1J5-OS-18b-2

データ解析コンペティションを用いたクラウドソーシングによる予測モデルの構築

Development of Predictive Model Using Crowdsourcing Competitions

馬場 雪乃[*1*2] Yukino Baba　則 のぞみ[*3] Nozomi Nori　齊藤 秀[*4] Shigeru Saito　鹿島 久嗣[*3] Hisashi Kashima

[*1]国立情報学研究所ビッグデータ数理国際研究センター
Global Research Center for Big Data Mathematics, National Institute of Informatics

[*2]JST, ERATO, 河原林巨大グラフプロジェクト
JST, ERATO, Kawarabayashi Large Graph Project

[*3]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, Kyoto University

[*4]株式会社オプト データサイエンスラボ
Data Science Lab, OPT, Inc.

Competitions for predictive modeling provide a new data mining approach where a large group of experts examine a wide variety of predictive models and compete with others to build models with the best performance. Competition hosts, who provide their own dataset and specify the problem to be solved, are not only able to select the best competition model but also to aggregate the submitted models to obtain one that outperforms all other individual models. In this paper, we report the results of a study conducted on CrowdSolving, a competition platform for predictive modeling in Japan. We hosted a competition for solving a link prediction problem and observed that (i) the prediction performance of the winning model was better than that of one using a state-of-the-art approach, and (ii) the model aggregated from the submitted models outperformed the winning model.

1. はじめに

データに潜む規則や関係を学習し未観測データの予測に用いる予測モデリングの研究では，様々な性質のデータや問題に対処するために日々新しいアルゴリズムが開発されている．一方，特定のデータを対象にした実際のモデリングにおいて洗練された現代的な予測アルゴリズムが常に最高の予測精度を達成するとは限らない．性能の向上はしばしば，既存手法の選択と特徴量の設計や正規化，サンプル選択などのデータ固有のヒューリスティクスの組み合わせによってもたらされる．このことは「どんな場合でもうまくいく方法はない」ことを示すノーフリーランチ定理にも支持される．

少人数のデータ解析技術者がデータに合ったモデルを広範囲に探索するのは困難であるが，近年ではデータ解析コンペティションを用いることで多人数による試行が実現可能となった．Kaggle[*1]などのデータ解析コンペティションでは，データ所有者が予測問題とデータを提供しコンペティションを開催することで，複数のデータ解析技術者に予測モデル構築を依頼することができる．技術者達は，コンペティション開催期間中に適宜モデルを提出し，提示される予測精度を他の参加者と競い合いながらモデルの改善に取り組み賞金獲得を目指す．コンペティションの開催によりデータ所有者は，予測モデルの広範囲探索を外部委託しより良いモデルの獲得が可能となる．

データ解析コンペティションは，たとえば研究者に対して様々なアルゴリズムのベンチマークを提供することを目的として，データマイニング系の主要国際会議である KDD に併設された KDD カップが 1997 年より毎年開催されている．2010 年に登場した Kaggle は，主に研究者を対象としていたデータ解析コンペティションの門戸を広く一般に広げたといえる．Kaggle は，データ所有者に対しては予測モデル構築の外部委託手段を提供しデータ解析技術者には腕試しと賞金獲得の場を与えることで，データ解析発注者と技術者の結びつけを支援している．Kaggle では，これまでに 100 を超えるコンペティションが開催され世界中から 12 万人が参加した．日本国内でも 2013 年にインフォコム株式会社が主催する CrowdSolving[*2]がサービスを開始し，2014 年現在までに 5 個のコンペティションが開催されている．

多くのデータ解析技術者を競わせ多様な予測モデルを試行させることで得られる恩恵は，競合結果としての最良のモデルの獲得だけではない．提出された個別の予測モデルを組み合わせることで優勝モデルを上回る予測精度のモデルを獲得できる可能性もある．たとえば，米国の DVD レンタルサービスで用いる推薦アルゴリズムの開発を題材にしたデータ解析コンペティション Netflix Prize では，多数の予測モデルを統合したモデルが優勝した [Töscher 09]．Netflix Prize では巨額の優勝賞金を獲得するために複数のチームが自発的に協働し，さまざまな予測モデルを作り上げ統合手法を構築したが，比較的少額の賞金付きコンペティションでは参加者の協働を期待するよりは独立にモデルを構築させ，最終的にコンペティション開催者が統合を行う方法の方が汎用性が高いだろう．

我々は，予測モデル構築の新しい方法としてのデータ解析コンペティションの有効性を検証するため，データ解析コンペティションプラットフォーム CrowdSolving において実際にコンペティションを開催し実験を行った．まず，複数のデータ解析技術者によるモデル探索の効果をはかるために，参加者から提出された個々の予測モデルと我々が用意した汎用的な予測手法の予測精度を比較した．また，参加者が独立に構築したモデルを簡単な手法により統合した際の精度を確認するため，全ての提出モデルを統合し作成した予測モデルと個々の予測モデルの予測精度を比較した．結果，単純な手法にデータ固有の

連絡先: 馬場 雪乃，国立情報学研究所ビッグデータ数理国際研究センター，ybaba@nii.ac.jp

*1 https://www.kaggle.com

*2 https://crowdsolving.jp/

ヒューリスティックを組み合わせた上位入賞モデルが汎用的手法の予測精度を上回ることを確認した．また，各予測モデルの出力を特徴量として用い学習した統合モデルが優勝モデルの予測精度を上回ることを確認した．

2. データ解析コンペティションの開催

2.1 コンペティション開催概要

データ解析コンペティションプラットフォーム CrowdSolving で実際にコンペティションを開催し，複数のデータ解析技術者がそれぞれ構築した予測モデルの予測結果を獲得した．コンペティションは 2013 年 8 月 14 日から 9 月 15 日までの 33 日間開催された．最終的に，16 名の参加者によって 134 個の予測モデルが構築されその予測結果を得た．また，上位入賞者 5 名の最終的な予測モデルのプログラムと予測手法の概要を記したレポートを得た．

コンペティションへの参加募集は CrowdSolving サイト上で実施した．参加希望者は予測に用いる正解付きの訓練データと予測対象のテストデータをダウンロードし，予測結果を提出することで，開催期間中いつでもコンペティションに参加することができた．優勝賞金の総額は 100,000 円に設定し，1 位から 5 位まで順に 50,000 円，20,000 円，15,000 円，10,000 円，5,000 円と傾斜配分することを参加希望者に告知した．CrowdSolving にて開催済みのコンペティションよりも低い金額設定[*3]にすることで，また，本コンペティションを参加者の技術向上を目指すための「チャレンジコンペティション」と銘打ちコンペティション終了後に上位入賞者の手法を公開すると告知することで，特に学習者の参加を促した．予測モデル構築手法としてのデータ解析コンペティションの有効性を検討するという本実験での目的においては，熟練者ばかりが参加するよりも経験が浅い学習者が参加した方がより一般的な結果が得られると考えたためである．なお，本コンペティションが学術研究の一環である旨は公表しなかった．

開催期間中に参加者が自身のモデルを改良していけるよう，一日に一度，各参加者が提出した最新の予測精度のランキングを公開した．ランキングでは，テストデータのうちの半分を対象として算出した予測精度を提示した．参加者には，テストデータのどの部分でランキング用の予測精度が算出されているのかは開示されていない．この方法は，最終的な勝敗を予想し難くすることで参加意欲を高めるために，Kaggle や CrowdSolving の他のコンペティションでも採用されている．

2.2 予測問題：Wikipedia 記事間リンク予測

コンペティションの題材には，グラフにおけるノード属性情報付きのリンク予測問題を採用した．この問題は関係予測の問題と二値分類問題が組み合わさったものであり参加者の創意工夫が期待できる．Wikipedia のデータを用い，記事をノード，記事間のハイパーリンクをリンク，記事が所属するカテゴリをノード属性情報として問題を設計した．リンクは有向とした．コンペティション参加者には，記事の属性と，リンクが存在する記事ペアの一部を正解付き訓練データとして提供した．リンクの有無を隠した記事ペア集合をテストデータとして提供し，テストデータ中の各ペアのリンク存在率を予測し提出するよう参加者に依頼した．参加者には，提供されるデータが Wikipedia の記事間リンクであることは開示した．外部データの利用を防ぐために，記事 ID をランダムに振り直すことで各記事が Wikipedia のどの記事に該当するのかはわからないようにし，また，記事の属性がカテゴリ情報であることは隠した．予測精度の指標には ROC 曲線下面積 (AUC) を採用した．

予測に用いるデータは，Wikipedia における様々なデータを構造化した DBpedia[*4]を用いて作成した．所属するカテゴリ名に “problem”, “science”, “medical”, “medicine”, “social” のいずれかの単語を含む英語版の記事 23,269 個を取得し，合わせて記事間のリンク情報も取得した．予測対象のテストデータにコールドスタート，すなわち訓練データには現れずテストデータで初めて現れる記事を含めることで問題を難化させた．参加者には訓練データとして，45,209 個の正解ラベル付き記事ペアと，23,269 個の記事それぞれについて 39,541 次元の属性情報を提供した．また，テストデータとしてリンクの有無を隠した記事ペア 78,426 個を提供した．テストデータ中，リンクが存在するペア（正例）は 39,118 個であった．参加者に正例と負例の比率は提示しなかった．

3. 実験

予測モデル構築手法としてのデータ解析コンペティションの有効性を検討するために，参加者から提出された個々の予測モデルと我々が用意した汎用的な予測手法の予測精度を比較し，複数のデータ解析技術者によるモデル探索の効果を確認した．また，提出された予測結果を統合して作った予測モデルと個々の予測モデルの精度を比較し，モデル統合の効果を確認した．

3.1 コンペティション結果概要

図 1 に，開催期間中に各参加者が提出した予測モデルの精度の変化を示す．初日は Participant-3 が AUC 0.6735 で 1 位となり三日めまでその順位を維持したが，四日めに Participant-11 が AUC 0.8249 と大きく精度を向上させ，さらに五日めに Participant-8 が 0.8953 という高い AUC 値を達成した．その後，七日めから参加した Participant-9 が，最初に提出したモデルは AUC 0.5468 と高い予測精度ではなかったが日を追うにつれモデルを改善し，コンペティション開始二週間後にはその時点での 1 位（Participant-8）のモデルの AUC 0.9316 に迫る精度 0.9256 を達成した．しかし，結局 Participant-8 は抜き去られることなく最終日まで 1 位の座を守った．

上位入賞者から提出されたレポートによれば，5 名の入賞者は「リンク予測を二値分類問題として捉え教師あり学習手法を用いて予測した」グループ（Participant-8, 10）と「リンク存在率の指標を設計し予測に用いた」グループ（Participant-9, 12, 15）に分けられる[*5]．最終的に AUC 0.94 を超えるという高い予測精度を達成した Participant-8 と 9 が全く異なる手法を採用していた事実は興味深い．

訓練データでは正例しか与えられていないため，前者のグループはいずれも擬似負例を作成していた．また，前者のグループは記事ペアの特徴を表現するために多くの種類の特徴量を設計し利用していたが，後者のグループは与えられたノード属性情報等の限られた情報を用いてリンク指標を設計したという違いがあった．教師あり学習手法を用いたグループは，二名ともランダムフォレストを学習手法として用いていた．二名は共通して，記事が所属するカテゴリ数を記事の特徴として，カテゴリ情報の類似度を記事ペアの特徴として使用したが，Participant-8 は記事の ID や記事ペアの ID の差も特徴量として用い，Participant-10 は訓練・テストデータでの記事の登場回数や逆向きリンクの有無を特徴量に用いた点が異なって

*3 本コンペティション以前に開催されたものはいずれも賞金総額 800,000 円であった

*4 http://wiki.dbpedia.org/

*5 予測手法の詳細は https://crowdsolving.jp/node/629/summary で公開されている．

いた．リンク存在率指標を設計したグループは，Participant-9 と 15 は記事特徴だけを用いたのに対して，Participant-12 は記事ペアの特徴だけを使用したという違いがある．

3.2 汎用的手法と提出された予測手法の比較

コンペティション開催に先立ち事前分析としてリンク予測のための汎用的手法を用いた実験を行い予測精度の評価を行った．これは，今回のデータに固有の性質を考慮せずにリンク予測の汎用的な手法を用いた場合にはどの程度の精度が達成可能かを見るためのベンチマークとしての意味合いを持つ．

本実験では二つの汎用的手法を用意した．いずれの手法でも，参加者に提供したものと同じ訓練データを用いて学習を行った．一つめは，複数のリンク予測指標を特徴量として用い，教師あり学習手法を用いて予測モデルを学習する手法である．記事間リンクと，記事とカテゴリ間のリンクのそれぞれについて，4 種類のリンク指標（共通隣接ノード指標，Jaccard 係数，Adamic/Adar 指標，優先的選択指標）を算出し[*6]，8 次元の特徴ベクトルを作成した．L1 正則化付きのロジスティック回帰を用いて各特徴量の重みを学習した．

二つめは，非線形次元削減手法 [Nori 12] である．これは，リンク情報とノード属性情報を用いる汎用的な多項関係予測手法であり，リンクが存在する記事同士の持つ属性が潜在空間で近くなるように属性から潜在空間への写像を学習する[*7]．

図 1 に提出された各モデルと汎用的手法の比較結果を示す．リンク指標を特徴量に用いた手法 (LM) は AUC 0.7434，非線形次元削減手法 (NLDR) は AUC 0.7166 であった．コンペティションの開始後三日間は，参加者から提出されたモデルの予測精度をこれらの手法が上回っていたが，四日めには Participant-11 に抜き去られ，最終的には上位 4 つの予測モデルが AUC 0.9 を超えるという汎用的手法を大きく上回る精度を達成した．これほどの予測精度の向上は，我々が用いた汎用的手法のような単一の予測手法の改善では困難であるが，上位入賞者が用いた既存の手法とリンク指標等の特徴生成のヒューリスティクスが今回のデータの性質を上手くとらえた結果，高い精度がもたらされたと考える．この結果は，モデルの広範囲探索を可能としたデータ解析コンペティションの利点を顕著に示している．

3.3 統合手法と提出された予測手法の比較

最後に，提出された予測結果を特徴量として用い新しい予測モデルを学習する統合手法と提出された個々の予測モデルを比較した．統合手法では，決定木を弱学習器とする集団学習手法の一つ Gradient Tree Boosting を用いて予測モデルを学習した [Friedman 01]．学習時には，一部のテストデータについての予測結果を特徴量とした予測モデルを構築し，残りのテストデータの予測に用いた．つまり，統合手法では参加者に提供した正解ラベル付き訓練データ以外に一部のテストデータに対する正解ラベルも使用することに注意されたい．

図 1 に各モデルと統合手法の比較結果を示す．統合手法の適用の際には，各日にちにおいて，その日までに提出された全ての予測結果を用いて学習を行った．ただし，提出された各モデルとは異なり，一部のテストデータ（図 1 では 5,000 個）を正解ラベル付きの追加訓練データとして用いた．統合手法の評価時には，訓練に用いなかった残りのテストデータを予測対象としている．また，図 1 では統合手法については，訓練に用いるデータのランダムサンプリング・学習・評価を 10 回繰り返した際の AUC の平均値を提示している．

初日に提出されたモデルの最高精度は AUC 0.6735 であるが，統合手法はこの時点で平均 AUC 0.8479 という高い予測精度を達成した．提出されたモデルがこの値を超えるのは五日めのことである．五日めに Participant-8 が AUC 0.8593 と大きく予測精度を向上させると，この予測結果を特徴量に含めた統合手法も精度を大きく向上させ 0.9342 となった．統合手法の五日め時点での予測精度に提出モデルが追いつくのはコンペティション開始後 22 日めのことである．最終的には，統合手法は優勝モデルの AUC 0.9459 という予測精度を大きく上回る 0.9823 を達成した．これは分類器の予測精度としては驚くほど高い値である．また，統合対象の予測モデルの精度が向上するほど統合手法の精度が向上するのは明らかであるが，5 個の予測結果しか提出されていなかった初日の時点でも統合手法が高い予測精度を達成した事実は注目に値する．

各提出モデルと統合手法では学習に用いる訓練データの数が異なるため，より公正な比較を行うために，同数の訓練データを用いて学習した場合の優勝手法と統合手法の予測精度を比較した．たとえば，元々の訓練データに追加してテストデータからランダムに選んだ 1,000 個の訓練データも用いる場合，統合手法では元々の訓練データに対する各提出モデルの予測結果を特徴量として 1,000 個の訓練データを使って統合モデルの学習を行う．優勝手法では，元の訓練データとあわせた 46,209 個の訓練データを使って学習する．ただし，統合手法と優勝手法では追加する訓練データの数は同じだが異なるサンプルを選んだ．優勝手法には擬似負例の作成が含まれているため追加訓練データとしては正例だけを用いた．一方，統合手法では擬似負例を作成しても，擬似負例に対する提出モデルの予測結果が得られないため，正負同数のサンプルを選び追加訓練データとした．両手法の評価時には，各手法で訓練に用いたデータをテストデータから除外し，残りを予測対象とした．また，この比較では，統合手法は全提出結果 134 件を特徴量として用いた．

優勝手法ではランダムフォレストの適用時に 10,000 個の決定木を構築しているが，学習時間短縮のために本実験では，決定木の数は 100 とした．代わりに，元々の訓練データに対して 10,000 個の決定木と 100 個の決定木で学習を行った際の AUC 値の差分を算出し，各訓練データ数における 100 個の決定木での AUC 値に差分を上乗せして優勝手法の予測精度とした．

結果を図 2 に示す．追加訓練データが 100 個未満の時には統合手法は優勝手法を上回ることはできなかったが，100 サンプル以上ではいずれの場合でも統合手法は優勝手法よりも高い予測精度を示している．追加訓練データの数が少ないときには，統合手法は少数のサンプルで学習しなければならないため予測精度が安定しないが，追加訓練データ数が増え安定した学習ができるようになると，統合手法は優勝手法を上回るという結果が得られた．

4. むすび

多人数のデータ解析技術者による広範囲探索という，予測モデル構築の新しい方法としてのデータ解析コンペティションの有効性検証を目的として，実際にコンペティションを開催し実験を行った．結果として，(1) 上位入賞モデルが我々が用意した汎用的手法を上回る予測精度を達成すること，(2) 統合手法が優勝手法を上回る予測精度を達成することの二点を確認した．一般にデータ解析コンペティションにおいては，賞金やその他の報酬や動機付けにより，参加者により精度の高いモデル

*6 各リンク指標の詳細は [鹿島 07] を参照されたい．
*7 三つ以上の多項関係データの予測のために提案された手法であり，今回の題材のような二項関係に特化した手法ではない．

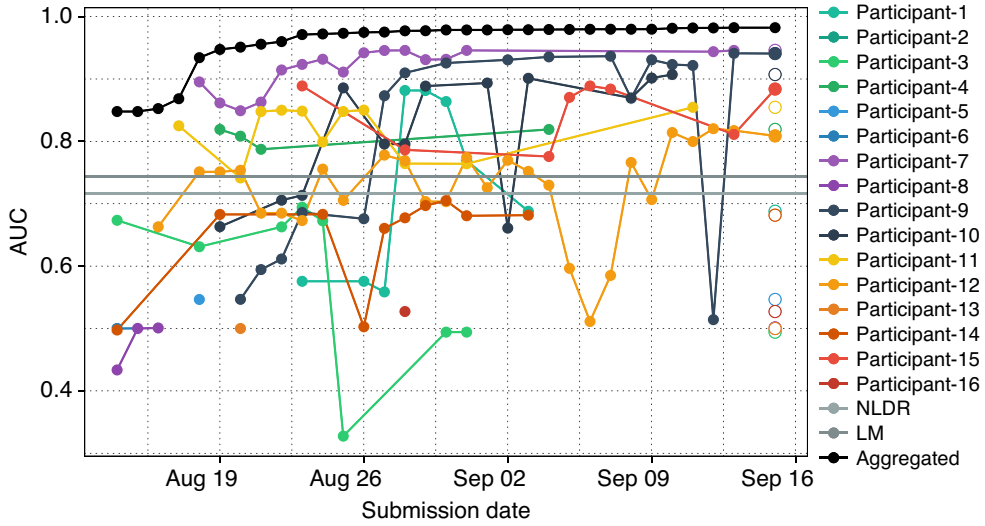

図 1: 開催期間中に参加者から提出された予測モデルの精度の変化と，汎用的手法と統合モデルの予測精度との比較．予測精度は，全てのテストデータを用いて測定した AUC．Participant1〜16 は，各日にちに各参加者から提出されたモデルの精度を示す．ただし，一日に複数個のモデルが提出された場合には，精度最高のモデルの AUC だけを示している．白丸は，参加者の最終ランキングの決定に使われた予測モデルの精度を示す（ただし，最終日にモデルの提出があった場合には白丸のプロットを省略した）．LM と NLDR は，著者らが事前分析として構築した予測モデルであり，それぞれ，リンク指標を特徴量にする手法と非線形次元削減手法である．Aggretaged は，各日にちまでに提出された予測結果を特徴量として用いる統合手法である．統合手法は，テストデータ中の 5,000 ペアを訓練に用いて構築した．また，統合手法の AUC は，訓練に用いる記事ペアのテストデータからのランダムサンプリング・学習・評価を 10 回繰り返した際の平均値を示している．

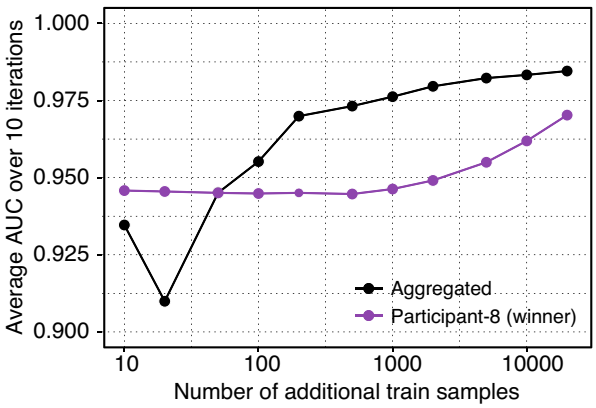

図 2: テストデータの一部を追加訓練データとして用いた際の優勝手法 (Participant-8) と統合手法 (Aggregated) の比較．統合手法は，最終日までに提出された全ての予測結果を特徴量として用いている．両手法の AUC は，追加訓練データのランダムサンプリング・学習・評価の手順を 10 回繰り返した際の平均値を示している．

を構築してもらうことが主たる目的であるが，今回の結論は，参加者の予測結果を統合するモデルを構築することで，一人の優勝者のモデルよりも精度の高いモデルを構築することができることを示唆している．

データ解析コンペティションにおける賞金設定の工夫についての理論的研究は既になされているが [Abernethy 11]，実際のコンペティション開催における報酬設定の効果を調査することは今後の課題の一つである．また，統合手法で上手く分類ができなかったサンプルを新たな訓練データとして参加者に提供し予測モデル構築を依頼するような，統合モデルの予測精度を高めることを目標としたコンペティションの設計も，データ解析コンペティションを実用的な予測モデル構築手法にするための重要な課題だと考えられる．

5. 謝辞

コンペティションの開催とデータ分析にご協力頂いたインフォコム株式会社の福江一起氏に感謝する．また，本実験で開催したコンペティションの全参加者，ならびに予測モデルのプログラムとレポートの提出にご協力頂いた，コンペティション上位入賞者に感謝する．

参考文献

[Abernethy 11] Abernethy, J. and Frongillo, R. M.: A Collaborative Mechanism for Crowdsourcing Prediction Problems., in *Advances in Neural Information Processing 24*, pp. 2600–2608 (2011)

[Friedman 01] Friedman, J. H.: Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, pp. 1189–1232 (2001)

[Nori 12] Nori, N., Bollegala, D., and Kashima, H.: Multinomial Relation Prediction in Social Data: A Dimension Reduction Approach, in *Proceedings of the 26th AAAI Conference on Artificial Intelligence* (2012)

[Töscher 09] Töscher, A., Jahrer, M., and Bell, R. M.: The BigChaos Solution to the Netflix Grand Prize (2009)

[鹿島 07] 鹿島 久嗣, ネットワーク構造予測, 人工知能学会誌, Vol. 22, No. 3, pp. 344–351 (2007)