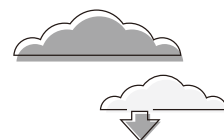


組織に創発現象を起こす クラウド型データ分析環境



西川大亮（新日鉄住金ソリューションズ（株））

本稿では、KDD Cup 2015 で2位に入賞したチーム（FEG&NSSOL@DataVeraci）の取り組み内容において、他チームと比べて特徴的な部分を中心に解説する。

グループ企業内での合同チーム

筆者の所属チームは、システムインテグレータの新日鉄住金ソリューションズ（株）（NSSOL^{☆1}）5名と、NSSOLの100%子会社で金融分野を中心にデータ分析やモデル化を主業とする（株）金融エンジニアリング・グループ（FEG）9名の計14名の合同で結成された。チームの中心メンバは、NSSOL、FEGともにデータマイニングの実務経験を持ち、一部はKDD CupやKaggleなどで開催されるオープンコンペティションへの参加経験があった。

例年のKDD Cupでは、技能向上を目的として両企業それぞれで主に個人で参加していた。今回14名の合同チームを結成した目的は、従来の技能向上に加え今後データマイニング分野の市場が拡大することを想定し、課題を個人ではなくチームで解決するための知見の獲得と、チーム戦に必要なクラウドサービスの実証検証であった。

データ分析統合環境

チーム名にあるData Veraci（ダータヴェラーチ）^{☆1}は、NSSOLが2014年から開発／運用している、データ分析を共同で行うためのクラウドサービス型の統合環境である。特徴は単なるデータ分析アプリケーションではなく、タスク管理や文書検索、テキストチャットなどデータ分析業務に必要なIT環境全体を用意して

いる点である。これによりデータ分析プロジェクトの素早い立ち上げと、遠隔地からの支援、IT環境の統一による効率化、セキュリティの確保^{☆2}などを狙っている。

今回は地理的に離れたオフィス間での共同作業ができた点や、検討時のソースコードがそのまま別のメンバで実行できることで他メンバの検討状況が容易に把握できた点で、クラウド型の統合環境が有効であった。また情報共有により重複のない仮説検証や、課題解決までの時間短縮にも効果があった。結果として個人では思い浮かばないような仮説を立て、各メンバが自然に役割を分担し実行できる体制が構築できた。

特徴抽出に注力

今年のKDD Cupのデータセットで主要なテーブルはユーザの受講ログと教材マスタのみであり、さらに教材マスタにはその種類と階層関係しかなく、この限られた項目からいかに特徴を抽出するかが重要なポイントとなった^{☆3}。

人数は多いが作業場所と作業時間はまちまちである今回のチームでは、各人がData Veraci内で行われている仮説検証状況を確認しつつ、新たな仮説を立案し検証するという形で、特徴抽出を分担して行った。

最初にチームで共有した特徴量は、受講ログの項目定義から機械的に集計パターンを網羅した約500次元の“ベースライン”特徴量である（図-1）。各メンバはこの特徴量をもとに、異なる観点からの仮説を立て特徴量を追加していった。特に以下の特徴抽出が精度向上に大きく貢献した。

^{☆1} 新日鉄住金ソリューションズの登録商標。

^{☆2} 実務上、委託元からデータ格納場所を社内限定される場合が多く、機器を社内で保有管理するクラウドサービスとしている。

^{☆3} 今回のKDD Cupの課題については、本特集「編集にあたって」を参照。

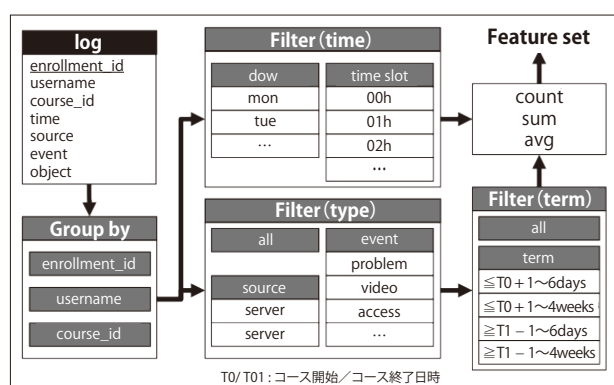


図-1 「ベースライン」特徴量

- **学習進捗**: ユーザごとの教材のアクセス順序を集計し、平均的なアクセス順序やタイミングを定義、進捗状況や平均からのズレを求める
- **未来ログ**^{☆4}: 今回は、講座に10日間アクセスがない状態を脱落と定義しているが、講座の開催期間に重複があるため、その10日間に別の講座を受けている場合があり、その回数を求める
- **周期性**: 受講ログの単純集計で失われるアクセスの周期性を、フーリエ変換により求める
- **アクセスパターン**: テキストを見る、ビデオを見る、問題を解くなどの受講ログの出現順序を状態遷移と見なし、その遷移確率や時間を求める
- **講座別モデル**: 講座単位で各教材のアクセスなどを特徴量化し講座単位で識別器を作成、その結果を特徴量として用いる

上記の取り組みの結果、最終的には約2,500次元からなるデータセットが構築できた。この特徴量に対してxgboost^{☆5}（勾配ブースティング木の一実装）単体で作成したモデルのプライバシースコアは0.9087であり、ほぼ特徴抽出のみで入賞できたことになる。

自発的な分担

今回は、チームメンバーに事前に役割を設定せず、メンバー間の調整で分担を決めていた。このような管理で十分であった理由としては、メンバーそれぞれが経験

ある実務者として、重要な課題を発見し解決する能力があったため、Data Veraci 内での議論やドキュメント、ソースコードの確認により、メンバーの対応余力や力量を理解し共有できたからである。

今回はメンバーの能力で解決したが、より組織的にデータマイニングを行うためには、スキルやプロセスを定義しその中で役割を決めていく必要があると考えている。

チーム戦でのモチベーション

多人数チームで参加する場合、個人で参加する場合と比べて結果に対する貢献度合いが不明確になる。加えて通常業務の合間に実行していることと、明確な役割がないことが影響して、モチベーションの維持が難しい。

今回はチームマージ期限までは個人で参加し、メンバー間で競争させる対策をとった。それでもチームマージ後のモチベーション維持が課題になるが、今回はチームマージの途中で1位になったため、順位の維持が明確な目的となり、モチベーションが維持しやすかったと考える。

また、他メンバーの活動を見て自分もやる気になる点や、グループ企業内とはいえ別組織との合同チームであることによる、取り組みへの責任感もモチベーション維持に寄与したと考える。

今後の課題と取り組み

今回は惜しくも2位であったが、参加によって共同作業の進め方など、業務にフィードバックできる多くの知見を得ることができた。今後も企業内のデータ分析実務者として、新手法のキャッチアップ、IT環境の整備、プロセスの洗練などを継続して行う中で、KDD Cup に代表されるオープンコンペティションに積極的にチャレンジしていきたいと考えている。

(2015年10月30日受付)

西川大亮 ■ nishikawa.daisuke.8tx@jp.nssol.nssmc.com

2001年北海道大学大学院工学研究科博士後期課程修了。博士（工学）。現在、新日鉄住金ソリューションズ（株）システム研究開発センターデータ分析グループリーダー。

☆4 実務上は予測時点では利用できないデータだが、競技であるため使用している。

☆5 <https://github.com/dmlc/xgboost>