様式 C-19

科学研究費補助金研究成果報告書

平成２２年３月３１日現在

機関番号：21602
研究種目：基盤研究（C）
研究期間：2007～2009
課題番号：１９５００１２８
研究課題名（和文）パターン間類似度に基づく不完全
データからの知識学習と理解に関する研究
研究課題名（英文）Knowledge learning and understanding from
incomplete data based on pattern similarity
研究代表者
趙　強福（Qiangfu Zhao）
会津大学・コンピュータ理工学部・教授
研究者番号：９０２６０４２１

研究成果の概要（和文）：理解できる知識を獲得するために、われわれはパターン間類似度に基づく多変量決定木の一種である最近傍識別木(NNC-Tree)とその構築方法を提案した。本研究の貢献は主に３つある。それは、（１）学習しながら重要特徴を選択する方法；（２）学習アルゴリズムが統一化できるデータのファジィ化方法；（３）高次元のデータを効率的に低次元に圧縮する方法、である。これらの方法を結合することによって、より効率的に多変量決定木を構築することができる。

研究成果の概要(英文)：To acquire understandable knowledge through machine learning, we have proposed the nearest neighbor classification tree (NNC-Tree) and an induction method. NNC-Tree is a kind of multivariate decision trees based on pattern similarity. In this project, we have proposed (1) a method for selecting important features through learning; (2) a method for unifying the learning algorithms through data fuzzification; and (3) a method for efficient dimensionality reduction. With these new contributions, we can induce multivariate decision trees more efficiently.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,400,000	420,000	1,820,000
2008 年度	700,000	210,000	910,000
2009 年度	700,000	210,000	910,000
総　計	2,800,000	840,000	3,640,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：学習と発見

１．研究開始当初の背景

知識の学習と理解を同時に行うために、われわれは最近傍識別木(NNC-Tree: Nearest Neighbor Classification Tree)とその構築方法を提案した(Q. F. Zhao, "Inducing NNC-Trees with the R^4-rule," IEEE Trans. on SMC-B, Vol. 36, No. 3, pp. 520-533, 2006)。NNC-Tree を使用することによって、従来の単一変量決定木よりも理解しやすいルールを抽出することができ、しかも、データ量が大きいときに、より高い認識率が得られる。しかし、これまで提案した構築法では、パターン間類似度を求めるために、ユークリッド距

離を利用していたので、それを計算するためにすべての特徴が必要であった。

データが不完全であっても NNC-Tree を効率的に構築するために、２つの手法が考えられる。1 つ目は、欠損値を既存情報から補間、予測などの方法で補欠してから学習する方法である。２つ目は、判断に重要でない欠損値を逆に増やして、性能を保ちながら判断コストを低くする方法である。

実際、既存情報から正しく欠損値が補欠できるのならば、対応する特徴が独立的なものではなく、他の特徴から簡単に推測できるものであると考えられる。この場合に、既存の情報だけを利用しても正しく判断できるはずである。逆に、既存情報から欠損値を補欠することができなければ、対応する特徴が独特なものであると考えられる。この特徴がなければ、正しい判断はできない可能性があるので、信頼できる判断するために、追加データが必要である。従って、欠損値補欠は、現実問題を解くためにあまり有効な手法ではないと考えられる。

一方、欠損値を増やすことによって判断コストを削減する方法においては、最も重要なポイントは「いらない」特徴の検出である。いらない特徴に関しては、欠損値があっても構わない。逆に、いらない特徴ではないのに、欠損値がある場合には、その値を欠損したままでは正しい判断はできない。従って、欠損値をただ増加する方向で考えるといけない。信頼できる判断結果を得るために、必要に応じてデータを追加しなければならない。

従って、不完全データから NNC-Tree を構築するためには、既存の手法をそのままでは利用できない。

２．研究の目的

本研究の目的は、以下の通りである：

１）判断に重要な特徴と「いらない」特徴を学習を通して自動的に選択する。
２）重要な特徴だけを利用して知識学習、パターン分類を行う。
３）以上の結果を利用し、不完全データから NNC-Tree を構築する。

当初の研究計画では、特徴の測定・計測コストなどをも加味し、最も経済的なパターン認識システムを構築することを目標の一つとしたが、現実データの入手が困難であるため、本研究は以上の 3 点に絞った。

３．研究の方法

数多くの特徴の中から、判断に重要な特徴を選択する問題は「特徴選択」という問題である。この問題はパターン認識の分野で広く研究されている。しかし、殆どの既存方法は、大局の意味で重要な特徴を選択する。即ち、すべてのパターンに対して、正しい判断を下すために必要な特徴は重要なものであると考える。しかし、あるコンセプトを表すために、汎用な特徴よりも、独特な特徴を利用したほうがより正確になると考えられる。例えば、インフルエンザの特徴は、心臓病の特徴と必ずしも同じものではない。むしろ、違う病気を診断するために、違う特徴を利用する方がよいと考えられる。

以上の考え方をベースに、本研究では「相対優勢」の概念を利用した新しい特徴選択法を提案した。相対優勢とは、ある特徴は、あるコンセプトを表すと、どちらかといえば、他のコンセプトを表すよりも適していると判断された場合、その特徴をそのコンセプトの優勢特徴であると考える。特徴ベクトルが疎である場合に、優勢特徴はあるコンセプトの「専用特徴」にすることさえできる。特徴ベクトルが疎ではない場合には、各コンセプトに対して、優勢特徴に対応する「インデックス集合」を作ることができる。このインデックス集合に指定されている特徴だけを利用すれば、パターンを効率的に分類することができる。また、これを学習過程に組み込めば、学習の効率化もできる。

以上の相対優勢法は、テキスト分類などの応用には直接適応することができる。なぜかというと、テキスト分類において、特徴量は対応する用語の出現頻度か確率なので、その大きさは用語の重要度を評価するために利用できるからである。しかし、他の応用については、特徴が重要であるかどうかは、その値から直接に判断できない場合は多い。例えば、体温という特徴は、値が高いときにインフルエンザを診断するための重要特徴となるが、白血球の数は、その値が低いときに免疫力低下の指標として知られている。このような場合、体温、白血球数を直接に特徴として使うよりも、「体温が高い」と「白血球数が低い」を重要特徴として利用したほうがよいと考えられる。

さまざまな特徴の重要度を計るために、われわれが特徴のファジィ化による特徴変換を提案した。この変換によって、数値で表されている特徴量をいくつかの「言語的」特徴に展開することができる。例えば、体温という特徴は、「低い」、「正常」、「高い」などの新しい特徴に展開することができる。新しい特徴の値は、それぞれの言語的特徴のメンバーシップ関数に、元の特徴の値を代入して得られる値であり、高ければ高いほどその言語的特徴が「ある」と考えられる。これらの新しい特徴を利用すれば、相対的優勢法で特定のコンセプトを表す優勢特徴を自動的に求

めることができる。実際、これで作られた新しい特徴ベクトルは、テキスト解析に使用されている BOW (Bag-of-words)モデルとよく似ていて、各言語的特徴は word そのものとなる。従って、提案手法を利用すれば、医療診断を含め、さまざまなパターン認識問題が統一した手法で解決することができる。

以上の議論をまとめると以下のようになる。まず、学習データを与え、それをファジィ化して、言語的特徴を求める。そこから相対優勢に基づく学習を行い、優勢特徴を選択する。特定のコンセプトは、その優勢特徴だけで表現されるので、すべての特徴を使用するよりも理解しやすくなる。特定のデータをあるコンセプトと分類される場合、そのコンセプトに対応する優勢特徴に欠損値がなければ、そのまま判断を認める。逆に、優勢特徴に欠損値がある場合、無理な判断を行わず、追加情報をユーザに求めることがより合理的であると考える。

特徴空間の次元は高いとき、ファジィ化を施すと、次元はさらに数倍増えるので、学習効率が悪くなる可能性がある。この問題を解決するために、われわれは次元圧縮についても検討した。要は、次元を大幅に圧縮すれば、その後ファジィ化によって特徴数を増やしても、学習コストを抑えることができると考えられる。次元が高く、データ数が多い場合に、効率的に次元圧縮するために、われわれはさまざまな方法について検討し、最終的に DMC(Discriminant Multiple Centroid)法を提案した。

４．研究成果

(1) 重みつき相対優勢法

前節では、本研究の基本的考え方を簡単に述べたが、本節では本研究の成果をより詳しく説明する。まず、本研究で提案した重みつき相対優勢(WCA: Weighted Comparative Advantage)法を説明する。

WCA はもともと教師信号がない場合のデータ（特にテキストデータ）クラスタリングのために提案した。クラスタリングのためによく知られている方法には k-means 法がある。この方法は、クラスタごとに代表点を求める方法である。仮に k 個のクラスタがある場合、k-means 法による学習は以下の操作の繰り返しとなる：

Step 1: 各データを、それに最も近い代表点に分類する。
Step 2: 各クラスタのデータの平均を求め、それを新しい代表点とする。

最初の代表点は、通常データからランダムに選ぶ。また、Step 2 で求められた新しい代表点が古いものと同じものであれば、繰り返しを中断し、学習を終了する。

k-means 法は簡単ではあるが、さまざまな問題に対して非常に良い結果が得られることが知られている。主な問題点は２つある。１つ目は、各クラスタはその代表点で表され、クラスタの特徴を陽に表現することはできないことである。即ち、仮に正しい分類されたとしても、各クラスタにあるデータを簡単明瞭に解釈することができない。２つ目は、計算コストである。特に適切な k がわからないとき、それを決めるために、k を小さい値から増やしながら k-means を繰り返して実行する場合には、計算コストは非常に高くなる。これらの問題を解決するために、われわれは WCA を提案した。

WCA は、基本的に k-means と同じように上記の Step 1 と Step 2 を繰り返して学習する。ただ、代表点は、すべての特徴量ではなく、相対的に優勢となる特徴だけで構成される。例えば、クラスタ２の代表点の第３番目の特徴量は、他のクラスタの代表点に比べて、最も大きければ、第３番目の特徴は、相対的に、クラスタ２の優勢特徴となる。公平に比較するために、各代表点は、ノルムが１となるように正規化される。また、特徴の値自体は、その特徴の重要性を示す指標であり、その相対的優勢となる特徴の重み(weight)として使われる。この考えをもとに Step 1 を実行すると、すべての特徴を一律に使うよりも k 倍速くなる。

任意のデータ x に対して、x と j 番目のクラスタセンター c_j との類似度は、以下のように求める：

$$S(x, c_j) = \sum_{i \in I_j} x_i c_{ij} \qquad (1)$$

ただし、I_j は c_j の優勢特徴のインデックス集合である。WCA においては、優勢特徴が一つのクラスタに独占されるので、x と k 個のクラスタセンターとの類似度を全部求めるために、m 回の乗算があれば十分である。ここで m はデータの次元数である。従って、WCA の学習スピードは k-means の約 k 倍となる。

WCA アルゴリズムの有効性を確認するために、公開されている文書データベースで実験をしてみた。使用したデータベースは、CLASSIC3 と NSF3 である. CLASSIC3 は SMART system の一部であり、3893 の文書データがある。その中に、MEDLINE, CISI, CRANFILED、３つの部分集合があり、それぞれは医療関連雑誌から取得した 1033 個の要約文、情報検索関連論文から取得した 1460 件の要約文、CRANFIELD は航空関連論文から取得した 1400 件の要約文、を含む。NSF3 は米国自然科学基金で奨励された 4303 件の研究要約を含む。

その中に、846 件は天文学、1954 件は生物学、1503 件はコンピュータ関係である。

表１データベースの説明

	データ数	次元
CLASSIC3	3893	19929
-Medline	1033	
-Cisi	1460	
-Cranfiled	1400	
NSF3	4303	18721
-Astronomy	846	
-Biology	1954	
-Computer	1503	

表２：実験結果(k=3)

	Precision	MER	CPU Time
WCA-CLASSIC3	0.949	0.974	3.665
k-means-CLASSIC3	0.934	0.969	8.893
WCA-NSF3	0.924	0.934	5.005
k-means-NSF3	0.904	0.919	16.716

実験では、k-meansとWCAをそれぞれ500回試行して、その平均結果を表２にまとめた。その中に、精度(Precision)、Mutual Exclusion Rate (MER)、計算時間(CPU Time)が表示されている。この結果からわかるように、WCAはk-meansよりも高速に学習することができ、しかもよい結果を得られている。実際、表２からわからないが、WCAで得られた代表点は、対応するクラスタの最も重要な特徴だけで構成されているので、k-meansで得られたものよりもわかりやすくなっている。

(2) 教師ありWCA

以上で説明したWCAを教師あり学習に適応するために、われわれはSWCA(Supervised WCA)を提案した。教師信号がある場合、各代表点の各特徴の重要度を決めるために、特徴の値だけではなく、その特徴のクラス間分散も使用できる。例えば、ある特徴は、クラス２の代表点でしか大きくならない場合（クラス間分散は小さい)、その特徴はクラス２の優勢特徴であると考えられる。逆に、ある特徴は、すべてのクラスの代表点においてほぼ同じ値を取る場合（クラス間分散は大きい)、この特徴はあまり重要でないと考えられる。従って、特徴の重要度はクラス間分散に反比例する。重要ではない特徴量を無視することによって、データを分類する際に必要とされる計算コストを減らすことができる。

また、WCAにおいては、相対的に優勢な特徴は一つのクラスタに独占されたが、これは一般的に有効ではない。特徴ベクトルが疎ではない場合や複数のクラスタに重要な特徴がある場合、WCAでは必ずしもよい結果が得られない。この問題を解決するために、優勢特徴のインデックス集合を拡張すればよい。即ち、インデックス集合I_jには、j番目のクラスにとって重要な特徴であればそのインデックスを含むが、同じ特徴は他のクラスにも利用される可能性がある。従って、$I_1, I_2, \cdots, I_k$ の積集合は必ずしも空ではない。このようにすることによって、高速性を維持しながら分類性能を向上することができる。

表３　SWCAの実験結果（CLASSIC3）

Method	Accuracy	Training Time	Test Time
SWCA	0.9949	0.2146	0.0095
DSM	0.9949	1.3287	0.0697

表４　SWCAの実験結果（NSF3）

Method	Accuracy	Training Time	Test Time
SWCA	0.9846	0.8372	0.0261
DSM	0.9845	2.6540	0.0892

SWCAの性能を確認するために、前の実験と同じデータを利用して実験を行った。ただし、この実験では、教師信号を使用するので、比較対象はWCAとk-meansではなく、LVQの一種であるDSM(decision surface map)アルゴリズムにした。表３と表４は結果を示している。これらの結果から、DSMに比べて、SWCAは分類性が劣らずに効率的に学習することができたことがわかる。

(3) データのファジィ化

WCAとSWCAを利用する際に、特徴の値はその特徴の重要性を直接的に反映できると仮定している。テキスト分類などの分野においてこの仮定は成立するが、一般的には成立しない。この問題を解決するために、われわれが特徴のファジィ化を提案した。

特徴をファジィ化するために、各特徴に対して、k-means法でまずクラスタリングを行う。これによって、対応する特徴の出現頻度の高い値をいくつか求めることができる。これらの値を中心に、「言語的値」を定義することができる。例えば、体温という特徴について、正常な人の体温は約36.5度前後、インフルエンザ患者の体温が39.0度前後、などが知られている。たくさんの人の体温データをクラスタリングすれば、これらの代表値はクラスタセンターになると思われる。これらの代表値

をクラスタリングで求め、それらを中心に、三角関数やガウス関数などを使ってメンバーシップ関数を定義すれば、与えられたデータをファジィ化することができる。

一つの特徴をファジィ化すると、k個の特徴に展開する。このk個の特徴は、メンバーシップ関数の値をそのまま使ってもよければ、k個の特徴の中で、最大となるものだけを１にし、他は0にすることもできる。いずれにしても、特徴の値は、その特徴の重要性を示すものとして考えられる。また、ある特徴に欠損値がある場合、ファジィ化で得られるk個の特徴はすべて０となる。通常、このようなデータには、正しい教師信号があれば、欠損値がそもそもいらないと考えられる。

以上のようにファジィ化を行えば、WCAとSWCAをどんな問題にも適応することができるようになる。言い換えれば、WCAとSWCAは、さまざまな問題を解決するための統一的アルゴリズムとして使用できる。この統一的アルゴリズムを利用すれば、各クラス（或はクラスタ）を代表する優勢特徴を学習しながら見つけることができる。また、優勢特徴だけを利用してパターンを分類しているので、分類も学習も効率的に行える。

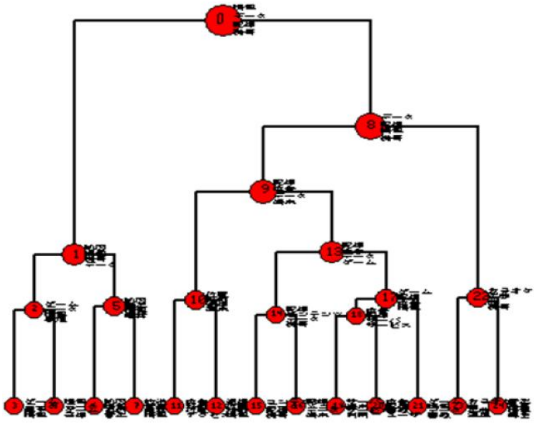

図1：NNC-Treeの構成例

(4) システム全体の構成

WCAやSWCAを再帰的に利用すれば、図1のNNC-Treeを構築することができる。構築過程自体はこれまでの方法と同じである。この例は２分木であるが、k=3にすれば3分木を得ることもできる。多分木の場合、ルートから離れている中間ノードには、必ずしもk個のクラスのデータを全部含まないので、その枝数はk以下となる可能性がある。システムの綺麗さを重視する場合には、k=2にした方がよい。実際、kを２に固定しても任意の問題を解決することができる。システムをより効率的に構築したい場合には、k>2を利用したほうがよい。なぜかというと、WCAやSWCAの計算コストは、k-meansやDSMのそれの約ｋ分の１であり、kを大きくすればするほど、学習のスピード向上が期待できる。

(5) 学習モデルの拡張

これまでの研究では、距離や類似度を直接に利用することができるNNC-Treeを利用してきた。実際、ここで提案した方法は、ニューラルネットツリー(NNTree: Neural Network Tree)にも適応できる。NNTreeは、NNC-Treeと同じ構成をしているが、各中間ノードには、NNCではなく、階層型ニューラルネット(MLP)に置き換えただけである。

MLPの中間ニューロンの役割について、従来、一つのニューロンが一つの超平面に対応し、パターンがその超平面のどっち側にあるかによって分類される。しかし、われわれの脳の中においては、各ニューロンがむしろパターンマッチングの役割をしているのではないかと考えられる。即ち、入力パターンがニューロンの重みベクトルとよくマッチする場合、ニューロンの効果的入力(effective input)が大きくなり、それがある閾値より大きくなるとパターンマッチングが成功したといえる。

従って、NNTreeも類似度に基づく多変量決定木の一種である考えられる。各中間ノードにあるMLPの中間層ニューロンが、学習によって特殊のパターンしか認識しなくなり（いわゆる専門家ニューロンとなる)、それを利用すれば、NNC-Treeと同じように、理解できる知識の学習ができる。これはわれわれの今後の課題として引き続き検討したい。

(6) 次元圧縮

問題空間の次元は非常に高い場合、前述のファジィ化を施すと、次元がk倍に増えるので、学習効率が悪くなる。この問題を解決するために、次元の圧縮が考えられる。即ち、元の次元を大幅に圧縮すれば、その後ファジィ化によってk倍に増えても、学習コストを抑えることができる。

よく知られている次元圧縮法には、主成分分析(PCA: Principal component analysis)と線形識別解析（LDA: Linear Discriminant Analysis）がある。教師信号がわかっているとき、LDAの方がより効果的であると知られている。LDAでは、クラス数はC個ある場合、もともとの次元が非常に大きい場合でも、C-1次元に圧縮することができる。問題は、次元が高く、データが多い場合には、LDAの変換行列を求めるための計算コスト非常に高いことである。これを解決するために、われわれはDMC(Discriminant multiple centroid)アルゴリズムを提案した。

DMCはTwo-Stage圧縮法である。第１ステー

ジでは、まず高次元のデータをk個のクラスタセンターに写像する。ここでkはクラス数Cより大きいが、データ数や次元数に比べて遥かに小さい整数である。第２ステージでは、写像されたk次元空間においてLDAを施す。すると、最終次元はLDA圧縮と同じように、C-1となるが、計算コストは非常に低くなる。

このように、次元が高いときに、まずDMCで次元圧縮を行い、C-1次元にする。この低次元空間でファジィ化をすれば、WCAやSWCAをそのまま利用できるようになる。ただ、これで得られた優勢特徴はどんなものなのか解釈しにくい。また、このように得られたNNC-TreeやNNTreeの性能はどうなるか、などについては、今後の課題となる。

(7) まとめ

以上をまとめると、本研究は以下の成果を得ることができた：

- WCAとSWCAによる学習の効率化と重要（優勢）特徴の自動選択。
- データのファジィ化による学習アルゴリズムの統一化。
- DMCによる次元圧縮の効率化。

今後の課題としては、より多くデータを利用して提案手法を実証することと、提案手法を実用化することである。近い将来、本研究の成果をさまざまな実問題に応用し、役に立つものにしたい。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）

① T. Chan, J. Ji and Q. F. Zhao, "**Learning to detect spam: Naïve-Euclidean approach**," International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 1, No. 1, pp. 31-38, 2009.
② H. Hayashi and Q. F. Zhao, "**A fast algorithm for inducing neural network trees,**" Journal of Information Processing, Vol. 49, No. 8, pp. 2878-2889, 2008. (in Japanese)
③ Jie Ji, T. Chan, and Q. F. Zhao, "**Clustering large sparse text data: a comparative advantage approach**" conditional accepted for publication in Journal of Information Processing Society of Japan.
④ Jie Ji and Q. F. Zhao, "**Applying naïve Bayes classifier to document clustering**," accepted for publication in Journal of Advanced Computational Intelligence and Intelligent Informatics.

〔学会発表〕（計 6 件）

① Jie Ji, Daichi Kunita and Qiangfu Zhao, "A Customer Intention Aware System for Document Analysis," to appear in Proceeding of IJCNN2010, Spain.
② H. Hayashi and Q. F. Zhao, "**Model reduction of neural network trees based on dimensionality reduction,**" Proc. of International Joint Conference on Neural Networks (IJCNN2009), pp. 1171-1176, Atlanta, 2009.
③ J. Ji, T. Chan and Q. F. Zhao, "**Fast document clustering based on weighted comparative advantage,**" Proc. of IEEE International Conference on Systems, Man and Cybernetics (SMC2009), pp. 547-552, Texas, 2009.
④ H. Hayashi and Q. F. Zhao, "**Induction of compact neural network trees through centroid based dimensionality reduction,**" Proc. of IEEE International Conference on Systems, Man and Cybernetics (SMC2009), pp. 974-979, Texas, 2009.
⑤ J. Ji, T. Chan and Q. F. Zhao, "**Comparative advantage approach to classifying sparse text data,**" Proc. of IEEE International Conference on Computer and Information Technology (CIT2009), pp. 3-8, Xiamen, 2009.
⑥ J. Ji, R. Shindo, Q. F. Zhao and Y. Kunishi, "**A study on criteria for extracting key terms in document clustering,**" Proc. IEEE International Conference on Systems, Man and Cybernetics (SMC2008), pp. 3674-3679, Singapore, 2008.

〔その他〕

① Y. Watanabe, "**Extracting understandable knowledge based on data fuzzification**," Master Thesis of The University of Aizu, Mar. 2009 (supervised by Q. F. Zhao).

６．研究組織

(1)研究代表者

趙　強福（Qiangfu ZHAO）
会津大学・コンピュータ理工学部・教授
研究者番号：90260421

(2)研究分担者

ユー　リュー（Yong Liu）
会津大学・コンピュータ理工学部・上級準教授
研究者番号：60325967