

Graduate School of Data Science



YOKOHAMA
CITY
UNIVERSITY
公立大学法人
横浜市立大学

研究紹介 Research

末廣 大貴 准教授

データサイエンス専攻

最近、メディアでもよく見聞きするキーワード「AI (Artificial Intelligence)」。その成功を支えているのが、「機械学習」と呼ばれる技術である。その名の通り、機械（コンピュータ）に、学習をさせる技術である。人間が様々な物事を経験して学習するのと同様に、コンピュータも「データ」を自らの経験として学習を行う。この機械学習の技術を理論的に支えているのが、データサイエンス専攻の末廣大貴准教授である。

機械学習とその理論とは

機械学習で実現したいことというのはなんだろうか。これも大体人間と同じである。例えば、試験に合格するためには、たくさんの

模試を受けるだろう。一方で、どれだけ多くの模試を受けても、本番の試験では全く同じ問題が出ることはない。模試の答えを覚えるのではなく、問題の傾向や解法といった汎用的な知能を得ることが重要だ。

機械学習においても、このような汎用的な知能を得ることが目的となる。例えばスマホのカメラを人に向けると顔の部分に四角い枠が表示されるだろう。これは、コンピュータが様々な画像データ（学習用のデータ）をもとに、画像内の「顔らしい部分」を特定する知能を学習し、予測したからである。

ただし、模試の成績がよくても本番では間違えてしまうことがあ



末廣 大貴 准教授

データサイエンス専攻

るように、手持ちのデータ（既知データ）でうまくいったからと言って、まだ見たことのないデータ（未知データ）でもうまくいくとは限らない。既知データに対する予測性能と、未知データに対する予測性能にはギャップがある。このギャップはどの程度か、どのようにすれば埋められるかを数学的に明らかにするのが機械学習の理論である。

機械学習技術を 理論で支える

実用化のスピードに負けない理論解析を

例えば「顔の画像」と「顔でない画像」を集めてきて、「顔」or「顔でない」の分類結果がなるべく正解するような計算式を求めた結果、集めたデータに対しては95%正解できたでしょう。

一方、見たことがないデータに何%正解できるかはわからない。もし、先程のギャップが「5%以下」であれば、見たことがないデータに対しても、最低90%は正解できるということが言える。

実は、このギャップを理論的に見積もることができれば、「データって何個集めてくれば良いの？」という素朴な疑問にも理屈で答えることができるようになる(Fig.1)。

理論解析は難しい

「だんだん話が小難しくなってきたと思いますが、その通り、機械学習の理論はなかなか難しいのです」と末廣准教授は理論研究の難しさを強調する。

機械学習を実現するには、統計、線形代数、最適化をはじめとした様々な数学的知識が必要だ。先程の「顔」or「顔でない」をうまく予測するような問題を「2クラス学習問題」などと呼ぶが、他にも様々な学習問題があり、理論的に十分な解析が行われていない学習問題もたくさん存在する。世の中の色々なところにAIが使われているが、機械学習が必要とされるタスクも多種多様になってきており、理論解析が追いつけていない。すでに膨大な数の学習問題が提案されており、なお増え続けている。一つの学習問題を解析するだけでも多くの知識と労力を必要とするのに、これらの学習問題に

対し個別に解析しては理論解析は遅れをとるばかりである。

機械学習問題をまとめて理論解析する

「そこで、個別に解析するのではなく、学習問題を“まとめて解析する”のです」

学習問題のほとんどは、手持ちの学習用データ（サンプル）に対して予測誤差を最小化する問題（経験誤差最小化問題と呼ばれる）を解く。末廣准教授が最近提案した方法は、大雑把に言えば、この経験誤差最小化問題が「同じ形に変換できる」のであれば、学習問題としても同じものとして理論解析を可能にするものであり、画期的な方法として注目を集めている。

従来も学習問題を別の学習問題に変換して理論解析を行うという研究は行われていた。しかし経験誤差最小化問題の変換に着目し、様々な学習問題を「まとめて解析する」ことを可能とした方法は末廣准教授の研究が世界初である。

これまで個別解析が行われていた様々な学習問題を「まとめて」理論解析できた例の一つを紹介しよう(Fig.2)。非常にポピュラーな学習問題として「マルチクラス学習問題」がある。これは、例えば、トラ、ゾウ、ウマという答えがつけられた画像をサンプルとして学習し、新しく手に入れた画像がトラなのか、ゾウなのか、ウマなのかを正しく予測することを目標とした問題だ。

一方、こちらも長く研究が行われている「マルチインスタンス学習問題」というものがある。これは、一つのデータが複数のデータの集合で与えられるような問題である。例えば、動画は画像（フレーム）の集合とみなせる。学習用のデータは、動画の中に対象物が映り込んだかどうか、という答えはついていないが、どのフレームに映っているかという答えはついていない。目標は、新しく手に入れた動画に対象物が映り込んでいるかどうかということを正しく予測するような問題だ。

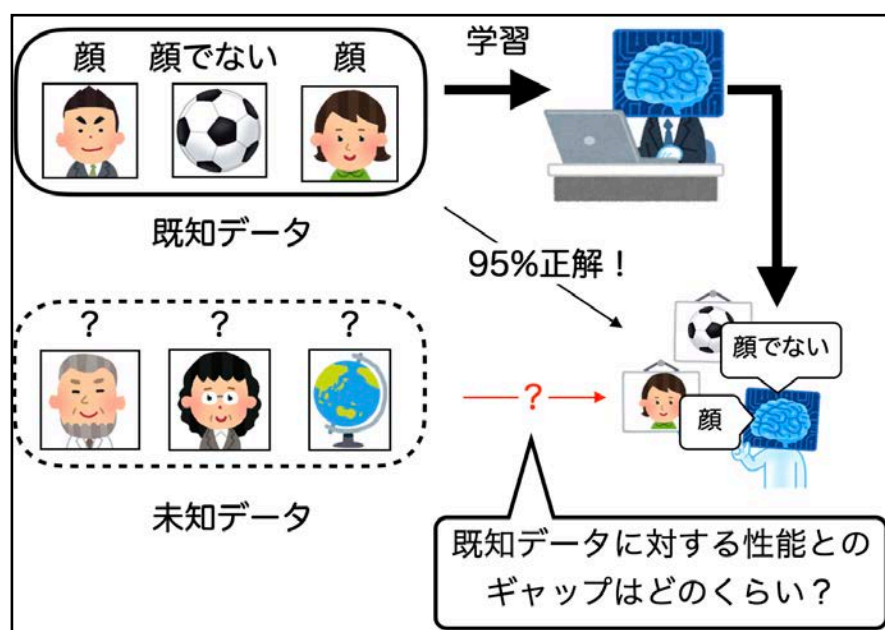


Fig.1 既知データに対する性能と未知データに対する性能のギャップ

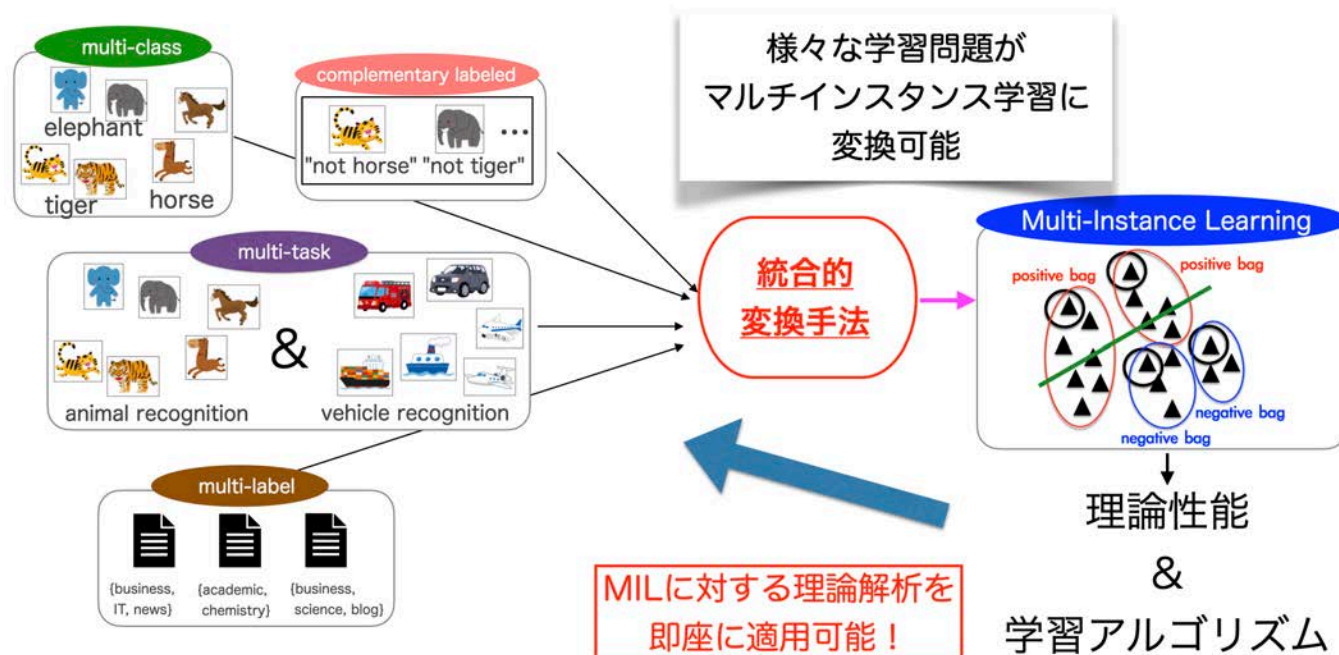


Fig.2 様々な学習問題を「まとめて」理論解析可能にした例

さて、このマルチクラス学習問題と、マルチインスタンス学習問題は、全く性質の違うような問題に見える。データの条件も、目標も全く異なるからである。しかし驚くべきことに、末廣准教授が提案した方法を用いると「マルチクラス学習問題」の経験誤差最小化問題が、「マルチインスタンス学習問題」の経験誤差最小化問題に変換できる。実はどちらの問題も20年来別々の問題として理論解析が行われていたのだが、「マルチインスタンス学習問題」に対する理論解析を十分に行えば、マルチクラス学習問題の理論解析も十分に行えたことになる。

「詳細は割愛しますが、マルチラベル学習問題、補ラベル学習問題、マルチタスク学習問題といっ

たポピュラーな学習問題に加えて、実に7つの学習問題が“マルチインスタンス学習問題”に変換できることもわかりました」

挫折から生まれた研究

「実は、初めから学習問題の統合化をやりたいと思っていたわけではありません。大きな挫折をもとに始まった研究なのです」

当初は時系列データにおける局所波形を使った分析を行っており、機械学習の枠組みでその定式化、理論解析に取り組んでいた。

「理論性能の証明を含め3年ほどかけてようやく論文として出版できたのですが、実はこの問題も“マルチインスタンス学習問題”として定式化できることが、研究成果が揃ってきた頃に気づきました。こ

れまでの成果が無駄になったというわけではありませんが、もっと早くに気づいていれば分析を大いに簡素化できた可能性が高く、正直かなりショックでした」と末廣准教授は当時を振り返る。そして、同じような犠牲者を生まぬためにも、「理論解析が同じようにできる学習問題はまとめるべきだ」という強い思いを持ったという。

理論から実データ解析へ

このような挫折の経験も含め、理論研究で得た知識は、実データ解析にも大いに活かされている。末廣准教授は、これまで様々な分野の研究者との交流を通じて、教育データ分析、画像認識、文字認識、医療画像解析など、多様な実

“スマホや車はもちろんボールペンにだって、普段身近に使っているあらゆる技術には理論があり、機械学習も理論なしには生活に定着できません”

データ解析に関わってきた。ドメインやデータの性質などによって、学習問題は大きく異なる。また、理論研究ではしばしば都合の良い仮定を置いて議論する。あまりに現実離れしているような仮定は少ないものの、そのため現実のデータではうまく行かないことも多い。特に医療画像解析分野は問題設定が非常に豊富で、実問題に取り組んでいるうちに新しい理論課題を手に入れることも多い。

機械学習を理論で支える

機械学習は「科学技術」として

はまだまだ幼い。幼いからこそ、その進歩は凄まじい。ほとんどの人がスマートフォンを持ち、テキスト、画像、音声、位置情報、購買情報といったデータが大量に蓄積される情報化社会、著しく性能が向上している計算機など、機械学習の発展はまさに絶頂期にある。

末廣准教授は、実用化のスピードに負けないよう、理論解析のスピードアップを可能とする技術の開発に日々取り組んでいる。

「スマホや車はもちろん、ボールペンにだって、普段身近に使っ

ているあらゆる技術には理論があり、機械学習も理論なしには生活に定着できません」

「理論保証を与えることで、安心・安全、信頼できるAIを理論的側面からサポートすることができるのです」と末廣准教授は自身の研究動機を熱く語っている。何より機械学習分野は研究者の探究心をくすぐる魅力的な未解決問題に溢れているのだ。



Activity Report



ヘルスデータサイエンス専攻 博士前期課程入学試験を実施しました

令和5年7月22日（土）に、ヘルスデータサイエンス専攻博士前期課程入学試験が実施され、10名が合格されました。



データサイエンス専攻 博士前期課程入学試験を実施しました

令和5年8月22日（火）に、データサイエンス専攻博士前期課程入学試験が実施され、12名が合格されました。



横浜市立大学サマーデザインワークショップ2023を開催しました

データサイエンス研究科が主催組織の一つとなっている「横浜市立大学サマーデザインワークショップ2023」が、令和5年9月11日（月）から13日（水）にかけて、対面（みなとみらいサテライトキャンパス）とオンラインによるハイブリッドで開催されました。参加者数は、ファシリテーター6名、見学者12名を含めて、全体で40名でした。

Schedule



- 令和5年9月23日（土）
データサイエンス専攻 博士後期課程入学試験およびヘルスデータサイエンス専攻 博士後期課程入学試験を実施します。
出願期間は8月21日（月）～8月23日（水）です。
合格発表日は10月6日（金）です。
- 令和6年2月3日（土）
データサイエンス専攻 博士後期課程入学試験を実施します。
出願期間は1月4日（木）～9日（火）です。
合格発表日は2月16日（金）です。
- 令和6年2月3日（土）
ヘルスデータサイエンス専攻 博士前期課程入学試験を実施します。
出願期間は1月4日（木）～9日（火）です。
合格発表日は2月16日（金）です。



公立大学法人 横浜市立大学大学院

データサイエンス研究科 ニュースレター
令和5年9月号

© 2023 データサイエンス研究科