会議報告

The 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining（KDD 2018）

開 催 地：ExCeL London（ロンドン，イギリス）
開催日程：2018 年 8 月 19 日（日）～ 23 日（木）
http://www.kdd.org/kdd2018/

1. KDD 2018

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 NIPS 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 11 回目の参加で，2009 年以降は続けて参加している．

開催地のロンドンは，イギリスの首都である．ここは，アルファベット社傘下の DeepMind 社や，UCL をはじめ多数の大学が存在し，データ科学の集積地の一つといってよいだろう．会場はロンドン塔よりさらに東にはずれた，ロンドン・シティ空港そばであった．オリンピック会場を展示場に改修した巨大な建物であった．今年も IT 系企業を中心としたスポンサーが 50 社以上あり，提供金額は約 120 万 USD であった．ダイヤモンドスポンサーは，ここ数年そうである中国のライドシェアリング DiDi に加え，小規模会計サービスの intuit であった．昨年に続いて日本企業のスポンサーはなかったようだ．

2 年前から 5 日間の開催となり，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，99 か国から 3 411 人で，これは 2016 年サンフランシスコの 2 792 人や昨年カナダ・ハリファックスの 1 675 人を超えた新記録であった．中国からの参加者が非常に多かった．筆者の見た印象では，日本からの参加者は 70 ～ 100 人ぐらいだと思う．

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．通常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．各種インドや中国のチャプターのセッション，ヘルスケアと深層学習の招待講演を集めた helth day と deep learning day，若手研究者が著名研究者と対話できる Networking with Experts Forum，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，AI の社会に与える影響などを扱う Global AI Initiatives セッションなどもあった．

2. 基調講演

4 件の基調講演があった．最初の Jeannette M.Wing は，Data for Good と題し，医療，気候変動，安全性評価などへのデータ科学，特に因果推論の利用についての講演であった．

二つ目は，著名な統計学者 David Hand による講演で，金融がテーマであった．金融分野は最も統計学の適用が早かったが，長い間経済学などのモデルに基づく理論駆動が主流であった．それに対し，最近は，データへの当てはめによるデータ駆動に変わり精度が向上した．しかし一方で，過学習などの危険を“Let the data speak for themselves”という著名な言葉をもじり，“If data can speak for themselves, they can also lie for themselves”といかにもイギリス的に戒めていた．また，衛星写真の石油タンクの影から備蓄量を予測できるなど多くの関連情報を扱えるようになったことも最近の傾向である．このように統計的予測に関する深い洞察に基づいている一方で，最後に巨大スクリーンがニョキニョキと長くなる指し棒を取り出して指し示すなど，ユーモアにもあふれていた．

三つ目は，ノーベル経済学賞受賞者である Alvin Roth による，求職者と雇用者などを結び付けるマッチング市場についてであった．学校への入学希望者を割り当てる問題で，希望者の希望順位ごとに選考をすると，第 2 希望がすでに埋まっているため，本当の希望順を申告しないことが有利になる場合がある．そこで，deferred acceptance では，仮割当ての概念を導入することで，この問題を解決するアルゴリズムである．移植する腎臓の割当てなどにも適用され，その有効性が示されている．こうしたマッチング市場のためのアルゴリズムは，シェアリング，就職，婚活などのインターネット上のサイトで，より複雑な条件を満たすように改良され，その重要性は高まるだろうとのことであった．

四つ目は，ノンパラメトリックベイズで著名な Yee Whye Teh が，データが少ない状況に対応するための超パラメータ推定とマルチタスク学習について講演した．超パラメータの最適化には，ガウス過程がよく用いられていたが，計算量が多い問題があった．そこで，確定的な関数を生成する分布を導入するニューラル過程を紹介していた．次に，Distral と呼ぶマルチタスク学習手法は，タスク間の共通モデルを正則化項に用いて事前知識とし，その一方で，タスク個別モデルに distillation 手法を適用して，共通モデルを獲得するものであった．

そのほか，2 年前から始まった，Applied Data Science トラックでは，データ分析の実務家の招待講演が企画され，12 件の講演があったが，筆者自身は聴講しなかっ

たので割愛する．

3. チュートリアル・ワークショップ

通常形式のチュートリアル29件に加え，本会議と並列して8件のハンズオンチュートリアルがあった．筆者は，公平性とファクトチェッキングに関するチュートリアルを聴講した．公平性は因果推論に基づいたもので，介入による効果の公平性を論じていた．ファクトチェッキングは，フェイクニュースへの対抗策として注目されており，外部情報源からエビデンスを収集し，その信頼性を評価して情報を集約することで実現が目指されている．現状の知識で，多くの推論過程を通じて同じ言明が導出されるなら，その言明はより信頼できるとか，異常検出技術を活用した方法などで信頼性評価を行っていた．

去年と同様に，今年もワークショップに1日が割り当てられ，27件のワークショップが開催された．筆者はデータジャーナリズムに関するワークショップを聴講した．ドイツのニュースサイトZEITは，印刷媒体の新聞社がその記事をオンラインで配信するのではなく，親会社の新聞社とはオンラインサイトは運営レベルで独立している．社会問題について意見が異なるが，環境などが類似している人をマッチングさせて，実際に会って議論を行うGermany Talksの紹介があった．ファクトチェッキングを目指すFACTMATA社の取組みの紹介では，LIMEなど機械学習のモデル説明をするアルゴリズムと，ジャーナリストの事実に基づくかどうかという基準との整合性についての定量評価の試みなどがあった．

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983件，応用データ科学トラックでは496で，いずれも過去最多であった．研究トラックでの採録数は，口頭107（10.9%），ポスター74（7.5%），応用データ科学トラックでは口頭40（8%），ポスター72（14.5%）であった．採択率の推移は，研究トラックは18.1%→17.4%→18.4%と低下傾向に歯止めがかかり，応用データ科学トラック3年目になって，19.9%→22.1%→22.6%と安定状態になった．分野別では，研究トラックは深層学習や転移学習の注目が高まっているようで，応用データ科学トラックでは自然科学への適用が伸びているようだ．

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“Adversarial Attacks on Classification Models for Graphs”，機械学習の安全性に関するもの，Runner-upは深層学習系で“Xiaolce Band：A Melody and Arrangement Generation Framework for Pop Music”であった．応用データ科学のベストペーパは“Real-time Personalization using Embeddings for Search Ranking at Airbnb”で，Runner-upは昨年も関連発表があった水道水汚染の分析“ActiveRemediation：The Search for Lead Pipes in Flint，Michigan”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Yehuda Korenの近隣法と行列分解を用いた協調フィルタリングアルゴリズムの論文，会議運営への貢献により与えられるサービス賞はJie Tang，今までの業績に対して与えられるInnovation賞はlifelong学習などで著名なBing Liu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，ロンドンと北京の大気汚染予測の問題で，中国勢が独占した．その他ベンチャー企業を対象としたStartup Award，社会への影響を評価するSocial Impact Awardなどがあった．

個人的に関心のあった一般発表をいくつかあげておく．

- Real-time Personalization using Embeddings for Search Ranking at Airbnb：利用者は類似した物件を順番に閲覧することから，閲覧系列にskip-gram系列を適用することで，物件の類似性を反映した潜在空間への埋込み表現を獲得できるというものであった．
- Q&R：A Two-Stage Approach toward Interactive Recommendation：「○○に関心がありますか」といった説明を能動的に生成し，その応答に応じて推薦内容を変える．
- Sequences of Sets：一連のメールの宛名の集合といった集合の系列から，次に現れる集合を予測する予測器．
- A Unified Approach to Quantifying Algorithmic Unfairness：Measuring Individual & Group Unfairness via Inequality Indices：予測結果と実際の施策との差によって生じる利益に基づいて公平性を定義する．

5. おわりに

会議関連のTwitterのtweetは`https://togetter.com/li/1257525`にまとめておいたので参考にされたい．2019年は，アメリカ，アラスカ州のアンカレッジにて8月3～7日に開催される．2020年は，アメリカ，サンディエゴで，2011年以来の開催となる．今年も参加者数は記録を更新しており，あらゆる分野の基盤技術としてその需要は高まり続けているので，今後も本会議は発展していくであろう．

〔神嶌 敏弘（産業技術総合研究所）〕