

特集 「人工知能研究のベンチマークとは—標準問題・データセット・評価手法—」

# 機械学習コンペティションの進展と今後の展開

## Current and Future Directions in Machine Learning Competitions

馬場 雪乃  
Yukino Baba

京都大学大学院情報学研究科知能情報学専攻  
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University.  
baba@i.kyoto-u.ac.jp

**Keywords:** machine learning, predictive modeling, crowdsourcing.

### 1. 機械学習コンペティションの進展

データから未知の事象を予測する技術である予測モデリングは、データ活用の現場から生まれた新しい問題・新しい種類のデータに対応する形で研究開発が進められてきた。例えば、データが逐次的に与えられる状況で予測モデリングを行うためにオンライン学習の研究が進み、また、ソーシャルメディアの普及に伴いソーシャルネットワーク解析手法の研究が促進された。現実の問題設定やデータは、予測モデリング研究の起爆剤としてなくてはならない存在である。このような研究開発上の要請に応えるために、企業などから提供される実際のデータを用いて予測モデリング手法の性能を競う機械学習コンペティションが、機械学習・データマイニング分野の国際会議などで併設開催されるようになってきた。参加者を競わせるコンペティション形式は、予測精度という定量的な評価指標をもつ予測モデリングに適しており、参加意欲をかき立てる意味でも効果的である。近年では、機械学習コンペティション専用のWebプラットフォームも登場し、コンペティションの数・参加者数ともに増加の一途を辿る。

機械学習コンペティションは、研究開発の促進だけでなく、予測モデリングの実用方法の確立の点でも重要な意義をもつ。実際のデータに予測モデリング手法を適用する際に、最先端の手法が常に高い性能を発揮するとは限らず、適切なモデリング手法の選択が求められる。また、データの前処理・特徴抽出の工夫も予測精度向上に貢献することが多い。特定のデータで高い予測精度を達成するためには、データの性質に合わせてテラーメイト方式で解析手法を構築する必要がある。このような研究開発と実用のギャップを埋める知見を共有する場としても、機械学習コンペティションは機能している。

さらに機械学習コンペティションは、特定の問題・データを提供し組織外のデータサイエンティストに解析を依頼する、予測モデリングのクラウドソーシングと捉えることもできる。従来の機械学習コンペティションでは参加者同士を競わせていたが、目的が優れた予測モデルの

獲得であるならば、参加者を協調させる仕組みを取り入れた新しい形式のコンペティションも考えられる。

本稿では、機械学習コンペティションが予測モデリング手法の研究開発に貢献した事例を紹介する。また、機械学習コンペティションを運営するうえでの課題と、コンペティションの新しい活用方法についての展望も述べる。

### 2. 機械学習コンペティションの概要

#### 2.1 機械学習コンペティション実施の流れ

コンペティションによって多少の違いはあるが、多くの機械学習コンペティションは次のような流れで開催される。まず、コンペティション参加者には、モデリング用データと評価用データが提供される。モデリング用データには予測結果の正解が含まれているが、評価用データに対する正解は与えられていない。参加者はモデリング用データを用いて予測モデルを構築し、モデルを用いて評価用データに対する予測結果を出力して提出する。コンペティション運営者は、提出された予測結果と、運営者だけが見ることができる評価用データの正解を照らし合わせ、予測結果のスコアを算出する。スコアはリーダーボードに提示され、公開される。

参加者は、リーダーボードからのフィードバックを参考に予測モデルを改善し、コンペティション終了まで結果提出を繰り返す。コンペティションによっては提出回数に制限が設けられているが、決められた回数以内であれば何度でも結果を提出することができる。コンペティション終了時点のスコアで、最終順位が決定する。予測モデルがリーダーボードに適応しすぎるのを防ぐため、コンペティション開催中は、中間スコアのみがリーダーボードに掲載されることもある。中間スコアは、評価用データの一部を用いて算出されるものである。この場合、最終順位は、残りの評価用データから算出された最終スコアにより決定される。

#### 2.2 国際会議併設の機械学習コンペティション

近年、機械学習・データマイニングに関連するさま

ざまな国際会議で機械学習コンペティションが開催されているが、その先駆けとなるのが、データマイニングの国際会議 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) に併設する KDD Cup である。1997 年に第 1 回目のコンペティションが開催され、以降 2015 年現在まで、毎年開催されている。

KDD Cup ではいずれの年度においても実践的な問題設定・データが題材とされている。例えば 1999 年の KDD Cup は、マサチューセッツ工科大学から提供された TCP データを用い、ネットワークへの侵入検知問題を扱った。2003 年は、プレプリントサーバ arXiv が提供する論文の LaTeX ソースデータや引用ネットワークを利用した、引用予測問題が対象となった。2011 年は、Yahoo! Music で収集されたユーザの楽曲評価データを利用して、音楽推薦手法が競われた。2015 年の KDD Cup では、中国の大規模オンライン公開講座プラットフォーム XuetangX からアクセスログが提供され、生徒の受講離脱率予測が題材となった。

KDD 以外にも、機械学習とデータマイニングの国際会議 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) では 1999 年から、ECML/PKDD Discovery Challenge というコンペティションが開催されている。KDD Cup と同様の実践的課題が扱われており、例えば 2015 年のコンペティションでは、自転車シェアリングシステムの各ステーションにおける自転車が在庫予測問題と、タクシーの GPS データを利用した目的地予測問題という実応用を強く意識した問題が扱われた。そのほか、CIKM, ICDM, PAKDD, RecSys, WSDM といったデータマイニングに関する複数の国際会議で機械学習コンペティションが開催されるようになり、コンペティション開催が一種のトレンドになりつつある。

### 2.3 機械学習コンペティションプラットフォーム

多数の機械学習コンペティションが開催されるようになった背景には、コンペティション開催を支援する Web プラットフォームの登場がある。そのようなプラットフォームの代表例に、Kaggle<sup>\*1</sup> がある。Kaggle は機械学習コンペティション開催で求められる基本的機能として、予測結果の評価、リーダーボード、ユーザ管理、データ公開などのツールを提供する。コンペティション主催者は Kaggle を利用することで、自分でシステムを用意することなく容易にコンペティションを開催することができる。Kaggle は国際会議併設のものに限らずさまざまな機械学習コンペティションの開催で使用されており、その数は 100 を超える。

Kaggle で開催される多くのコンペティションでは企業が主催者である。企業は自身が所有するデータを参

加者に提供し、予測問題の解決を依頼するためにコンペティションを開催している。例えば米国の保険会社 Allstate は、交通事故車の情報から保険請求額を予測するコンペティションを開催し、ドイツの医薬品メーカー Merck は、化合物の活性予測コンペティションを開催した。参加者にとってコンペティションは、さまざまな分野における実データを用いて解析の腕試しができる格好の場である。また、企業にとっては社外の専門家リソースを活用できるという利点がある。Kaggle で開催されるほとんどのコンペティションは賞金付きであり、賞金総額が 100 万ドルを超えるものも存在する。

### 2.4 特化型の機械学習コンペティション

国際会議併設の機械学習コンペティションや、Kaggle で開催されるコンペティションでは、開催期間はほとんどの場合数か月程度である。一方、特定の問題を対象として、数年にわたって開催された機械学習コンペティションも存在する。そのようなコンペティションの代表例を二つ紹介する。一つめは、米国でオンライン DVD レンタルサービスを運営する Netflix 社が、2006 年から開催した Netflix Prize である<sup>\*2</sup>。Netflix Prize では、ユーザの映画評価データを利用した映画推薦が題材となった。同社はすでに Cinematch という推薦システムを開発しており、Cinematch による推薦精度を 10% 以上改善した参加者に対して 100 万ドルの賞金を支払うという条件のもとで、コンペティションが開催された。開始から 3 年後の 2009 年にこの条件を満たすチームが登場し、コンペティションは終了した。

もう一つが、大規模画像認識のコンペティション ImageNet Large Scale Visual Recognition Challenge (ILSVRC) である<sup>\*3</sup> [Russakovsky 15]。2010 年から毎年開催されているこのコンペティションでは、ImageNet という大規模画像データセットを利用して、画像中の物体認識問題などに参加者は取り組む。企業からの参加も多く、注目度の高いコンペティションである。

## 3. 機械学習コンペティションの貢献

### 3.1 特定の問題に適した予測モデリング手法の研究促進

機械学習コンペティションは、予測モデリング手法の研究開発の促進と実用方法の確立に大きく貢献してきた。例えば Netflix Prize と ILSVRC では、それぞれ推薦と画像認識という特定の問題においてコンペティションを開催し、多くの研究者を巻き込むことに成功した。結果、新しい予測手法が多数提案され、実用面での数多くの知見が蓄積された。

Netflix Prize には 5 万人以上が参加し、推薦手法の研

\*1 <https://www.kaggle.com>

\*2 <http://www.netflixprize.com/>

\*3 <http://www.image-net.org/challenges/LSVRC/>

究を大きく進展させた。特に、行列分解と  $k$  近傍法が予測精度向上に貢献し [Rendle 12a], 推薦問題におけるこれらの手法の有効性を明らかにした。優勝チームによるレポートでは, Netflix Prize が研究コミュニティにもたらした貢献が二つあげられている [Koren 09]。一つめが, Netflix Prize のデータセットを活用して, 多くの研究者が新しく推薦手法の研究開発に取り組み, 多数の研究成果が発表されたことである。実際, 優勝チームによる論文だけでも 3 本が KDD, ICDM というデータマイニングの一流国際会議に採択されている [Bell 07a, Bell 07b, Koren 08]。二つ目が, これまでは高度な推薦手法の研究が進められてきたが, 実は評価値の平均と分散などの基本統計量から構成される単純な予測モデルが, 予測精度向上に貢献することを明らかにした点である。予測モデリングの現場において, 高度な手法よりも単純な手法が時に効果を上げることは経験的に知られていたが, 多くの研究者が高度な手法を検討した Netflix Prize コンペティションにおいても同様の結果が得られたことは意義深い。

Netflix Prize が, 予測モデリングの実用において大きく寄与した点の一つがある。それは, モデル統合の威力を実証したことである。モデル統合とは, 複数の予測モデルの出力を受け取り予測結果を返す, 新しいモデルを構築する手法である。例えば, 一部のデータを使って複数の予測モデルをそれぞれ学習したのち, 各モデルの出力を特徴量とする別の予測モデルを残りのデータで学習することで, モデル統合を実現できる。Netflix Prize の優勝チームは, 100 個以上のモデルを 2 段階で統合し, 最終的な予測モデルを構築した [Töschler 09]。Netflix Prize での成功を受けて, 現在では機械学習コンペティションや予測モデリングの現場において, モデル統合は基本的なテクニックとして活用されている。

ILSVRC は, 特に深層学習の発展に貢献してきた。オーガナイザの報告によると, 2012 年に畳込みニューラルネットワーク (CNN) [Krizhevsky 12] が画像分類と物体位置検出の両方で優勝したのが, 大規模画像認識研究におけるターニングポイントとなった [Russakovsky 15]。これを受けて ILSVRC 2013 では多くのチームが CNN を採用し, 2014 年のコンペティションでは, CNN をベースに 22 層から成るネットワークを構築した GoogLeNet [Szegedy 14] が, 六つ中二つのサブタスクで優勝した。そのほかのタスクでも CNN を活用したチームが優勝し, 大規模画像認識のさまざまなタスクで CNN が性能を発揮することを明らかにした。Netflix Prize と同じく ILSVRC も, さまざまな研究者が大規模画像認識研究に取り組むきっかけとなり, 特に深層学習の研究を進めるうえでの重要なベンチマークコンペティションとなっている。

### 3.2 予測モデリング手法の汎用性を報知する場の提供

予測モデリング手法の研究者にとって, 機械学習コ

ンペティションやそのデータセットは, 考案した手法の汎用性を広くアピールする機会となっている。実際, Kaggle で開催された, 脳波測定結果から手の動きを予測するコンペティションにおいて 2 位入賞者は, 「自分が考案した手法 (Recurrent Convolutional Neural Network [Liang 15]) が時系列データでも有効であるか検証するためにコンペティションに参加した」と述べている\*4。

機械学習コンペティションを通じてその汎用性を実証してきた手法の一例が, 行列分解と教師付き学習を組み合わせた予測手法 Factorization Machine (FM) である [Rendle 10]。考案者の Steffen Rendle 自らが FM を用いて, Kaggle で開催された生徒の回答予測コンペティションで優勝し\*5, KDD Cup 2012 のトラック 1 (ソーシャルメディアの友人関係予測問題) で 2 位, トラック 2 (広告のクリックスルー率予測) で 3 位と高い成績を修めることで, さまざまな予測問題における FM の有効性をアピールした [Rendle 12b]。さらに, タグ推薦を扱った ECML/PKDD Discovery Challenge 2009 のデータセットを用いて, FM が優勝手法と同等の性能であることも示している [Rendle 10]。FM はライブラリが公開されており\*6 [Rendle 12a], 各種機械学習コンペティションでの成功を受けて, 現在もコンペティションで広く活用されている。FM を拡張した Field-aware Factorization Machines (FFM) も, Kaggle で開催された二つの広告クリックスルー率予測コンペティションで優勝している。FFM もライブラリが公開されており\*7, さまざまなコンペティションで利用されている。

同様に, コンペティションでその威力を発揮してきた予測手法に, 正則化付き勾配ブースティングを用いた XGBoost がある\*8 [Chen 15]。XGBoost は, Kaggle で開催された文章分類コンペティション, マルウェア分類コンペティションなどで優勝し, その汎用性を裏付けた。FFM も XGBoost も, その詳細は学術論文ではなく, スライドやライブラリで最初に公開されたという点は興味深い。従来の予測手法の提案は, その詳細と評価結果を記した論文を通じて行われるのが一般的であった。一方, 機械学習コンペティションの活発化により, まずライブラリを公開し, コンペティションを通じてその性能を利用者に評価してもらうという新しい形式の提案・評価方法が採用され始めている。機械学習コンペティションが, 予測モデリング手法の新しい研究開発プロセスを牽引しているといえるだろう。

\*4 <https://www.kaggle.com/c/grasp-and-lift-eeg-detection/forums/t/16617/team-daheimao-solution>

\*5 <http://blog.kaggle.com/2012/04/20/viva-libfm-steffen-rendle-on-how-he-won-the-grookit-challenge/>

\*6 <http://www.libfm.org>

\*7 <http://www.csie.ntu.edu.tw/~r01922136/libffm/>

\*8 <https://xgboost.readthedocs.org/>

### 3.3 実用上の知見の蓄積

国際会議併設の多くの機械学習コンペティションでは、会議期間中にそのコンペティションに関するワークショップが開催され、実際に用いた解析手法について入賞者が発表を行う。ワークショップを通じて、参加者間で予測手法や特徴量、データの処理方法についての情報が共有され、同種の予測問題に挑むデータ解析者にとっての貴重な情報源となっている。

Kaggle ではコンペティションごとにフォーラムが設置されており、関連する論文の紹介や、特徴量や予測手法に関する議論などが活発に行われている。参加者による試行錯誤の記録からは、多くを学ぶことができる。入賞者による手法の紹介記事も定期的に掲載されており、特に初学者やコンペティションの新規参加者にとっては、予測モデリングに取り組む際の足掛かりとなっている。

## 4. 機械学習コンペティション運営上の課題

### 4.1 予測モデルの評価方法

予測モデリング手法の研究開発を促進するために機械学習コンペティションは有用ではあるが、実際にコンペティションを運営する際の課題は数多く残されている。例えば、提出された予測モデルの評価方法である。予測モデルがリーダーボードに適応しすぎるのを防ぐため、コンペティション開催中は一部の評価用データから算出した中間スコアだけを公開する方法が、Kaggle などで用いられている。中間スコアを用いたコンペティションでは、中間スコアに適応しすぎて残りの評価用データから算出される最終スコアが大きくなって下がる事態を防ぐために、参加者は過学習に注意しながら予測モデルを構築しなければならない。つまり、中間スコアを過信しすぎると最終的に入賞できない可能性があり、これは参加者の立場からすると、フィードバックを信じられない不便な状況である。

一つの解決策として、Ladder アルゴリズムと呼ばれる方法が提案されている [Blum 15]。従来のコンペティションでは、新しい予測結果を提出すると決められた評価指標によって予測精度が算出され、それがスコアとなった。Ladder アルゴリズムは、算出した予測精度が、提出時点までの予測精度の最良値を超えている場合は、従来と同様に予測精度をそのままスコアとして割り当てる。それ以外の場合、提出時点までの最良値を予測結果のスコアとして採用し、予測結果自身の予測精度は利用しない。このような単純な修正により、より正しいスコアを提示できることが理論的に示されている。

参加者の間でも、予測モデルの評価方法は議論の的になっている。2015 年の KDD で開催された、KDD Cup の歴代優勝者によるパネルディスカッションでは、あらかじめ決められた評価用データを用いる現在の評価方法は実際の予測モデリングの運用からは乖離していること

が指摘された\*9。一方、評価用データを固定するのではなく変化させる方法も考えられるが、このやり方ではリーダーボード上の順位変動が大きくなり参加意欲をそいでしまうという懸念も示された。コンペティション運営者は評価方法を設計するうえで、予測モデリングの実運用に則した方法を用いるとともに、参加者のやる気を引き出すよう配慮しなければならない。適切な評価方法について、機械学習コミュニティ内でのさらなる議論が求められる。

### 4.2 コンペティションの公平性担保

入賞者に多額の賞金が支払われるコンペティション、あるいは、入賞が名誉となる人気コンペティションの運営者は、不正行為をする参加者の存在に注意しなければならない。例えば ILSVRC 2015 では、あるチームが複数アカウントを作成し、制限を大幅に上回る数の予測結果を提出したことが問題となった\*10。多数の予測結果を提出し多くのフィードバックを獲得すると、予測モデルの構築で有利となるため、コンペティションを公正に運営するうえでこのような行為は問題となる。一方、ILSVRC 2015 の場合は、当該チームが論文を公開したためにそこから不正行為が発覚したが、通常は複数アカウントの検出は簡単ではなく、コンペティション運営上の大きな課題となっている。

別の種類の不正行為として、外部データの利用がある。一部のコンペティションでは、主催者から提供されるもの以外のデータを利用した予測モデル構築を禁じている。参加者ごとにアクセス可能な外部データには差があるという不平等性に配慮した措置でもあるし、また、特に企業が主催者の場合に、得られた予測モデルの実運用を考えての措置でもある。特に問題となるのは、外部データから正解が判明する場合である。実際、IJCNN 2011 に併設で開催された Social Network Challenge では、外部データから正解を割り出し、つまり「カンニング」することで優勝するチームが登場した [Narayanan 11]\*11。このコンペティションでは、写真共有サービス Flickr における友人関係予測問題が扱われた。提供されたデータではユーザ ID が仮名化されていたが、データ上の各ユーザと Flickr ユーザのひも付けができれば、正解が判明する。優勝チームは、Flickr のデータを独自にクロールしたうえで、ネットワーク構造を利用してコンペティションデータ上のユーザと Flickr ユーザのマッチングを行い、正解を獲得した。

予測結果のみを提出させるコンペティションでは、外部データ利用の有無を確認するのは困難である。入賞者にソースコード提出を要求する方法が考えられるが、実

\*9 <https://www.linkedin.com/pulse/present-future-kdd-cup-competition-outsiders-ron-bekkerman>

\*10 <http://www.image-net.org/challenges/LSVRC/announcement-June-2-2015>

\*11 なお優勝チームがとった方法は、コンペティション主催者によって許可されている。

際の検証にはコストが掛かる。現状は、データの出所の秘匿化が、コンペティション用データと外部データのひも付けを難しくするための有効策となっている。

#### 4.3 データからのプライバシー情報漏えい

コンペティション用データを参加者に公開する際に運営者は、データからのプライバシー情報の漏えいにも配慮しなければならない。実際に、**Netflix Prize** のデータにプライバシー情報漏えいの危険性があることが報告されている [Narayanan 06]。この報告は、ある **Netflix** ユーザによる映画評価結果がごくわずかにわかっているだけで、そのユーザと **Netflix Prize** データ中の仮名化されたユーザの対応付けが可能であることを明らかにした。また、**Netflix Prize** 参加者へのデータ提供により映画視聴情報に関するプライバシーを侵害されたとして、**Netflix** のユーザ数名が訴訟を起こしている\*12。プライバシー侵害を一つの理由として、予定されていた2回めのコンペティションは中止となった\*13。このように、プライバシー情報の漏えいリスクは、企業がコンペティションへのデータ提供を躊躇する一因となっており、機械学習コンペティション普及上の大きな課題として残されている。コンペティション運営者は、コンペティションの公平性とプライバシー情報漏えいの危険性に配慮したうえでコンペティション用データを整備することが求められている。

### 5. 機械学習コンペティションの新たな展開

#### 5.1 クラウドソーシングによる予測モデリング

予測モデリング技術を競うために開催される機械学習コンペティションであるが、コンペティションに問題・データを提供する主催者にとって機械学習コンペティションは、世界中のデータサイエンティストを活用して予測モデリングを依頼する場だと捉えることもできる。このように考えると、コンペティション主催者の目的は優れた予測モデルを獲得することであり、参加者を競わせる形式はあくまで一つの手段にすぎない。そこで、参加者を協力させる仕組みを取り入れることも考えられる。実際、チームでの参加は従来の機械学習コンペティションでも認められており、コンペティションが進むなかでコンペティション参加者同士がチームを組むケースも見られる。チーム構築をさらに促進するため、チームメイト推薦などの工夫は効果的と考えられる。予測モデリングでは特徴選択・モデル設計・モデル最適化など多様なスキルが求められるため、特にスキルに基づいた推薦は有効だろう。

参加者同士の協力を促すための報酬メカニズムも提案さ

れている [Abernethy 11]。このメカニズムでは、ある参加者がモデルを提出するとそれが他の参加者に公開され、別の参加者がそのモデルの改善を行うことができる。これを何度も繰り返し、最終的に各ステップにおけるモデル改善の度合いに応じて報酬が支払われる。勝者総取りとは異なるこのような報酬メカニズムの採用により、参加者が自身の専門性を生かして自然と協力することが期待される。

参加者から提出したモデルを主催者が統合することでより良い予測モデルを得るという使い方も考えられる。著者らは実際に、国内の機械学習コンペティションプラットフォーム **CrowdSolving** においてリンク予測問題を出題し、得られた百数十個のモデルを統合すると、優勝モデルよりも予測精度の高いモデルを獲得できることを確認した [Baba 14]。統合時に学習したモデルごとの重みを確認すると、単独で予測精度の高いモデルに常に大きい重みが与えられるとは限らず、精度は低くても他のモデルと異なる傾向の予測をするモデルに大きな重みが与えられる場合があった。この結果は、統合モデルの精度向上には予測の多様性が重要であることを示唆している。モデル統合を前提とするコンペティションでは参加者から多様なモデルを獲得するために、参加者ごとに異なる部分データを与えることや、参加者ごとに異なる初期手法を提示するなどの工夫が有効と考えられる。

#### 5.2 探索型データ解析におけるコンペティション

予測モデリングとは別の種類のデータ解析方法に、探索型データ解析がある。探索型データ解析はデータの視覚化・集計などを通じてその中身を調べ、知見を獲得する作業である。探索型データ解析を対象にしたコンペティションも開催され始めている。例えば2000年の **KDD Cup** では、予測モデリングのほかに「購買履歴データを使って、高額購入者の特徴を分析してください」というような課題も対象とされた [Kohavi 00]。また、**Kaggle** でも、米国の人口統計データの分析コンペティションなどが開催されているし、**CrowdSolving** でも、電子書籍配信サービスのログ分析コンペティションが開催されている。

予測モデリングの場合には、予測精度という定量的な評価指標を用いてコンペティションの勝敗を決定することができる。一方、探索型データ解析では、適切な評価指標を用意するのが難しい。2000年の **KDD Cup** では、提出結果それぞれについて専門家が、プレゼンテーション・内容の正しさ・重要性などの観点で評価を行ったが、提出数が増えると多大な労力が必要となる。解決策として期待されるのが一般のユーザに評価を依頼する方法で、実際に **Kaggle** では、分析結果に投票する機能が導入されている。また、一般のユーザによる分析結果評価の信頼性は高いという実験結果もある [Baba 15]。探索型データ解析のコンペティション普及のため、ユーザを巻き込んだ評価プロセスの洗練化が求められる。

\*12 <http://www.wired.com/2009/12/netflix-privacy-lawsuit/>

\*13 <http://blog.netflix.com/2010/03/this-is-neil-hunt-chief-product-officer.html>

## 5.3 学習教材としての機械学習コンペティション

予測モデリングの専門家が切磋琢磨する場が機械学習コンペティションであったが、実は予測モデリングの学習者にとっても、機械学習コンペティションは有益な学習教材である。実践的な課題とデータが提供され、結果を提出すると速やかに評価が返される機械学習コンペティションの仕組みは、予測モデリングの実用技術を身につける教材として適している。データサイエンティスト教育における機械学習コンペティションの有用性は注目されはじめており、例えば **Kaggle** では、教育現場でのコンペティション開催を支援するための **Kaggle in Class** というツールが提供されている。このツールを用いると、予測結果の評価やリーダーボード、フォーラムなどの **Kaggle** が提供する機能をそのまま用いて独自のコンペティションを開催することができる。

教育目的により特化した機械学習コンペティションのプラットフォームとして、著者らが運営するビッグデータ大学<sup>\*14</sup>がある。**Kaggle** と異なりビッグデータ大学では、ほとんどのコンペティションでは賞金が定められていない。データ解析の実践演習の場として設計されており、講義などでの利用も積極的に進められている。初学者でも参加しやすいように多くのコンペティションでチュートリアルが提供され、また、学習用教材として入賞者からのレポートも公開されている。一方、学習効果を上げるためには継続参加を促す動機付けが求められる。ソースコード付きのチュートリアルの公開や、講義との連携、複数のコンペティションの結果を反映したユーザーランキングの提供が動機付けに効果的であることが経験的に明らかになってきている。適切な動機付けの仕組みを取り入れた教育型機械学習コンペティションプラットフォームの構築は、データサイエンティスト人口の拡大、ひいては予測モデリング手法の研究開発の促進とその利活用の活性化のために重要だと考えられる。

## ◇ 参考文献 ◇

- [Abernethy 11] Abernethy, J. D. and Frongillo, R. M.: A collaborative mechanism for crowdsourcing prediction problems, *Advances in Neural Information Processing Systems 24 (NIPS)* (2011)
- [Baba 14] Baba, Y., Nori, N., Saito, S. and Kashima, H.: Crowdsourced data analytics: A case study of a predictive modeling competition, *Proc. 2014 Int. Conf. on Data Science and Advanced Analytics (DSAA)* (2014)
- [Baba 15] Baba, Y. and Kashima, H.: Crowdsourcing data understanding: A case study using open government data, *An adjunct to the Proc. 3rd AAAI Conf. on Human Computation and Crowdsourcing (HCOMP)* (2015)
- [Bell 07a] Bell, R. M. and Koren, Y.: Scalable collaborative filtering with jointly derived neighborhood interpolation weights, *Proc. 2007 7th IEEE Int. Conf. on Data Mining (ICDM)* (2007)
- [Bell 07b] Bell, R., Koren, Y. and Volinsky, C.: Modeling

- relationships at multiple scales to improve accuracy of large recommender systems, *Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)* (2007)
- [Blum 15] Blum, A. and Hardt, M.: The ladder: A reliable leaderboard for machine learning competitions, *Proc. 32nd Int. Conf. on Machine Learning (ICML)* (2015)
- [Chen 15] Chen, T. and He, T.: Higgs boson discovery with boosted trees, *Proc. NIPS 2014 Workshop on High-energy Physics and Machine Learning (HEPML)* (2015)
- [Kohavi 00] Kohavi, R., Brodley, C. E., Frasca, B., Mason, L. and Zheng, Z.: KDD-Cup 2000 organizers' report: Peeling the onion, *ACM SIGKDD Explorations Newsletter*, Vol. 2, No. 2, pp. 86-93 (2000)
- [Koren 08] Koren, Y.: Factorization meets the neighborhood: A multifaceted collaborative filtering model, *Proc. 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)* (2008)
- [Koren 09] Koren, Y.: The BellKor Solution to the Netflix Grand Prize, [http://www.netflixprize.com/assets/GrandPrize2009\\_BPC\\_BellKor.pdf](http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf) (2009)
- [Krizhevsky 12] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems 24 (NIPS)* (2012)
- [Liang 15] Liang, M. and Hu, X.: Recurrent convolutional neural network for object recognition, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2015)
- [Narayanan 06] Narayanan, A. and Shmatikov, V.: How to break anonymity of the Netflix Prize dataset, *CoRR*, Vol. abs/cs/0610105 (2006)
- [Narayanan 11] Narayanan, A., Shi, E. and Rubinstein, B. I.: Link prediction by de-anonymization: How we won the Kaggle Social Network Challenge, *Proc. 2011 Int. Joint Conf. on Neural Networks (IJCNN)* (2011)
- [Rendle 10] Rendle, S.: Factorization machines, *Proc. 2010 IEEE 10th Int. Conf. on Data Mining (ICDM)* (2010)
- [Rendle 12a] Rendle, S.: Factorization machines with libFM, *ACM Trans. on Intelligent Systems and Technology*, Vol. 3, No. 3, pp. 57:1-57:22 (2012)
- [Rendle 12b] Rendle, S.: Social network and click-through prediction with factorization machines, *KDD Cup 2012 Workshop Proceedings* (2012)
- [Russakovsky 15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L.: ImageNet large scale visual recognition challenge, *Int. J. Computer Vision*, Vol. 115, No. 3, pp. 211-252 (2015)
- [Szegedy 14] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going deeper with convolutions, *CoRR*, Vol. abs/1409.4842 (2014)
- [Töscher 09] Töscher, A., Jahrer, M. and Bell, R. M.: The Big Chaos solution to the Netflix Grand Prize, [http://www.netflixprize.com/assets/GrandPrize2009\\_BPC\\_BigChaos.pdf](http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf) (2009)

2016年1月20日 受理

## —— 著者紹介 ——



馬場 雪乃 (正会員)

2009年東京大学大学院情報理工学系研究科修士課程修了。2012年同研究科博士課程修了。2012年より東京大学大学院情報理工学系研究科特任研究員。2014年より国立情報学研究所およびJST ERATO河原林巨大グラフプロジェクト特任助教。現在、京都大学大学院情報学研究所知能情報学専攻助教。博士(情報理工学)。データマイニング、ヒューマンコンピューテーションの研究に従事。

\*14 <http://universityofbigdata.net>