会議報告

Knowledge Discovery and Data Mining（KDD 2019）

開 催 地：Dena'ina Convention Center および
William Egan Convention Center
（アンカレッジ，USA）
開催日程：2019年8月4日（日）～8日（木）
Web：http://www.kdd.org/kdd2019/

1. KDD 2019

KDDはデータマイニングに関する国際会議で，この分野では最難関会議と位置付けられている．ICMLやNeurIPSなどの機械学習の国際会議ではデータマイニングに必要なアルゴリズムや手法が中心である．それに加え，データマイニング分野の会議では具体的な実問題も対象とし，その定式化やモデル化についての発表もなされる．2008年までは北米のみだったが，それ以降は，北米で2回，その他で1回の周期となっている．筆者は12回目の参加で，2009年以降は続けて参加している．

開催地のアンカレッジは，アラスカ州最大の都市である．人口は30万人ほどで，本会議の開催によって人口が1%増加したそうである．そのため，いくつものホテルでオーバブッキングが生じ，多数の参加者が路頭に迷うというハプニングがあった．市の中心部のダウンタウンにある二つの会場は2ブロックほど離れていた．今年もスポンサーは50社以上あった．ダイヤモンドスポンサーは，ここ数年そうであった中国のライドシェアDiDiの代わりにBaiduと，去年もそうであった会計サービスのintuitであった．非IT系企業の割合は減っているようで，TARGET，BOSCH，Bloombergぐらいであった．今年は日本からも，SmartNewsやNAVER LINEがスポンサーになっていた．

5日間の開催期間のうち，初日がチュートリアル，2日目がワークショップ，そして本会議が3日間である．参加者数は，51か国から3 194人で，去年のロンドン3 411人よりは少ないものの，2016年サンフランシスコ2 792人や2017年カナダ・ハリファックスよりは多かった．中国からの参加者はますます増えていた．また，筆者の見た印象では，日本からの参加者は100人前後のようだった．

二つのトラックに分かれた一般発表のほかにも多様なセッションが組まれている．講義形式のチュートリアルに加え，ハンズオンチュートリアルも行われた．ヘルスケアと深層学習の招待講演を集めたhealth dayとdeep learning dayが去年は開催されたが，さらに自然科学を扱うearth dayも加わった．中国などのローカルチャプターのセッション，応用データ科学招待講演など，去年もあったセッションのほかに，社会問題を扱うSocial Impactワークショップが開催された．

2. 基 調 講 演

2件の基調講演と1件のパネルがあった．最初の基調講演は，Peter Leeによる医療部門でのMicrosoft Researchのデータ分析活用についてであった．技術的な詳細について説明はなく，遺伝子分析，カルテの作成支援，医療論文の知識化，医療画像処理などの話題についてであった．最後に患者の安全性は正解率指標だけでは保証できないことに触れ，Microsoft社の倫理委員会が紹介された．

二つ目の講演は説明可能性に配慮した予測モデルで著名なCynthia Rudinである．前半は簡潔かつ，予測精度の高いモデルの存在定理についてであった．仮説空間中で，データに対する損失関数がしきい値以下になる仮説集合をRashomon集合と呼び，全仮説空間に対するRashomon集合の比率をRashomon比とする．最初の定理は，このRashomon比が大きければ，複雑なモデルの仮説と同等の損失の仮説が存在するには，どれくらい単純なモデルでもよいかを示すものである．そして次の定理で，Rashomon集合が大きければ，簡潔なモデルで複雑なモデルと同等の予測精度が達成できることを示した．実験的にRashomon比を推定する手法を開発し，さらにRashomon比と経験損失をプロットしたRashomon曲線を使い，適切な仮説集合を選ぶ方法について論じた．後半は，制約付き最適化を用いるが，効率的に解ける簡潔で説明可能性の高い学習手法を紹介した．

さらに，応用データ科学基調講演では，データ分析の実務家が24人招待された．深層学習で著名な，Apple社のRuslan Salakhutdinovは背景知識として，明示的に値を与えた記憶ユニットをRNNに加えることで，背景知識をニューラルネットワークで利用する方法を示したが，まだ効率的とはいえないようだった．強化学習で周囲を見渡しながら地図を作製するのに，長期と短期の二つのエージェントを利用する方法についても述べていた．

Airbnbの Elena Grewalは，2012年の創始期から，各国に拠点があるほど拡大するまでデータ科学チームを率いた経験について語った．初期的な結果は，分析チームに対する信頼を勝ち取るのに重要であるとか，データサイエンス職といっても，評価指標を定義するなどの分析職，アルゴリズムを作製する職，因果関係の調査をする職など，その内容はかなり異なるとのことであった．

3. チュートリアル・ワークショップ

講義形式のチュートリアルは 29 件で去年と同じだったが，本会議と並列して行われるハンズオンチュートリアルは 8 件から 15 件へと増えた．ここではファクトチェックに関するチュートリアルを少し紹介する．フェイクニュースを，意図的であること，信憑性（authenticity）を確認できないこと，そしてニュースであることの 3 要素で定義し，噂や誤解との差を明確にしていた．フェイクニュースの検出手法としては，信頼できる知識グラフを用いる知識ベース，文章や内容の特徴に基づくスタイルベース，ネットワーク上の拡散の経路に基づく方法，書き手の信頼性に基づく方法を紹介していた．

今年も 34 件のワークショップが開催された．対話システムの評価についてのワークショップで，ランキング学習で著名な Thorsten Joachims の講演を聴講した．オフラインデータでオンラインの効果を予測するには，利用者に提示したリストで上位のものほど閲覧されやすいバイアスを補正する必要がある．これを補正するために，閲覧される確率をランダム化実験で推定し，この確率を使って逆傾向スコアを求めて補正するランキング学習の手法を紹介していた．

4. 一般発表・受賞

KDD には，手法・理論・モデルなどの提案や改良を対象とした研究（Research）と，手法などを実問題に適用した事例を対象とした応用データ科学（Applied Data Science）の二つのトラックがある．投稿数は，研究トラックは 1 179 件，応用データ科学トラックでは 670 件で，いずれも記録を更新した．研究トラックでの採録数は，口頭 111 件（9.4%）ポスター 63 件（5.3%），応用データ科学トラックでは口頭 47 件（7%）ポスター 100 件（15%）であった．採択率の推移は，研究トラックは 17.4%→ 18.4%→ 14.8%と大きく低下したが，応用データ科学トラックは 22.1%→ 22.6%→ 22%と安定している．研究トラックには 54 か国から投稿があり，その 70%はアカデミアのものであった．ワードクラウドでは依然として深層学習は強いが，解釈可能性，ネットワーク埋込み，メタ学習などが目新しいように思う．

受賞についてまとめておく．研究トラックのベストペーパは“Network Density of States”，Runner-up は“Optimizing Impression Counts for Outdoor Advertising”であった．応用データ科学のベストペーパは“Actions Speak Louder Than Goals：Valuing Player Actions in Soccer”で，Runner-up は“Developing Measures of Cognitive Impairment in the Real World from Consumer Grade Multimodal Sensor Streams”であった．

引用数の多かった 10 年前の論文に与えられる Test-of-Time 賞はネットワーク分析の雄 Jure Leskovec らによる“Cost-effective Outbreak Detection in Networks”，会議運営への貢献により与えられるサービス賞は Balaji Krishnapuram，今までの業績に対して与えられる Innovation 賞は非常に多様な業績のある Charu Aggarwal であった．データ分析コンペティションの先駆けである KDD Cup は，通常の人間がモデルをつくる通常トラックと自動機械学習を行うトラックがつくられた．NTT ドコモの落合桂一らが通常トラックの一つで 1 位となったが，他は中国勢が占めた．

個人的に関心のあった一般発表をいくつかあげておく．

- Paper Matching with Local Fairness Constraints：論文と査読者の適正の総和を最大化すると，非常に悪い割当てが生じることがあるので，最低限の適正を保証したうえで割当てを決定する．
- Revisiting kd-tree for Nearest Neighbor Search：kd-tree は高次元では弱いので，軸に平行でない分割線を効率的に利用するために，データをランダム回転させる．
- Environment Reconstruction with Hidden Confounders for Reinforcement Learning based Recommendation：強化学習による推薦では，行動が別サイトの価格などの交絡因子の影響を受けるので，利用者や推薦器に加えて交絡因子のエージェントをも加える．
- Combining Decision Trees and Neural Networks for Learning-to-Rank in Personal Search：クエリの困難さに応じて，高精度の NN と，高速な GBDT のランク器を使い分けて速度と精度を両立させる．
- Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search：望ましい分布への近さと，最低表示回数を保証する公平性規準でのランク器の提案．

5. お わ り に

発表論文は無償で公開されており，短い紹介ビデオとともに SIGKDD のサイトに掲載されている．会議関連の Twitter の tweet は https://togetter.com/li/1382881 にまとめておいたので参考にされたい．2020 年は，アメリカ，サンディエゴにて 8 月 22 ～ 27 日，2011 年以来の 3 度目の開催となる．2021 年は，シンガポールとの発表があった．今年も参加者数は記録を更新しており，また日本からの参加者も増えているようで，来年も盛り上がるであろう．

〔神嶌　敏弘（産業技術総合研究所）〕