決定木の再構成に多くの時間を必要とする。

この他の ID3 の問題点として、単純な構造を持ったルールしか学習できないという問題に対して、複数の分類クラスを持った事例集合から決定木の一つである花つき回帰木を学習する手法 [27]、背景知識を利用することにより、構造を持った属性値を記述可能にした決定木を生成する EG2 [28] 等、多数の手法が提案されている。

3 ファジィ帰納学習法

3.1 ファジィ決定木

収集される数値にはノイズが含まれることが多く、境界付近における判断は、不確実な判断になりがちである。しかしながら、クリスプな境界に基づいた二者択一的な判断では、この不確実さを表現することができない。また、「暑い」、「寒い」や「若い」、「年取った」等の人間の主観に依存する言葉の場合、その境界をクリスプに定義すること自体が困難である。このような問題に対処するため、属性の表現にファジィ集合を持ち込んだファジィ決定木 [35] [49] が提案されている。

ファジィ決定木はファジィルールの集合を木構造に表現したものであり、従来の決定木と同様に、ノードとノード間を結ぶ枝から構成される。ノードには分岐ノードと末端ノードの 2 種類のノードが存在し、分岐ノードには、当該分岐ノードに到達した事例を分類するための属性が割り当てられ、下位方向に属性を表現するファジィ集合の数に対応した枝が存在する。また、

最上位に位置するノード(ルートノード)以外のノードに関しては、上位方向にひとつの枝が存在する。一方、末端ノードには、分類クラスと確信度を組にした確信度付きの分類クラスが割り当てられており、上位方向にひとつの枝が存在する。また、各枝には、連結する上位のノードに割り当てられている属性に対応して、ひとつのファジィ集合が割り当てられている。このファジィ決定木の表現においては、ルートノードから分岐ノードを経由して到達する末端ノードまでがひとつのファジィルールに対応する。

以上の表現を持つファジィ決定木の一例を図 1 に示す。図 1 においては、「温度」及び「湿度」を判断することにより、現在の状況が「不快」であるか「快適」であるかを判断するファジィルールの集合を表しており、「温度」及び「湿度」が属性、「不快」及び「快適」が分類クラス、「寒い」、「適当」、「暑い」が属性「温度」を表現するファジィ集合、「普通」及び「高い」が属性「湿度」を表現するファジィ集合になっている。これらのファジィ集合は、図中に示すようなメンバーシップ関数により定義される。この図のファジィ決定木は、5 つのファジィルールを表現しており、ルートノード N_1 から末端ノード N_5 までが、「温度」が「寒い」、「湿度」が「普通」ならば、「確信度 0.5 で不快、確信度 0.5 で快適」といったファジィルールを表現している。

3.2 ID3 の問題点による分類

ファジィ決定木を学習する多くのファジィ帰納学習法は、ID3 をその原点としている。本論文では、ファジィ帰納学習法を、ID3 に内在する問題点の克服という観点から分類し、各手法を紹介する。

ID3 の一番の問題は、利用可能な値が離散値に限定されることである。この問題を克服する手法は、属性及び分類クラスのどちらを拡張するか、ファジィ集合を手動及び自動のどちらで決定するか、というふたつの観点から分類することができる。

IDF [35]、ファジィ ID3 [49] は、属性、手動に分類され、数値やファジィ値を属性値として扱うことを可能としている。また、これらの値を利用可能とする属性を表現するファジィ集合は、利用者が手動で与えることにしている。これらの手法は、このファジィ集合に基づいて事例集合をファジィ分割しつつ、ファジィ決定木を成長させる。このとき、属性の選択基準としては、相互情報量を利用するものの、ID3 が事例の個数に基づいて相互情報量を計算するのに対して、各ファジィ集合に対する帰属度に基づいて相互情報量を計算する。

IDF 及びファジィ ID3 においては、ファジィ集合は手

図1 ファジィ決定木

動で設定されるものの、適切なファジィ集合を予め設定することは難しく、試行錯誤を通したファジィ集合の調整が必要である。この問題に 대응、属性、自動に分類される手法として、ファジィ集合を自動的に生成する改良 IDF[39]、ファジィ集合を自動的に調整するチューニング型ファジィ ID3[50] 及び学習型ファジィ ID3[12] といった手法が提案されている。改良 IDF においては、属性の選択に先立って、属性を表現するファジィ集合を自動獲得する。本手法においては、分類クラスの分離という観点で、十分な分離が行われていない属性の領域を再帰的に取り出し、細分化していくことにより、台形型メンバーシップ関数で記述されたファジィ集合を生成する。次いで、AIC を基準として、隣接するファジィ集合を統合していくことにより、データへの過度の依存を避けた、適切な数のファジィ集合を生成する。一方、チューニング型ファジィ ID3[50] では、メンバーシップ関数を、Z 関数、II 関数、S 関数で定義するとともに、ファジィ ID3 において行われていた代数積に基づく帰属度計算を、Dombi のパラメータ付き t -norm に変更する。これらの関数や演算に含まれるパラメータの値を、属性の選択に先立って、最急降下法を用いて調整することにより、適切なファジィ集合の調整を可能としている。また、学習型ファジィ

ID3[12] では、各分岐ノードで利用する帰属度演算を、パラメータ付き t -norm に変更し、そのパラメータを調整する。この調整を行っても相互情報量の値が 0 以下になる場合に、背景知識として定義される上位概念を表すファジィ集合を加えた相互情報量の再計算を行う。この再計算を行っても、相互情報量が正数にならない場合に、当該ノードを分類不能なノードとすることにより、適切なファジィ集合の調整を可能としている。

このような属性に対する拡張とは別に、制御問題への適用を意識して、分類クラス、自動、に分類される手法として、連続型 IDF [38] [41]、ファジィ c 平均法に基づいた手法[46]、スプライン関数に基づいた手法[17]、射影追跡 ID3[25]、が提案されている。連続型 IDF は、数値として表される分類クラス(連続クラス)に対して、AIC を評価基準とするクラスタリングを行い、クラスター平均を中心とする三角型メンバーシップ関数を定義し、分類クラスを表現するファジィ集合を生成する。このファジィ集合を利用して、分類クラスを確信度の付いたラベルに変換し、改良 IDF に基づいてファジィ決定木を学習する。次いで、末端ノードに割り当てべき数値を、末端ノードに伝播した事例の分類クラス及び当該ノードにおける確信度に基づい

て推定する。一方、ファジィ c 平均法に基づいた手法は、はじめに、連続クラスに対して、ファジィ c 平均法に基づいたクラスタリングを行う。次に、各連続クラスを、生成されたクラスターとそのクラスターに含まれる確信度に置き換え、ファジィ決定木を学習する。また、スプライン関数に基づいた手法は、属性を表現するファジィ集合のメンバーシップ関数を、 B -スプライン関数を用いて定義し、暫定的に定めた離散的な分類クラスを用いて、ファジィ決定木を学習する。このファジィ決定木をファジィルールの集合とみなして、ニューラルネットワークの学習則を適用することにより、結論部に割り当てるべき数値を学習する。さらには、射影追跡 ID3 は、連続クラスの誤差の 2 乗和が最小となるような属性の射影を求め、分岐ノードに割り当てる新たな属性を生成する。このような属性の割当を、誤差の 2 乗和がしきい値以下になるまで繰り返すことにより、ファジィ決定木を学習する。

以上の離散値しか扱えないという第一の問題点に加えて、ID3 には、欠損値が扱えない、バッチ的な学習しかできない、単純な構造を持ったルールしか学習できない、属性値の多い属性を選択しやすい、といった問題点がある。以下においては、これらの問題点を克服する手法に分類される手法を紹介する。

欠損値の問題とは、利用される事例には、必ずしもすべての属性値が与えられていない場合を想定した問題である。この問題に分類される手法として、*unknown* に基づいた手法[36]が提案されている。*unknown* に基づいた手法は、任意のファジィ集合との帰属度が 1 となるメンバーシップ関数を持ったファジィ集合 *unknown* を、欠損値を表すファジィ集合として定義する。このような定義を行うことにより、属性値が与えられていない属性の判断を留保し、下位に割り当てられている属性に判断をまかせるといった効果を得ることができる。

バッチ学習の問題とは、学習を開始するのに先立って、十分な事例が収集されていない場合を想定した問題である。この問題に分類される手法として、洗練型 IDF[40]が提案されている。洗練型 IDF は、学習時に伝播した事例が少なかった末端ノードを、監視対象ノードとし、学習時に利用した事例を廃棄せずに、当該ノードに格納する。また、この監視対象ノードに伝播した評価対象に対して、他の評価方法に基づいて、真の分類クラスを決定し、この評価対象の属性値と真の分類クラスからなる新たな事例を生成する。さらには、学習時に与えられなかった離散値や学習時に与えられた範囲を超えた数値が、属性値として与えられる場合にも、新たな事例を生成し、このような状況が発生し

た分岐ノードに、その事例を格納する。これらのノードに対して、個別に改良 IDF を適用することにより、部分的にファジィ決定木を再構成する。この洗練型 IDF は、精度が悪く考えられるファジィ決定木の一部分を動的に改良することにより、漸進的な学習を可能としている。

単純なルールの問題とは、ルールの各条件が必ずしも「and」で結ばれていない、各条件が必ずしも固定的な値として定義されていない、等の複雑な構造を持ったルールを想定した問題である。この問題に分類される手法として、OR 型演算に基づいた手法[13]、探索型 IDF[42]が提案されている。OR 型演算に基づいた手法は、従来の AND 型演算子に加えて、分岐ノードにおける演算子に、パラメータ付きの *t-conorm* を導入した手法である。各分岐ノードにおいては、パラメータ付きの *t-norm* 及び *t-conorm* のパラメータ調整を含めて、相互情報量が最大となる属性及び演算子を選択する。また、*t-conorm* が選択される場合には、当該属性をノードに持つサブファジィ決定木を生成する。OR 型演算に基づいた手法は、条件部の記述に AND 型のみならず、OR 型の演算子を許すため、ファジィ決定木の条件部を柔軟に記述することができ、コンパクトなファジィ決定木を構成することができる。一方、探索型 IDF は、パラメータの値によってその値が変動する属性値を対象とする手法である。探索型 IDF は、決定するパラメータの組を染色体表現し、ファジィ決定木の精度及びサイズに基づいた適合度で評価することにより、GA に基づいてパラメータの値を探索する。また、パラメータの値を固定して、属性値を決定し、改良 IDF を用いてファジィ決定木を学習する。このように、GA と改良 IDF を繰り返し利用することにより、パラメータの組とファジィ決定木を同時に学習することができる。探索型 IDF は、画像処理関数のパラメータを調整しつつ、画像に内在するパターンの特徴量を抽出して、診断を行う問題向けのルールの学習に適用することができる。

属性値が多い属性を選択しやすい問題とは、相互情報量に基づいた属性選択では、属性値の数が多い属性の方が、属性値の数が少ない属性よりも選択されやすいという現象に起因する問題である。この問題に分類される手法として、ファジィ C4.5[15][16]が提案されている。ファジィ C4.5 は、C4.5 に導入された *gain_ratio* を、帰属度をベースに計算するようにした手法である。また、パラメータの値に基づいて、属性の重みを考慮できる *gain_ratio* のさらなる拡張を行った選択基準も導入しており、人の直感に合った属性選択を可能としている。

図2 手法の関連

以上に分類してきた各ファジィ帰納学習法の関連を、前節に示したクリスプな帰納学習法との関連と合わせて、図2に示す。図においては、実線が始端の手法を終端の手法に改良することを示しており、その改良点を実線上に記述されている。また、破線が始端の手法を終端の手法で吸収することを示している。

最後に、ID3における問題点とは直接関係しないものの、ファジィ帰納学習法を理論的に解析していく上で必要となる関連研究を紹介する。論文[17]は、ファジィ集合の直積をメンバーシップ関数の積として定義し、各ファジィ集合に対するメンバーシップ関数の合計値が1になる場合に、相互情報量が非負になることを示している。また、論文[13]は、パラメータ付きの t -norm を利用する場合には、相互情報量が単峰性の分布になることを示している。一方、従来の帰納学習の分野においては、PAC 学習といった枠組の下、特定の確率モデルを過程することにより、ある精度以上の判断規則の集合を学習するには、どの程度の数の事例が必要で

あるかを、簡単な問題に対して示している。このような研究がなされているものの、ファジィ帰納学習法を理論的に解析する道具立ては、必ずしも十分に揃っておらず、さらなる発展が期待される。

4 今後の研究動向

これまでの節において、決定木形式の規則性を学習する ID3 を中心として、実問題に適用する際に発生する問題に打ち勝つ手法を概観した。本節においては、これらの手法に関連する最近の話題に触れるとともに、決定木形式の規則性を学習するタイプ以外の方法に関しても若干触れ、今後の研究動向を予測する。

決定木形式の規則性を発見する大部分の手法は、計算量の問題から数千程度の規模の事例集合を対象としている。これに対して、近年、データベースに格納されるデータは、数百万から数千万にも達しており、すべてのデータを対象とするのはなかなか難しい。この問題に対して、OLAP ツールは、GUI を通して、指定

した属性でのデータの絞り込みや、属性値の出現頻度の分布をグラフ化することにより、人間による大規模データ絞り込みの支援を可能としている。また、パソコンの低価格化にともなって、数十台、時には数千台のパソコンを並列に結合した PC クラスタシステムが登場するようになり、各パソコンに部分問題を解かせて結果を集約し、大規模データを扱う方法も研究されている。この他、計算量自体を少なくすることを目指した手法として、関連のある属性同士の組を記述する相関ルールを発見するアプリアリアルゴリズム[1]が提案されており、本手法を拡張した手法も多数提案されている。今後、個人レベル、個別業務レベル、各社レベル等、種々のレベルで、情報の電子化が進むと考えられ、大規模化の流れは、より一層進展すると考えられる。このため、大規模データを対象とする学習研究は益々重要になると考えられる。

決定木形式の判断規則の集合の利点として、人間による可読性の良さがあげられる。しかしながら、事例及び属性の規模が大きくなるにつれて、学習される決定木のサイズも大きくなる傾向にあり、学習結果を直感的に理解することが難しくなっている。一方、3次元可視化ツールの発展にともなって、リアルタイムに種々の角度からデータを見ることが可能になっており、これらのツールを利用した知識発見法に関する研究が盛んになっている。今後、情報の大規模化にともなって、学習結果の大規模化が進むと考えられる一方、学習結果に直接触れる利用者の増加にともなって、より分かりやすく結果を表示する必要性が高まると考えられ、学習結果の可視化に関する研究の重要性が一層増すと考えられる。

以上の大規模化に関連した流れとは別に、より一層の精度向上を目指す動きとして、学習法は従来のものを利用し、その前処理を工夫することにより、精度の向上を目指す手法が提案されている。属性集合の中から重要な属性の集合を抽出する属性選択は、いくつかの手法[22][30]が提案されており、この属性選択により精度が向上することが確認されている。また、事例集合に対するランダムサンプリングを行って事例を入れ替えながら複数回の学習を行う bagging や、誤った分類を行った事例の重みを調整しながら複数回の学習を行う boosting、といった手法[56]も提案されており、これらの手法を利用することにより、精度が向上することも確認されている。一方、決定木形式の帰納学習法とは別に、各事例を本来の事例の次元よりも高次元空間上に配置し、分類クラスの分離を最もよくする超平面を決定する SVM[2] にも注目が集まっており、決定木形式の帰納学習法よりも精度の高い分類を行える

といった結果が報告されている。このような精度向上を目指した研究は、今後も続くと考えられるものの、あらゆる対象に対応する一般的な手法を発見することは難しく、対象に応じた手法を選択し、対象に応じた工夫を行うことが重要になると考えられる。

関係データベースの発展にともなって、関係データベースに格納されるデータから規則性を発見するニーズが高まり、その要素技術のひとつとして、帰納学習法が利用されるようになった。一方、関係データベースでは扱うことができない形式のデータも多数蓄積されるようになり、より柔軟な形式のデータから、知識を発見したいというニーズが高まっている。そのひとつとして、テキストデータを分析するテキストマイニング[9]が研究されている。テキストマイニングにおいては、自然言語処理技術と帰納学習法を融合した手法が比較的多く提案[47][43]される一方、自己組織化マップを利用して、テキストデータを可視化する手法も提案されている。営業担当者の日報、コールセンターに寄せられる顧客の声、等のテキストデータを分析したいというニーズは、ナレッジマネジメントの重要性が認識されるにしたがって高まっており、今後この分野の研究が一層活発化すると考えられる。

また、ネットワーク環境の発展にともなって、各種レベルの Web ページがネットワーク上に乱立するようになっており、所望の情報やサービスがネットワーク上にあるとしても、その情報やサービスを利用することが難しいという現象がしばしば発生している。今後の利用者増加にともなって、利用者の平均的なレベルは低下すると考えられ、この現象は一層顕著になると考えられる。このため、知的な Web の実現を指向した Semantic Web[52]が検討されはじめており、Semantic Web に関連する技術として、XML データからのマイニング[54][57]等の研究が行われている。帰納学習法は、Semantic Web を実現する有望な要素技術のひとつであり、この方面での利用が今後活発化すると考えられる。

この他、マイニング結果に対する人間によるフィードバックを含めたアクティブマイニング[20][29]や、ゲノム情報等の医療分野への帰納学習法の適用も進みつつあり[18][48]、これらの研究も加速すると考えられる。

5 まとめ

コンピュータ及びネットワーク環境の普及にともなって、利用可能なデータが急速に蓄積されている。今後、この傾向は益々顕著になることが予想されており、これらのデータを利用するニーズも高まっている。本

論文では、これらのニーズに応える要素技術として、その研究が加速している帰納学習法、特に、ファジィ帰納学習法を概観すべく、ID3系の各手法の関連を明らかにした。また、ファジィ帰納学習法に関連した研究動向を紹介するとともに、今後研究が加速すると予想される分野を紹介し、今後の研究の方向性を明らかにした。

参 考 文 献

- [1] R. Agrawal and R. Strikant: "Fast algorithms for mining association rules", Proceedings of the 20th VLDB Conference, 487-499 (1994).
- [2] N. Cristianini and J. Shawe-Taylor: "An introduction to support vector machine and other kernel-based learning methods", Cambridge (2000).
- [3] 荒木 大, 小島 昌一: 「数値データによる決定木の帰納学習」, 人工知能学会誌, 7, 6, 60-68 (1992).
- [4] 浅見 徹, 山口 典男, 松本 一則, 鶴木 八寿, 橋本 和男: 「入力情報からの特徴抽出過程における専門家知識の最適化手法について」, 信学技報, 92, 84, 7-12 (1992).
- [5] J. Cheng, U. M. Fayyad, K. B. Irani, and Z. Qian: "Improved decision trees: A generalized version of ID3", Proceedings of the 5th International Conference on Machine Learning, 100-108 (1988).
- [6] J. Cendrowska: "PRISM: An algorithm for inducing modular rules", Knowledge-Based Systems, 1, 255-276 (1988).
- [7] B. A. Draper, C. E. Brodley, and P. E. Utgoff: "Goal-directed classification using linear machine decision trees", IEEE Transactions on Pattern Analysis and Machine Intelligence, 16, 9, 888-893 (1994).
- [8] Eui-Hong (Sam) Han and G. Karypis: "Centroid-based document classification: Analysis and experimental results", Proceedings of the 4th European Symposium, PKDD2000, 424-431 (2000).
- [9] R. Feldman and H. Hirsh: "Mining text using keyword distributions", Journal of Intelligent Information Systems, 10, 281-300 (1998).
- [10] R. Ghani, R. Jones, D. Mladenic, K. Nigam, and S. Slattey: "Data mining on symbolic knowledge extracted from the Web", Proceedings of the KDD-2000 Workshop on Text Mining (2000).
- [11] J. Grefenstette: "Genesis: A system for using genetic search procedures", Proceedings of a Conference on Intelligent Systems and Machines, 161-165 (1984).
- [12] 林 勲, 小澤 順: 「背景知識を用いた AND/OR 演算子学習型ファジィ ID3」, 第11回ファジィシステムシンポジウム, 137-140 (1995).
- [13] 林 勲, 小澤 順: 「学習型ファジィ ID3の高速化に関する一考察」, 第12回ファジィシステムシンポジウム, 447-450 (1996).
- [14] E. B. Hunt, J. Marin, and P. J. Stone: "Experiments in induction", New York Academic Press (1966).
- [15] 堀 和憲, 馬野 元秀, 佐藤 浩, 宇野 裕之: 「ファジィ決定木生成法ファジィ C4.5とその改良」, 第15回ファジィシステムシンポジウム, 333-336 (1999).
- [16] 堀 和憲, 馬野 元秀, 瀬田 和之, 宇野 裕之: 「ファジィ決定木生成法ファジィ C4.5とその改良(その2)-節点に応じて修正利得を変化させる方法-」, 第17回ファジィシステムシンポジウム, 739-742 (2001).
- [17] 市橋 秀友, 三好 哲也, 長坂 一徳, 白井 正志: 「ファジィ ID3を用いた Bernstein 多項式による多次元写像の近似法」, 日本経営工学会誌, 47, 2, 69-76 (1996).
- [18] S. Imoto, T. Goto, and S. Miyano: "Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression", Proceedings of the Pacific Symposium on Biocomputing 2002, 175-186 (2002).
- [19] 人工知能学会編集: 「人工知能ハンドブック」, オーム社, 67-97 (1990).
- [20] 北村 泰彦: 「アクティブ情報収集システムに関する検討」, 人工知能学会研究会資料, SIG-FAI/KBS-J-14, 85-90 (2001).
- [21] 「KINOsuite_IDTF ホームページ」, <http://www2.toshiba.co.jp/datamining/>.
- [22] H. Liu and H. Motoda, eds.: "Feature extraction construction and selection: A data mining perspective", Kluwer Academic (1998).
- [23] R. S. Michalski: "On the quasi-minimal solution of the general covering problem", Proceedings of the 5th International Symposium on Information Processing, A3, 125-128 (1969).
- [24] R. S. Michalski and R. L. Chilausky: "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, 4, 2, 125-161 (1980).
- [25] 三好 哲也, 市橋 秀友, 下村 武: 「射影追跡 ID3 による感性データ処理」, 第13回ファジィシステムシンポジウム, 825-826 (1997).
- [26] R. S. Michalski, I. Mozetic, J. Hong, and N. Lavrac: "The multi-purpose incremental learning system AQ15 and its testing application to three medical domains", Proceedings of the AAAI-86, 1041-1045 (1986).
- [27] 長木 悠太, 鈴木 英之進: 「花つき木構造による多次元クラスデータからの回帰学習」, 第14回人工知能学会全国大会, 323-326 (2000).
- [28] M. Núñez: "The use of background knowledge in decision tree induction", Machine Learning, 6, 3, 231-250 (1991).
- [29] 岡田 孝: 「アクティブマイニングのためのルール群表現法」, 人工知能学会研究会資料, SIG-FAI/KBS-J-36, 219-224 (2001).
- [30] 折原 良平: 「条件付確率に基づく属性選択」, 第12回人工知能学会全国大会, 19-21 (1998).

- [31] P. W. Pachowicz and J. Bela: "Improving recognition effectiveness of noisy texture concepts", Proceedings of the 8th International Workshop on Machine Learning, 625-629 (1991).
- [32] J. R. Quinlan: "Induction of decision trees", Machine Learning, 1, 71-99 (1985).
- [33] J. R. Quinlan: "C4.5: Programs for machine learning", Morgan Kaufmann (1992).
- [34] A. D. Shairo: "Structured induction in expert systems", Addison-Wesley (1987).
- [35] 櫻井 茂明, 荒木 大: 「ファジィ理論を適用した知識獲得」, 第15回知能システムシンポジウム, 169-173 (1992).
- [36] 櫻井 茂明, 荒木 大: 「ファジィ決定木生成アルゴリズムによる未知データの取り扱い」, 第46回情報処理学会全国大会, 2, 71-80 (1993).
- [37] 櫻井 茂明, 荒木 大: 「帰納学習によるファジィ決定木の生成」, 電学論 C, 113, 7, 488-494 (1993).
- [38] 櫻井 茂明, 荒木 大: 「ファジィ決定木を用いた数値予測」, 知識のリフォーメーションシンポジウム, 71-80 (1993).
- [39] 櫻井 茂明, 荒木 大: 「ファジィ帰納学習アルゴリズムの改良」, 電学論 C, 116, 9, 1057-1063 (1996).
- [40] 櫻井 茂明: 「ファジィ帰納学習における知識の洗練」, 電学論 C, 117, 12, 1833-1839 (1997).
- [41] 櫻井 茂明: 「ファジィ帰納学習による数値予測規則の学習」, 電学論 C, 118, 9, 1369-1375 (1998).
- [42] 櫻井 茂明: 「探索型ファジィ帰納学習によるパラメータ付きファジィルールの学習」, 電学論 C, 119, 10, 1180-1187 (1999).
- [43] S. Sakurai, Y. Ichimura, A. Suyama, and R. Orihara: "Acquisition of a knowledge dictionary for a text mining system using an inductive learning method", Proceedings of the IJCAI 2001 Workshop on Text Learning: Beyond Supervision, 45-52 (2001).
- [44] 榎木 哲夫: 「概念クラスタリング」, 日本ファジィ学会誌, 8, 3, 463-467 (1996).
- [45] 末田 直道, 櫻井 茂明, 島田 毅, 當田 遂充, 渡辺 貞一: 「事例ベース推論を用いた画像処理プログラム自動合成システム」, 電学論 C, 121, 4, 742-747 (2001).
- [46] 高木 浩, 亀井 且有: 「ID3の概念を用いた制御ルール獲得手法の提案とその応用」, 第13回ファジィシステムシンポジウム, 593-596 (1997).
- [47] P. -N. Tan, H. Blau, S. Harp, and R. Goldman: "Textual data mining of service center call records", Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, 417-423 (2000).
- [48] 津本 周作: 「データマイニングにおけるメタ分析」, 人工知能学会研究会資料, SIG-FAI/KBS-J-43, 261-266 (2001).
- [49] 馬野 元秀, 岡本 宏隆, 田村 坦之, 鳩野 逸生, 河内 二三夫, 梅津 祐久, 木下 淳一: 「ID3に基づくファジィ・ルールの抽出とその推論方法」, 第9回ファジィシステムシンポジウム, 857-860 (1993).
- [50] 馬野 元秀, 北辻 佳恵, 鳩野 逸生, 田村 坦之: 「チューニング型ファジィ ID3によるファジィ決定木の生成」, 第12回ファジィシステムシンポジウム, 451-454 (1996).
- [51] U. Fayyad, G. G. Grinstein, and A. Wierse, eds.: "Information visualization in data mining and knowledge discovery", Morgan Kaufmann (2001).
- [52] 浦本 直彦: 「Semantic Web -機械のための Web-」, 人工知能学会誌, 16, 3, 412-419 (2001).
- [53] P. E. Utgoff: "Incremental induction of decision trees", Machine Learning, 4, 161-186 (1989).
- [54] T. Miyahara, T. Shoudai, T. Uchida, K. Takahashi, and H. Ueda: "Discovery of frequent tree structured patterns in semistructured web documents", Proceedings of the PAKDD 2001, 18, 47-52 (2001).
- [55] J. Wnek and R. S. Michalski: "Hypothesis-driven constructive induction in AQ17-HCI: A method and experiments", Machine Learning, 14, 139-168 (1994).
- [56] B. Ženko, L. Todorovski, and S. Džeroski: "A comparison of stacking with MDTs to bagging, boosting, and other stacking methods", Proceedings of the ECML/PKDD01 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning, 163-174 (2001).
- [57] L. Xiao and D. Wissmann: "Information extraction from HTML: Combining XML and standard techniques for IE from the Web", Proceedings of the IEA/AIE 2001, 165-174 (2001).

(2002年1月15日 受付)

(2002年7月20日 採録)

[問い合わせ先]

〒212-8582 神奈川県川崎市幸区小向東芝町1

(株)東芝 研究開発センター 知識メディアラボラトリ

櫻井 茂明

E-mail: shigeaki.sakurai@toshiba.co.jp

TEL: 044-549-2398

FAX: 044-520-1308

著者紹介

さくらい しげあき
櫻井 茂明 [正会員]

1989年東京理科大学理学部応用数学科卒、1991年同大学院修士(数学)課程了。同年、㈱東芝入社。1998年新情報処理開発機構つくば研究センタ出向。2000年㈱東芝研究開発センター帰任。現在、同社知識メディアラボラトリーにて、テキストマイニングシステムの研究開発に従事。電気学会、人工知能学会、日本ファジィ学会、日本感性工学会各会員、博士(工学)。

The past, the present, and the future of fuzzy inductive learning methods

by

Shigeaki SAKURAI

Abstract :

The development of computers and network environments has been accompanied with the increase of data stored electrically. The importance of inductive learning methods, which discover knowledge from data, is more increasing. The paper focuses on a representative inductive learning method, ID3, which acquires a rule set described by a decision tree, and discusses relations between ID3, other inductive learning methods, and fuzzy inductive learning methods. Lastly, the paper introduces current topics in related study fields and shows future directions of fuzzy inductive learning fields.

Contact Address : Shigeaki SAKURAI

*Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, Kanagawa 212-8582, Japan*

TEL : 044-549-2398

FAX : 044-520-1308

E-mail : shigeaki.sakurai@toshiba.co.jp