様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　6 月 18 日現在

研究種目： 若手研究（スタートアップ）
研究期間： 2008 ～ 2009
課題番号： 20800045
研究課題名（和文）　組合せ構造に基づいた新しい学習手法の開発

研究課題名（英文）　A Novel Learning Method Based on Combinatorial Feature of Data

研究代表者
原口　和也（HARAGUCHI KAZUYA）
石巻専修大学・理工学部・助教
研究者番号： 80453356

研究成果の概要（和文）：本研究課題では、データの組合せ的な特徴に基づいた新しい学習手法の開発を目標としてきた。このためデータの特徴を良く表すような組合せ構造を抽出し、それを解析することで学習を行うという手順を考えた。分類と呼ばれる学習問題に対し、2 部グラフ構造に基づいた学習アルゴリズムを提案した。また、その学習能力が既存のものと比較しても遜色無く、ある特殊な場合においては優れることを実験的に示した。

研究成果の概要（英文）：In this project, we have aimed at establishing a novel learning method based on combinatorial feature of data. For classification, an essential learning problem, we proposed a learning algorithm based on bipartite graph structure. In computational experiments, we observed that its learning ability is competitive with previous methods and is even superior in some special cases.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,200,000	360,000	1,560,000
2009 年度	1,140,000	342,000	1,482,000
総　計	2,340,000	702,000	3,042,000

研究分野：数理工学
科研費の分科・細目：情報学基礎
キーワード：アルゴリズム、機械学習、情報可視化

１．研究開始当初の背景

(1) 大量のデータから意味のある知識を発見するための手法が、データマイニング、バイオインフォマティクス等の発見科学の領域において近年ますます求められている。とりわけ、分類（classification）やクラスタリング（clustering）は重要な問題として知られる。分類・クラスタリングに対する従来手法は、データ（ベクトルの集合）は距離空間に存在するという幾何的なコンセプトを暗に含んでいる。

(2) しかしデータを幾何的に捉えるアプローチだけがすべてであろうか。たとえばクラスタリングでは、データベクトル間の距離を定める関数の定義によって計算結果が大きく変わるという問題点がある。また、特に数値で表されるデータの場合、スケーリングの問題が必ずついて回る。そもそも 1 つの距離はデータ全体から見ると局所的な情報しか持っていない。そのような距離だけを手掛かりに学習を行う方法では、将来完全な学習システムを構成できるかどうか甚だ疑問で

ある。
　もちろんデータを幾何的に捉えるアプローチは自然であるし、実用におけるいくつもの成功例もあることから、これを否定する意図はない。しかし、幾何的ではない、他の視点からデータを捉えた学習手法があってもよいのではないだろうか。

２．研究の目的
　距離はベクトル間の「近さ」という関係を数値化したものである。本研究課題では学習の本質に迫るべく、より抽象化されたベクトル間の関係を考える。すなわち、ベクトル間の距離という数値を考えず、代わりに二項関係（大小関係、同値関係など）を定義する。そうするとノルムの大きさ等に依らず、組合せ構造がデータに対して構築される。そのような組合せ構造に基づいた学習手法を考える。すなわち、(A) データの特徴を表現する組合せ構造を抽出し、(B) その構造を分割することによって学習（分類・クラスタリング）を行うという、組合せ構造に基づいた新しい学習手法のフレームワークを提案したい。

３．研究の方法
(1) 第２節で述べた(B)について、離散数学の分野では既にグラフ分割、半順序分割など、数多くの組合せ構造に関する分割アルゴリズムが研究されており、これを学習に利用する。すなわち(B)で用いる組合せ構造のプロトタイプとその分割アルゴリズムを定めれば、(A)においてデータを組合せ構造に変換することによって学習を行うことができるはずである。

(2) そこで本研究課題では学習問題として分類に焦点を絞り、提案するフレームワークに基づいた学習アルゴリズムを開発する。また上記（B）で用いる組合せ構造として、全順序関係を採用する。ベクトルに全順序を定め（第２節（A）に対応）、得られた鎖を分割する（第２節（B）に対応）ことによって学習を行うが、このための方式を確立するための研究を行い、アルゴリズムを開発する。開発されたアルゴリズムの実装を行い、ベンチマークデータなどを用いて分類器の汎化誤差を実験的に評価する。

４．研究成果
(1) 第２節（A）に関して、データベクトルに全順序を与えるための数理モデルとして２部グラフの枝交差数最小化問題を考える。まず、データの特徴を示す次のような２部グラフを考える：前もって与えられた決定表のエントリ（決定表の１行に対応）を一方の節点集合、各データベクトルをもう一方の節点集合とし、各データベクトルとマッチするすべてのエントリに対して枝を張る。

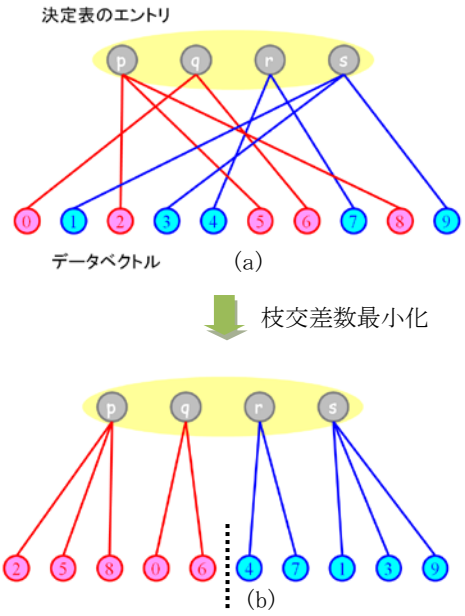

図１：枝交差数最小化に基づく全順序決定のアイディア

　このような2部グラフを図1(a)のように配置する。この図ではデータベクトルの色（および接続する枝の色）はそれが属するクラスを示すものとする。青に多く接続するエントリをより右に、そうでないエントリをより左に置く。このようなエントリの順序に対し、枝交差数が最小となるようにデータベクトルを整列すれば、図 1(b)のように青と赤のベクトルが分離される。このように得られたデータベクトルの全順序は、クラス分離に有用な情報を持つことが期待される。
　我々は、全順序上に適当なしきい値を定め、左右どちらに落ちるかにしたがって新しいデータベクトルを分類することにより、この２部グラフを２値分類器として用いることを提案した。また計算実験の結果、提案手法は決定木やＳＶＭなどの分類器と同等の汎化誤差を達成することを確認した。（これらの結果は、第５節の論文③に主たる内容として含まれる。）
　また研究過程で得られた知見から、上記２部グラフの枝交差数最小化がデータの線形分離性と密接な関係があることを予想した。この予想に基づき、提案した２値分類器を多値分類器に拡張した。ECOC 法（Error Correcting Output Codes）などの従来の汎用拡張法と比較して、元の多値問題をより少ないクラスの問題に分解する点が、提案手法の特徴の一つである。計算実験の結果、特にクラス数が多い場合において、（元来多値問題への拡張が自然に定まる）決定木、および

ＳＶＭのある種の拡張を上回る汎化誤差を達成することを確認した。(これらの結果は、第 5 節の論文⑤に主たる内容として含まれる。)

(2) 研究成果の意義を述べる。
① 従来の学習手法が幾何的なコンセプトに基づいているのに対し、本研究は組合せ構造に基づいた全く新しい学習手法のフレームワークを創出しようとするものである。本研究課題では提案フレームワークに基づいた学習アルゴリズムの一例を開発し、その可能性を確認することができた。

その手始めとして全順序関係に着目したのは、従来の学習手法が超平面を構成する過程で、暗にベクトルに対して全順序関係を定めていることによる。全順序はデータ解析において何かしら本質的な役割を果たしているのかもしれない。我々はその全順序を陽に定めることで、学習を行った。

② なお、枝交差数最小化問題は一般に NP 困難であり、計算実験では適当な近似アルゴリズムを選択する必要があった。我々が選んだのは、その近似性能の高さが既に報告されている重心法である。重心法を用いる場合、提案した 2 値分類器は幾何的な超平面分類器として解釈できるというメリットがある。今後の課題の一つとして、枝交差数最小化が、学習の改善につながることを示すことが挙げられる。

③ 枝交差数最小化は、情報可視化の分野で広く用いられる要素技術の一つである。当該分野の従来研究では、枝交差数最小化自体はデータの解析に用いられなかった。すなわち、分類・クラスタリング・主成分分析などの解析ツールによって前処理されたグラフデータを、描画するためのものに過ぎなかった。本研究の提案は、「枝交差数最小化を通して学習するアルゴリズム」であり、可視化技術の新しい可能性を見出すものとして解釈できる。

(3) 次に今後の展望を述べる。
① 既に述べたように、枝交差数最小化が学習の改善につながることを示すことは、主たる課題の一つである。2 部グラフの枝交差数はケンドール距離と密接な関係にあり、このことをベースにした議論が期待される。これまでにＰＡＣ学習などの枠組みで多くの学習基準が提案されており、それらとの関係を見出すことを目標としたい。

② 本研究課題では組合せ構造に基づいた学習と題し、データベクトルに対する「もっともらしい」全順序の決定について研究を進めてきたが、「もっともらしい」半順序の決定についても検討したい。(なお、上述の 2 部グラフから定まるデータベクトル間の半順序関係については、第 5 節の論文④で議論している。)

このアイディアの背景は、以下のようにまとめられる。

機械学習のこれまでの研究は、数値ベクトルの集合として与えられるデータ（数値データ）を対象とするものが中心であった。しかし世の中には、味、色、におい、文字など大きさを持たず、一般には数値化して取扱うのが難しい、記号ベクトルの集合として与えられるデータ(記号データ)も数多く存在する。ゲノム配列などはその好例であろう。例えば記号データから分類器を学習する場合、これまでに用いられてきたアプローチは以下の 2 つに大別される。

(i) 決定木や決定表など、記号データを直接取扱うことのできる分類器モデルを用いる。

(ii) 記号データを数値データに変換し、数値データに対する分類器モデルを用いる。数値データに陽に変換せずとも、カーネル関数を用いて暗に変換し、分離平面を構成する手法（SVM など）はこの類であろう。

数値ベクトル間に適当な半順序関係を定めることで（例: componentwise order)、任意の数値データを半順序集合と見なすことができる。よってアプローチ（ii）では、記号データは距離空間上の数値データに変換されるだけでなく、暗に半順序集合に埋め込まれるものと解釈できる。

変換によって得られる数値データは、元の記号データの当該距離空間における分布・配置などの幾何的な特徴を持つと考えられる。一方、そのように暗に得られる半順序集合は、記号データのトポロジカルな、あるいは構造的な特徴を持つのではないか。このことから、距離空間を経由することなく、記号データの特徴を表す半順序集合を何らかの方法で直接抽出し、学習に利用することはできないだろうかという着想を得た。

そもそも、記号データを数値データに無理矢理変換するアプローチ（ii）が認められるのであれば、「もっともらしい」半順序集合にデータを埋め込み、学習に利用する手法も許容されてよいはずである。そしてそのような手法は、ベクトル間の距離や順序の概念を(一般には)考慮しないアプローチ（i）と、距離を考慮するため、結果的に順序の概念も暗に含むアプローチ（ii)の中間に位置する手法として、直感的に位置づけられる。

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５件）
① Kazuya Haraguchi, Seok-Hee Hong and Hiroshi Nagamochi, “Visual Analysis of Hierarchical Data Using 2.5D Drawing with Minimum Occlusion,” Department of Applied Mathematics and Physics, Kyoto University, Technical Reports, 査読無, 2009-010.
② Kazuya Haraguchi, Seok-Hee Hong and Hiroshi Nagamochi, “Classification via Visualization of Sample-feature Bipartite Graphs,” Department of Applied Mathematics and Physics, Kyoto University, Technical Reports, 査読無, 2009-011.
③ Kazuya Haraguchi, Seok-Hee Hong and Hiroshi Nagamochi, “Bipartite graph representation of multiple decision table classifiers,” Proc. SAGA 2009 (LNCS 5792), 査読有, 2009, 46-60.
④ Kazuya Haraguchi, Seok-Hee Hong and Hiroshi Nagamochi, “Visualization can improve multiple decision table classifiers,” Proc. MDAI 2009 (ISBN: 978-84-00-08851-4), 査読有, 2009, 41-52.
⑤ Kazuya Haraguchi, Seok-Hee Hong and Hiroshi Nagamochi, “Multiclass Visual Classifier Based on Bipartite Graph Representation of Decision Tables,” Proc. LION 4 (LNCS 6073), 査読有, 2010, 169-183.

〔学会発表〕（計３件）
① Kazuya Haraguchi, Seok-Hee Hong and Hiroshi Nagamochi, “Classification by Ordering Data Samples,”Kyoto RIMS Workshop on Acceleration and Visualization of Computation for Enumeration Problems (AVCEP08), 2008.
② Kazuya Haraguchi, Seok-Hee Hong and Hiroshi Nagamochi, “Visualized Multiple Decision Table Classifiers without Discretization,” 4th Korea-Japan Workshop on Operations Research in Service Science, 2009.
③ Kazuya Haraguchi, Seok-Hee Hong and Hiroshi Nagamochi, “Learning Classifier by Edge Crossing Minimization,” Int’l workshop on Multi-dimensional Visualization, 2010.

６．研究組織
(1)研究代表者
原口　和也（HARAGUCHI KAZUYA）
石巻専修大学・理工学部・助教
研究者番号：80453356

(2)研究分担者
なし

(3)連携研究者
なし

(4)研究協力者
永持　仁（NAGAMOCHI HIROSHI）
京都大学大学院・情報学研究科・教授
研究者番号：70202231