4 「超長方形分割」による関係データ解析

関係データ解析を行う機械学習の全く新しいパラダイム「スーパーベイズ法」を提唱

NTT コミュニケーション科学基礎研究所（以下、CS 研）はさまざまな機械学習技術の研究に長年取り組んで来た。本稿では CS 研が考案した、関係データ解析にパラダイムシフトを起こすような新たな機械学習の手法について紹介する。

関係データ解析とは

EC サイトなどでどのユーザーが何を購入したか、のような関係を行列で表したものを「関係データ」と呼んでいる。複数の頂点とそれらを繋ぐ辺で表現できるネットワーク型データも関係データとして表現することが可能だ。SNS でどのユーザー同士がつながっているか、といった関係を示すのに適している。

こうした関係データを解析し、「特定のユーザー群が特定の商品群を購入しやすい」といった傾向の把握が可能であることが知られている。このことを利用し、EC サイトならおすすめの商品を提示する、SNS なら興味のありそうなユーザーの投稿を紹介する、といったことが行われている。

「長方形分割」を用いた関係データ解析

図 1 左に、行方向がユーザー群、列方向が商品群となっている関係データの例を示す。商品の購入有無を黒と白で示している。このままでは購買傾向を把握できないため、行と列を適切に並び替え、黒／白がそれぞれなるべく集まるようにし、赤い補助線を引いたものを図 1 右に示す。いくつかの黒い長方形と白い長方形でクラスタリングされていることがわかる。同じ長方形に含まれるユーザーは似たような購買傾向を示すと推測できるため、あるユーザーが特定の商品を購入したことがなくても、その商品を購入しそうであるかを推測できる。

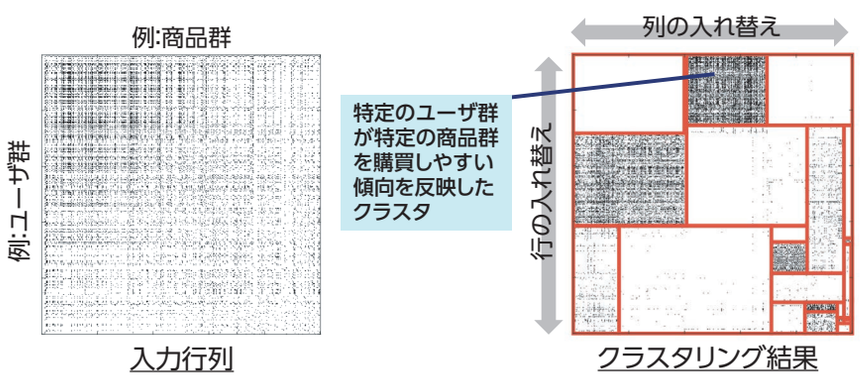

図 1　関係データ解析

NTT コミュニケーション科学基礎研究所メディア認識研究グループ
研究員　**中野 尤裕**氏

ノンパラメトリックベイズ法を利用する機械学習

行と列の並び方は、それぞれデータ数の階乗だけ存在する。データ数が増えると単純作業で行と列を並べ替え、長方形分割を行うのは難しい。そこで、実用的な長方形分割の手法が大きな研究テーマとなってきた。なかでも CS 研が 2000 年代前半から力を入れてきたのが“ノンパラメトリックベイズ法”を利用する機械学習であった。

「ノンパラメトリックベイズ法は

さまざまな研究機関が研究してきた手法です。CS 研はその黎明期から強みを発揮してきました。私は学生の頃にノンパラメトリックベイズ法に魅了され、当時すでにノンパラメトリックベイズ法に強かった CS 研に入所しました。それ以来継続して研究しています。」（中野氏）

この機械学習の手法を簡単に説明すると、まず縦軸・横軸共に 0 から 1 の仮想的な平面上で、図 2 左に示すようなランダムな長方形分割を行う。次にこの仮想平面上に、入力行列の行と列に対応する座標を、やはりランダムに生成する。こうして得られた長方形分割によるデータの説明のしやすさを評価し、不足があるようなら図 2 中央に示すように補助線を少し修正する。そして図 3 に示すように、入力行列の行と列に対する座標を再びランダムに生成し、よりデータを説明しやすくなったか、よりデータに適合する長方形分割になったかを評価する。

このような作業の反復により、最適な長方形分割の状態を探索する。この手法により、3 × 3 の入力行列に対し長方形分割を行うイメージを図 4 に示す。

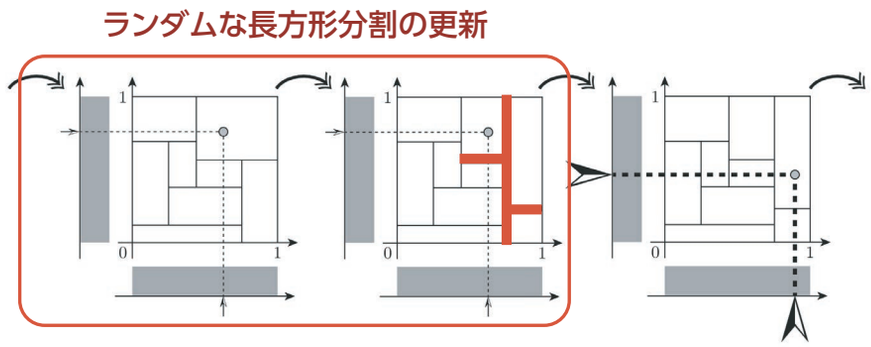

図 2　仮想平面上のランダムな長方形分割を更新

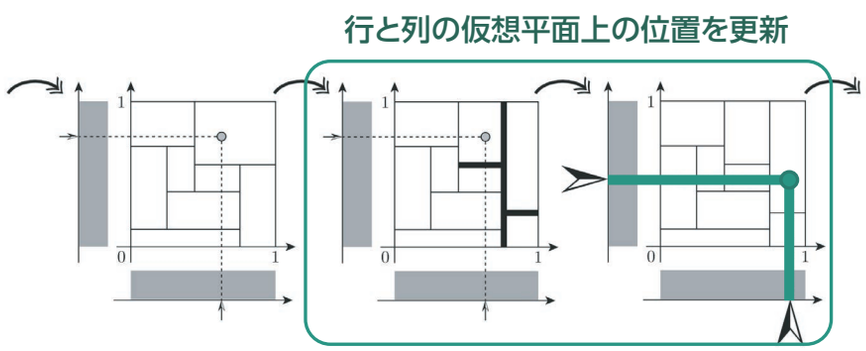

図 3　入力行列の仮想平面上の位置を更新

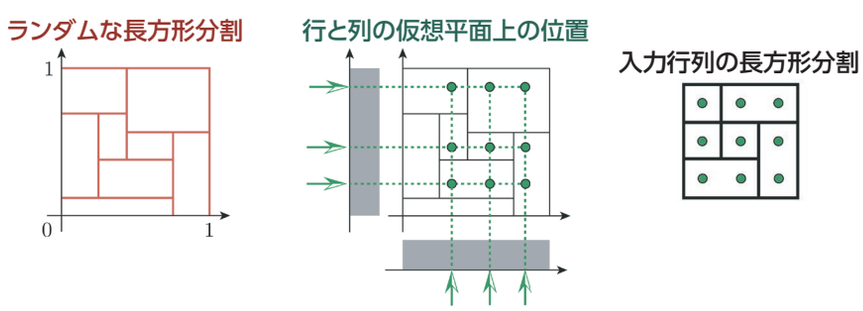

図 4　ノンパラメトリックベイズ法による長方形分割の例

ノンパラメトリックベイズ法による機械学習の課題

ノンパラメトリックベイズ法を利用した機械学習は広く使われているが、課題もあるとして、中野氏は次のように述べている。

「あえて極端に言えば、関係データに適合する長方形分割をアルゴリズムが奇跡的に見つけることを期待した手法とも言えます。もちろん無限に更新と評価を繰り返せば、理論的にはいつかは最適解に辿り着きます。ですが実用的な反復回数では悪い局所解に捕まりやすい、という問題が当初より認識されています。」

局所解に捕らわれることを回避する「超長方形分割」を提案

実用的な反復回数で最適な長方形分割を得ることが難しいという問題に対し、CS 研では「ありとあらゆる分割パターンを表現できるような特別な分割」を生成することを考えた。「極値組合せ論」という数学の分野で注目されている「超順列（superpermutation）」、「超木（supertree）」といった考え方にヒントを得ているという。簡単に言えば「とてつもなく冗長な集合を用意すると、その中に正解が含まれる」という仮定に基づき、局所解に捕ま

ることを回避するという発想だ。

「ノンパラメトリックベイズ法を用いる従来手法では、データを説明するのに必要充分なサイズの長方形分割を行います。これに対し、我々が提案する新たな手法では、あえて非常に複雑かつ冗長な長方形分割を行います。これを『超長方形分割(superrectangulation)』と呼んでいます。」(中野氏)

仮想平面上で超長方形分割を行った例を図 5 に示す。多数の長方形で構成されていることがわかる。色分けは見やすくするためのものであり、色自体に意味はない。

超長方形分割を活用する“スーパーベイズ法”を提唱

CS 研はこの超長方形分割を活用し、従来のノンパラメトリックベイズ法による機械学習を置き換えるような新たな機械学習法を提唱している。

「ノンパラメトリックベイズ法を置き換える新たなパラダイムの提唱という意味を込め、この新たな手法を“スーパーベイズ法”と呼んでいます。」(中野氏)

超冗長な長方形分割に最適解が含まれると仮定し、仮想平面上での分割を 1 度だけ行う。あとはその仮想平面上に入力行列の行と列に対する座標をランダムに生成し、評価する作業を繰り返すことにより、最適解を見つけ出す(図 6)。ノンパラメトリックベイズ法による従来手法では長方形分割も繰り返し更新するのと比較し、問題が単純化されている。

新手法による関係データ解析の実験

スーパーベイズ法による機械学習

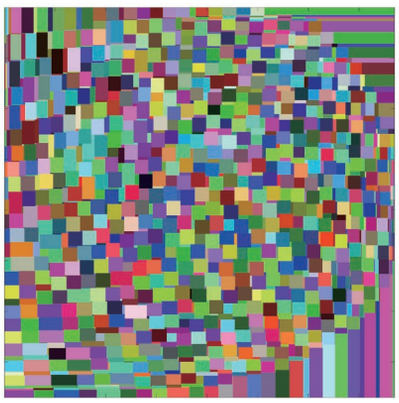

図 5　超長方形分割の例

の性能を評価した実験の結果を、表 1 に示す。実験データには SNS など各種 Web サービスのデータを利用した。いずれもネットワーク型データであり、たとえば Facebook であれば、誰と誰がつながっているかを示す。

実験ではまずランダムに選択したデータで 500 × 500 の行列を構成し、そのうち一部のデータを隠す。そして関係データ解析により、隠された部分のデータを予測した。この予測をデータセットごと、また機械学習の手法ごとに 10 回行った。

表 1 のうち、左の 3 列にはノンパラメトリックベイズ法を用いた 3 つの手法による機械学習、一番右にはスーパーベイズ法による機械学習の、予測性能を示している。左から 2 列および 3 列目は、「Nakano」と中野氏の名前が記載されていることから明らかなように、CS 研の研究成果による実験結果となっている。3 つめの PCRP は従来最も良い性能を実現していた手法だ。

各数値は左側の「1.2565」などの数値が 10 回の試行における平均値であり、この値が 1 に近いほど良い予測性能であることを意味する。また右側の「± 0.0017」など

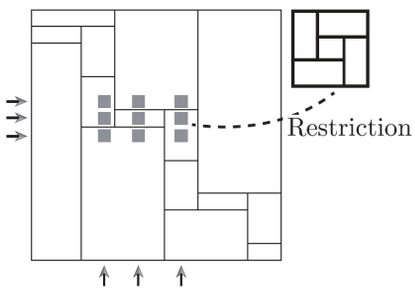

図 6　冗長な長方形分割に含まれる解

の数値は、標準偏差を示している。Twitter データの解析に関する性能を除けば、どのデータに対してもスーパーベイズ法が最も良い予測性能を示していることがわかる。

「ノンパラメトリックベイズ法については、およそ 15 年に渡る研究の結果到達した性能です。スーパーベイズ法はいきなりその性能を超えたと言えます。正確には、予測性能は既存の最も優れた手法と比較して大きな差はないのですが、スーパーベイズ法は標準偏差が全般的に小さいという点が重要です。分散が少ない、すなわち解析のバラツキが少ないということであり、悪い局所解に捕まりにくいことを意味します。」(中野氏)

性能も使い勝手も良く、さまざまな用途に適用可能

従来手法と同程度の性能と小さな標準偏差を実現した新手法のメリットについて、中野氏は次のように述べている。

「ノンパラメトリックベイズ法の場合は何度も機械学習を実行し、その中で一番良い結果を採用する必要がありました。しかしスーパーベイズ法の場合はバラツキが少ないため、1 回の機械学習であってもそれなりにリーズナブルな結果であるこ

表１　ベンチマーク結果

	Nonparametric Bayes			Super Bayes
	MP (Roy and Teh, 2009)	BBP (Nakano et al., 2020)	PCRP (Nakano et al., 2021)	Zigzag (proposed)
Wiki	1.2838 ± 0.0094	1.2712 ± 0.0056	1.2583 ± 0.0041	**1.2565** ± 0.0017
Facebook	1.1944 ± 0.0217	1.1818 ± 0.0197	1.1545 ± 0.0187	**1.1493** ± 0.0095
Twitter	1.2316 ± 0.0209	1.2146 ± 0.0058	**1.2057** ± 0.0092	1.2077 ± 0.0071
Epinions	1.4098 ± 0.0064	1.4006 ± 0.0044	1.3955 ± 0.0061	**1.3951** ± 0.0054

とが期待できます。また、ノンパラメトリックベイズ法もさまざまなユースケースで利用されている手法ですが、実用上はユースケースに合わせていろいろな工夫が必要でした。これに対しスーパーベイズ法はさまざまな用途に適用しやすいことも特徴の１つです。」

国際会議“AISTATS 2022”でスーパーベイズ法を発表

本研究はノンパラメトリックベイズ法に関する研究を積み重ねてきたCS研ならではのものであった。論文の著者にはNTTフェローである上田氏、CS研の前所長である山田氏も含まれている。

この論文の内容は2022年３月に行われた国際会議“AISTATS（International Conference on Artificial Intelligence and Statistics）2022”において発表された。AISTATS 2022では、投稿された1685の論文のうち492件が採録され、そのうち44件には追加の発表機会が与えられた。CS研の論文はその44件に含まれている。

「2012年頃に深層学習への注目が高まって以降、ノンパラメトリックベイズ法を活用した機械学習の分野は相対的に存在感がやや小さくなってしまったかもしれません。スーパーベイズ法はこの状況を覆す手法になるのでは、と期待しています。AISTATS 2022においても、予想外のことにより今までの状況をひっくり返すといったような意味を持つ“Flips the script”という言葉を使った評価をもらっています。」（中野氏）

スーパーベイズ法の普及に向け適用先を増やしていく

今後はスーパーベイズ法の適用先を広げるなど、普及に取り組むとして、中野氏は次のように述べている。

「スーパーベイズ法を提唱したものの、まだ広く知られてはいません。ノンパラメトリックベイズ法を置き換える新たな手法としてスーパーベイズ法を広めるため、関係データ解析以外の用途、たとえば巡回セールスマン問題のような複雑な組合せ最適化問題を解くような用途など、アプリケーションの例を増やすことに取り組んでいます。」

研究の出口

スーパーベイズ法は特定の用途に限らず利用可能です。実用上の様々な工夫を直観的に組み込みやすく、今すぐに実用的に使ってもらえる技術です。

ではどのような用途に最も適しているかというと、それは膨大なデータから未知の何かを推測するような用途ではないかと考えています。データの規模が大きく、組合せが複雑になればなるほど、スーパーベイズ法の強みが活きると思います。

例として新薬の開発などを目的とする生体情報の解析を挙げることができます。縦軸を遺伝子の発現パターン、横軸を病気の発症リスクとする関係データから、将来の病気のリスクを予測する、といったことも考えられます。

また、生物の進化の道筋を木の枝のように可視化した系統樹というものがあります。全体として動物や植物などの進化を表現する一方、細かな枝が個別の種の進化を表現しています。こうしたものを超長方形分割で表現するといった利用法もあるのでは、と考えています。

（中野氏 談）