基般

03

メンバの技術的バックグラウンドの多様性を活かした組織的データ分析

米川　慧（（株）KDDI 研究所 ）　秋山卓也（（株）KDDI 研究所 ）

体制と進め方

本チームは（株）KDDI 研究所のさまざまな部署に所属する社員を中心に構成された 12 人のチームである．KDD Cup では，個人・チームともに予測データの 1 日あたりの提出回数が 5 回までと制限されている．そのためチームで取り組むにあたっては，さまざまな予測結果からどれを提出するかといった意思決定や，スコアを次に繋げるための情報共有が重要となった．また競技を進めるにつれ，今回のテーマでは機械学習モデルの選択や各モデルのパラメータチューニングよりも，予測の元になる特徴量（説明変数）の充実度が重要であることが分かってきた．そこで本チームでは，メンバの役割を特徴量担当と機械学習担当に分け，相互の連携を図ることで，より効果的な特徴量を作り出した．**図 -1** に取り組みの全体概要を示す．

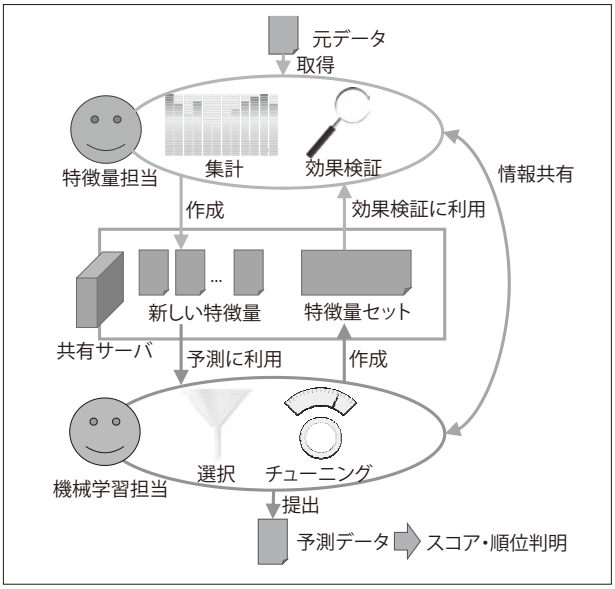

図 -1　チーム体制の概要図

特徴量担当

特徴量担当は元データから，機械学習モデルのインプットとして有効な特徴量を考案，作成した．特徴量担当は，ネットワークからアプリケーションまで多様な技術的バックグラウンドを持つメンバで構成されており，各々が異なる技術的知見から特徴量を考案することで多種多様な特徴量を生み出した．

機械学習担当

機械学習担当は元々データ分析の知見を持つメンバによって構成され，共有された特徴量の選択，機械学習モデルのチューニングを通じて予測データを作成した．機械学習モデルとして主に使用したものは，XGBoost ☆1，H2O Deep Learning ☆2，Regularized Greedy Forest ☆3 である．

情報共有

特徴量担当と機械学習担当との連携を強める上で，一番重要なのは相互の情報共有である．特徴量担当による新たな特徴量の共有，機械学習担当による予測結果と利用した特徴量セットの共有，特徴量担当による効果検証の共有を円滑にすることで，特徴量の質を高めていくための組織的な分析ループを実現した．特に競技中盤以降，誤った予測をしがちなユーザへの対応で効果を発揮し，100 位前後から一時は 4 位まで順位を上げた．本チームでは情報共有として，週次でミーティングを開催するとともに，プロジェクト管理ツールの 1 つである Backlog ☆4 のフ

☆1　GitHub dmlc/xgboost，https://github.com/ dmlc/xgboost
☆2　H2O.ai –H2O Deep Learning，http://h2o.ai/product/deep-learning/
☆3　Regularized Greedy Forest（RGF）in C++，http://stat.rutgers.edu/home/tzhang/software/rgf/
☆4　Backlog，http://www.backlog.jp/

ァイル共有機能，メーリングリスト機能および連絡板機能を利用した．

代表的な分析例

ここでは，前章のようなチーム体制のもとに作成され，特にスコアの上昇に貢献した代表的な特徴量を紹介する．これらは作成した特徴量の中のごく一部ではあるが，メンバの技術的バックグラウンドの多様性を活かした特徴量の典型的な例である．このように組織的なデータ分析のもとに作られた特徴量が，スコアの上昇に大いに貢献し，総合 6 位という結果をもたらした．

受講ユーザ数予測に基づく離脱率指標

この特徴量は，各講座の離脱判定期間中の受講ユーザ数（＝離脱しないユーザ数）を予測し，その予測結果と他の特徴量を組み合わせてユーザごとの離脱率を推定したものである．離脱判定期間中の受講ユーザ数予測には ARIMA[1] と呼ばれる時系列予測モデルを利用した．ARIMA は無線基地局の建設計画等のためのトラフィック予測等に用いられるモデルである．

類似ユーザからの推定特徴量

この特徴量は，予測したいユーザの受講講座の組合せから，似たような講座の受け方をしているほかのユーザを推定し，それら類似ユーザの受講ログを利用して作成したものである．これは，予測したいユーザの受講ログが少ない場合に有効な特徴量を作成できないという課題から取り組んだものである．類似性の導出には，協調フィルタリングと呼ばれる，e コマースで商品のレコメンドなどによく用いられる技術を利用した．**図 -2** は協調フィルタリングを利用した特徴量作成の例である．

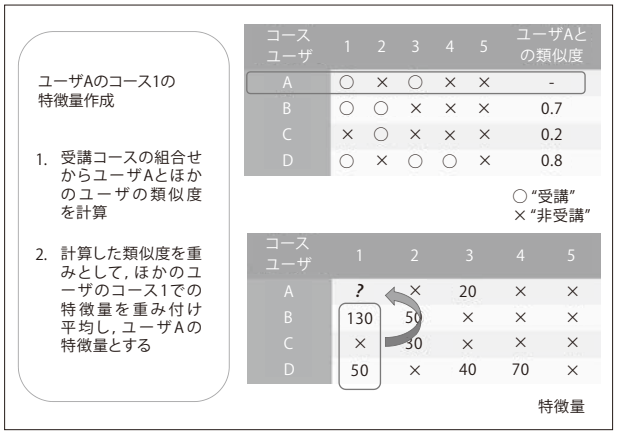

図 -2　特徴量作成の例

参考文献
1）Box, G. E. P. and Jenkins, G. M.: Time Series Analysis : Forecasting and Control, Holden-Day, San Francisco (1970).
（2015 年 10 月 30 日受付）

米川　慧 ■ ke-yonekawa@kddilabs.jp
2014 年 KDDI（株）に入社．クラウドサービス基盤の運用保守に従事後，（株）KDDI 研究所へ出向し現在に至る．

秋山卓也 ■ ta-akiyama@kddilabs.jp
2013 年 KDDI（株）に入社．光回線の運用保守に従事後，（株）KDDI 研究所へ出向し現在に至る．