様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　６月１０日現在

機関番号：31308
研究種目：若手研究（B）
研究期間：2010～2012
課題番号：22700015
研究課題名（和文）順序関係に着目した記号データの新しい学習法

研究課題名（英文）A Novel Learning Method That Highlights the Order Relation between Data Examples

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

研究成果の概要（和文）：本研究では機械学習における分類問題において、新しい分類器モデルの構築を目指してきた。着目したのは、分類器がデータ空間を分割し、部分空間のランキングを生成する点である。真のランキングを有する人為的なデータにおいて、真のランキングと分類器ランキングのケンドール距離が、汎化誤差と高い相関を持つことを確認した。また解析の結果、当該距離はAUCと分類器の複雑さに相当する値の和で表されることがわかった。

研究成果の概要（英文）： We have worked on developing a novel classifier model for the classification problem, which is one of the most significant issues in machine learning. We have remarked that classifier partitions the data domain into subspaces and induces a ranking on them. Conducting computational experiments with an artificial data set such that the true ranking is given, we observed a high correlation between the generalization error and the Kendall-tau distance from the ranking induced by the classifier to the true ranking. As a result of analysis, we showed that this distance is represented as the sum of the normalized AUC and the value that reflects the complexity of the classifier.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,100,000	330,000	1,430,000
２０１１年度	1,000,000	300,000	1,300,000
２０１２年度	900,000	270,000	1,170,000
総　計	3,000,000	900,000	3,900,000

研究分野：総合領域
科研費の分科・細目：情報学、情報学・情報学基礎
キーワード：計算論的学習理論

１．研究開始当初の背景

本研究では、機械学習において最も重要な問題の一つである２値分類問題を考える。この問題では、事例の集合が訓練集合として与えられる。各事例はベクトルで表され、正もしくは負、いずれかのクラスに属する。この問題の目的は、まだ見ぬ事例のクラスを精度良く予測するような分類器（事例空間から正負クラスへの関数）を、訓練集合から学習することである。事例が記号ベクトルで表されるようなデータを記号データ、事例が数値ベクトルで表されるようなデータを数値デー

タと呼ぶことにする。
従来、記号データから学習を行うためには、以下のいずれかのアプローチが用いられてきた。
（ア）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。決定木や決定表はそのようなモデルの例である。
（イ）数値データに変換して数値用の分類器モデルを用いる。特徴空間に超平面を引くことで分類器を構成する SVM などがこのようなモデルの例として挙げられる。

２．研究の目的

上記従来手法に関して、一般に（ア）はデータベクトル間の距離や順序関係に関する概念を用いず、（イ）は両方の概念を併せ持つ。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（ア）が認められるのであれば、「もっともらしい」順序付き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（ア）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（イ）の中間に位置する手法として、直感的に位置づけられるであろう。
本研究課題では上記（ア）（イ）のいずれとも異なる新しい学習手法の確立を目指し、順序関係のみに着目した記号データ学習に挑んだ。当該研究期間における目標として以下の３つを掲げた。
（１）順序関係に着目した記号データからの学習アルゴリズムの開発。
（２）提案手法と従来手法の比較。具体的には、本質的に何が異なり、どの点でどちらが優れているか等を議論する。
（３）学習結果をグラフによって可視化し、ユーザの知識発見をサポートするためのソフトウェアの実装。

３．研究の方法

本研究課題は、原則的に研究代表者一人によって遂行するものである。すなわち、仮説生成、実験および解析、およびその考察を一人で行った。
また研究の円滑な遂行を図るため、外部からの意見を取り入れるべく、連携研究者および関連分野の専門家との研究打合せを行った。また当該分野の最新の研究動向に関する情報収集のため、関連会議や研究会への出張を行った。

４．研究成果

本研究では、任意の分類器モデルがデータ空間を分割し、部分空間のランキング（順序付け）を暗に生成することに着目してきた。例えば、超平面モデルは分離平面からの距離によってランキングを生成し、決定木モデルは一つ一つの葉、決定表モデルは一つ一つの行に点数を与えることによって、線形順序に基づいたランキングを生成する。分類器をランキング空間における点とみなし、ランキング空間の探索を通じて分類器を構成する試みを行った。

（１）はじめに、何をもって良いランキングとするかの基準を定めなければならない。このため真のランキングが与えられた理想的なデータを考える。このようなデータでは、分類器が生成するランキングと真のランキングのケンドール距離が、汎化誤差と高い相関を持つのではないかと予想した。そしてその予想の妥当性を実験によって示した。
理想的なデータというのは一般に存在しないので、実験的に生成しなければならない。事例の定義域を n 次元 0-1 ベクトルの集合とし、適当に定めた論理関数によって事例のクラスが決定されるものとする。また真のランキングとして 2 値ランキング（任意の正事例は任意の負事例より上位にランク付けされ、同じクラスの事例は同位にランク付けされるようなもの）を考えた。
実験結果の一部を図１に示す。提案指標と汎化誤差の相関係数は、訓練誤差と汎化誤差の相関係数よりも高い。訓練誤差は汎化誤差の近似値として良くないことが知られているが、少なくともそれよりは見込みがあることが示されたという点で、有意義な結果だと考えられる。

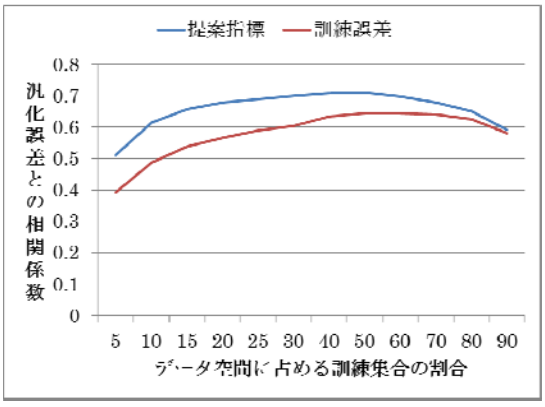

図１：汎化誤差との相関係数の比較

（２）一般のデータには真のランキングなるものが存在しない。そのためヒューリスティクスに基づいてランキング空間を探索する必要がある。ランキング間の距離に基づいて

決定木を構成するアルゴリズムのプロトタイプを開発し、その有望性を実験的に示した。

任意の正事例が任意の負事例より上位にランク付けされたランキングを分離ランキングという。またランキングにおける同値類の個数をそのランキングの次数という。よって（1）で述べた2値ランキングは、次数が2であるような分離ランキングである。

開発したアルゴリズムは、すべての事例が同位にランク付けされたようなランキングから始め、最も近い分離ランキングに近づくように葉の分枝を繰り返し、分枝できなくなった後は、2 値ランキングに最も近づくように枝刈りを続けるものである。

表1に評価実験の結果を示す。C4.5 は当該分野で広く用いられている決定木生成アルゴリズムである。C4.5 にはいくつかのパラメータがあるため、6 通りのパラメータ値で決定木を生成し、そのうち最小の汎化誤差を比較の対象とした。C4.5 に有利な条件を与えているにも関わらず、開発したアルゴリズムは勝るとも劣らない分類器を生成していることがわかる。

表１：提案手法と C4.5 の汎化誤差の比較
（単位はパーセント，データは UCI Repository of Machine Learning より）

	提案手法		C4.5
	枝刈り前	枝刈り後	
ADULT	20.26	24.57	**13.78**
BCW	4.85	6.83	**4.75**
CHESS	**0.46**	4.72	0.57
CREDIT	18.46	**13.76**	13.88
GLASS	8.41	7.58	**6.75**
HABER	34.98	**26.85**	28.05
HEART	24.74	24.32	**22.53**
HEPA	18.04	**15.86**	16.14
IONO	16.56	**5.3**	8.61
MONKS-1	**3.7**	16.67	21.99
MONKS-2	**21.3**	32.87	32.87
MONKS-3	10.19	**2.78**	**2.78**
MUSH	**0**	1.48	**0**
PIMA	29.81	26.77	**25.63**
SPAM	8.91	10.46	**7.16**
TTT	**5.44**	17.59	14.31
VOTING	6	4.39	**3.65**
WDBC	6.61	7.37	**6.26**

（3）オッカムの剃刀の観点から、2 値ランキングを真のランキングと見なすことは決して悪い仮定ではないと考えられる。そこで解析を行った結果、分類器が生成するランキングと2値ランキングの間のケンドール距離は、分類器の学習性能の評価指標の一つである AUC と分類器の複雑さに相当する値の和で表されることがわかった。

分類器が生成した訓練集合 S のランキングを分割 $S=S_1 \cup \ldots \cup S_k$ で表す。ここに k はランキングの次数で，$i<j$ ならば S_j に属する事例は S_i に属する事例より上位にランク付けされているものとする。また $S=S^+ \cup S^-$ で，S^+ と S^- はそれぞれ正例と負例の集合とする。

解析の結果、分類器が生成したランキングと2値ランキングのケンドール距離は、

$$-2|S^+||S^-|AUC - \frac{1}{2}\sum_{i=1}^{k}|S_i|^2 + \text{定数}$$

となる。AIC や MDL など、導出量に類する指標が汎化誤差と高い相関を持つことは既に学習理論の分野で広く知られており、我々の視点の妥当性が示唆されている。

（4）提案手法の一部をデータマイニングソフトウェア Weka の上で利用できるように実装作業を行った。

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１件）

① Kazuya Haraguchi, Constructing A Classifier by Searching The Ranking Space. 石巻専修大学 研究紀要 第24号, 査読無, 2013, pp. 7-13, http://www.isenshu-u.ac.jp/library/kenkyu/24/start.html

〔学会発表〕（計０件）
無し

〔図書〕（計０件）
無し

〔産業財産権〕
○出願状況（計０件）
無し

○取得状況（計０件）
無し

〔その他〕
無し

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・准教授
研究者番号：80453356

(2)研究分担者
無し

(3)連携研究者
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