基般

01

国際チームにおけるモデリング手法とチームガイドライン

小嵜耕平（(株）リクルートテクノロジーズ ）

InterContinental Ensemble

KDD Cup 2015 の優勝チーム「InterContinental Ensemble」は 9 人から構成される国際チームである．メンバはそれぞれアメリカ，オーストリア，日本，中国，シンガポールに分散している．多くのメンバがデータマイニングの競技プラットフォーム「Kaggle」でのコンテストや Netflix Prize などのコンテストでの受賞歴を持つ，コンテスト愛好者の集まりである．本稿では優勝チームの用いたモデリング手法について解説し，チームとしてアンサンブルモデルを作り上げる上で効果的であったチームガイドラインについて解説する．

モデリング手法

アンサンブルとは，複数のモデルを組み合わせることで新しいモデルを作る手法の総称である．アンサンブルにおいては個々のシングルモデルの多様性を高めることが重要となる．多様なアルゴリズムと多様な特徴量セットを用いてチームメンバが独立にシングルモデルを作成し，これらのシングルモデルをアンサンブルすることで最終的な予測モデルを構築した．

チームのメンバがシングルモデルの学習に用いたアルゴリズムは Gradient Boosted Decision Tree（GBDT），Neural Network（NN），Random Forest（RF），Extremely Randomized Tree（ET），Factorization Machine（FM），K-Nearest Neighbor（KNN），Kernel Ridge Regression（KRR），Logistic Regression（LR）の計 8 種類である．これらのアルゴリズムを用いて合計 64 個のモデルを用意した．アンサンブルには多段のスタッキングを用いた．スタッキングは個別のモデルの予測結果を特徴量として使用し（これはメタ特徴量と呼ばれる），予測モデルを構築する手法である．多段のアンサンブルモデルを構築する理由は，各アンサンブルモデルがお互いに補い精度を改善できる余地を残すほどの多様性を持っていると期待できるためである．最後に構築したモデルから線形回帰による三段目のアンサンブルモデルを構築した（**図 -1**）．

チームガイドライン

チームでは情報共有のツールとして Skype，Dropbox，GitLab（Wiki やバージョン管理システムなどのコラボレーションツールを提供するサービス）を用いた．これに加えて，共通のチームガイドラインを持つことでメンバ各自が独立してチームに貢献できる体制を作った．ガイドラインは次の 4 点の決め事で構成される．

1. 訓練データに対するメタ特徴量は，必ず事前に共有された 5 分割交差確認のインデックス（分割方法を定義）に従って生成する．
2. モデル作成に使った特徴量を Dropbox で共有し，必要に応じてコードを GitLab で共有する．
3. 各モデルの交差確認セットとテストセットにおける予測結果を Dropbox で共有する．これらはアンサンブルモデルにおけるメタ特徴量として扱う．
4. 各モデルの交差確認時のスコアと Public Leaderboard Score（コンテスト側が暫定的なランキングを提供するためにテストデータの一部で評価したスコア．以下，LB スコアと省略）を Wiki に記録する．

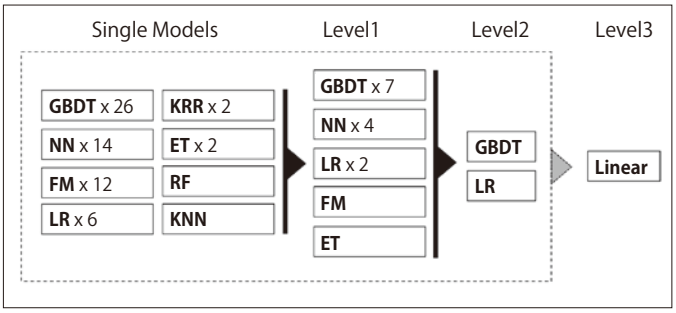

図 -1　最終的に構築したモデルの構成

図 -2　LB スコアと交差確認におけるスコアの関係

このガイドラインによって，チームメンバは特徴量抽出・シングルモデル・アンサンブルモデルのどの部分にも貢献できるようになった．たとえば，特徴量抽出やシングルモデルを構築するアイディアがなくても，アンサンブルモデルにおけるアイディアがあればほかのチームメンバが作った特徴量やメタ特徴量を Dropbox からダウンロードしてアンサンブルモデルを作ることができる．

4. の記録はチームメイトのモデルが過学習しているか否かを判断する上で役に立つ．**図 -2** は各モデルの LB スコアと交差確認のスコアを散布図にプロットし，ロジスティクス回帰分析の結果を重ねたものである．図中の矢印は交差確認のスコアと LB スコアとの差異が大きいモデルである．このモデルは訓練データセットにとってのみ有効なモデルであり，過学習していると見なせる．このようなモデルはアンサンブルモデルに用いても同様に過学習を引き起こしてしまうため，アンサンブルモデルから排除した．この結果，アンサンブルにおける過学習を避けることができた．

チームにとって重要なこと

チーム InterContinental Ensemble はチームガイドラインが非常に効果的に機能したため，チームワークによって優勝を勝ち取ることができた．メンバ間では，Skype 上でチームマージ後からコンテスト終了時までの 3 週間の間に 1,905 回のメッセージのやり取りが行われた．その一方で特別なデータ分析基盤は何も用いなかった．主催からは機械学習のクラウドプラットフォーム提供に関する告知が参加者になされたが，これも利用しなかった．チームにおいて何よりも重要なのは，自身に足りないものを補い合い，チームとして機能するためのコミュニケーションと最小限の枠組みであると思われる．

（2015 年 10 月 31 日受付）

小嵜耕平 ■ i@ho.lc

2011 年に奈良先端科学技術大学院大学情報科学研究科博士前期課程修了（工学），2014 年に同博士後期課程単位認定退学．2015 年（株）リクルートテクノロジーズに入社．