会議報告

The 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining（KDD 2016）

開催地：Hilton San Francisco Union Square
（サンフランシスコ，米国）
開催日程：2016 年 8 月 13 日（土）～ 17 日（水）
http://www.kdd.org/kdd2016/

1. KDD 2016

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，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．ほとんどは北米で開催され，その他の地域で開催されたのはパリ（2009），北京（2012），シドニー（2015）である．筆者は 9 回目の参加で，2009 年以降は続けて参加している．

開催地のサンフランシスコは，シリコンバレーにも近く世界の IT 産業の中心地といってよいであろう．KDD は 2001 年の第 7 回に一度行われており，今回が 2 回目である．会場のホテルは，ダウンタウンにあり，中心地の Union Square から 2 ブロックほど離れている．今年も多数の IT 系企業がスポンサーになっており，筆者の知る限り日本企業では日本電気が初めてスポンサーとなっていた．提供総額も，74.3 万 USD と去年から 3 倍ほどになっている．

例年は 4 日間の開催だが，今年は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 日間と開催期間が 5 日であった．参加者数は，2014 年のニューヨーク 2 134 人や 2015 年のシドニー 1 119 人を上回り最終日には 2 792 人と大幅に増加した．ACM の中でも SIGGRAPH などの上位 5 位の巨大会議に近づいているとのことで，データマイニングへの注目がうかがえる．筆者の見た印象では，日本からは数十人が参加していたようだ．

今回は，産業寄りのさまざまな新企画があった．まず，今までは理論面を扱う「研究」と，応用面を扱う「産業・政府」との二つのトラックがあったが，後者が「応用データ科学」という名称に変わった．通常のチュートリアルに加え，技術会議でよく行われている，Spark や深層学習ライブラリなどのソフトウェアを実際に使ってみるハンズオンチュートリアルが 8 件行われた．カンファレンスの重要な目的の一つは新たな人の交流をつくり出すことであり，そのための新企画があった．SIGKDD にはインド，中国，豪＋ NZ の国別チャプターと，オースチンやシアトルの地方チャプターがあるが，中国とインドのチャプターでの交流会があった．従来からのジョブマッチングに加え，ベンチャーキャピタリストとの会合もあった．

2. 招 待 講 演

基調講演は，通常のもの 3 件に加え，昼食中の講演とチューリング賞受賞者の講演があった．最初の基調講演は数学者 Jennifer Chayes によるグラフ上の確率についてであった．ソーシャルネットワークなどは時間とともに変化するが，その変化が収束した極限での状態について分析するため，グラフの系列の収束性と極限を測度論に基づいて議論する．このようなグラフの収束性についてはいろいろな研究があり，数十種類もの収束性が提案されているとのことであった．

二つ目はシリコンバレーらしく，ベンチャーキャピタリスト Greg Papadopoulos の講演であり，3 部構成の内容であった．最初は，データ科学の現状についてであり，画像処理などに非常に大きな進展があったことが紹介された．しかし，その大きな進展の多くは，新たな入力を得たことによっていると見ているとのことであった．そして，データ科学は確率的な系の記述を与えるが，それはその系自体が真に確率的な挙動をしていることを意味していないと認識していた．慧眼である．サン・マイクロシステムズの CTO を務めた経歴もあるように，技術を見る目は素晴らしい．Google に最初に資金を提供した Andy Bechtolsheim が，PageRank についてのプレゼンの途中でその可能性を認め，プレゼンを遮り資金の提供を決めた逸話が思い出された．このように，技術に関して高度な知見と洞察力を備えた投資家がいることこそがシリコンバレーの強みなのだと感じた．続いてベンチャーキャピタルの投資傾向などについて簡単な紹介があり，今年は全体の投資額は大きく減少しているが，人工知能関連は若干上昇しているようである．最後は，基盤技術はコモディティ化することを前提に，特定の事業と特定の技術を結び付けることと規模の拡大ができるようにすることが重要であると締めくくった．

同じ日に，公開鍵暗号で，チューリング賞を 2015 年に授与された Whitfield Diffie の講演があった．内容は，第二次世界大戦以降の情報セキュリティの歴史についてであった．全くスライドを使わず，またほとんどメモを見ることもなく 1 時間以上にわたってよどみなく講演するのに驚かされた．

三つ目は，やはり外すことのできない深層学習について Nando de Freitas が講演した．AlphaGO で脚光を浴

びたDeepMind社の中心人物の一人である．前半は，深層学習でよく知られた成果の俯瞰で，一般画像認識，自然言語処理，sequence 2 sequence，GAN，深層Q学習などの紹介であった．後半は最新の内容について取り上げた．部分問題について学習した部品を組み立てて，より複雑な問題を解く枠組みを用いることで，少数の訓練データで学習が可能になる手法や，オンライン学習手法の最新手法の紹介などがあった．

今年から，Applied Data Scienceという企業からの招待講演が企画され，12件の講演があった．そのうちスモールワールドの分析で著名なDuncan Wattsは二つの話題について講演した．従来の社会学とは異なり，計算機を利用することで能動的な実験もできる計算的社会学が始まったとの前置きがあった．一つ目の話題は，伝染病の感染モデルで，クチコミの伝播をモデル化することが多いが，統計的な性質を調べてみるとかなり異なる．感染は同時に数名程度に伝播する程度だが，クチコミでは少数が大量に伝播する人がいるとのことであった．後半は作業チームの規模についてで，人数はある程度増やしても，チームの効率はあまり上がらなくなる現象を調査した．クラウドソーシングを利用し，地図上に災害や犯罪の発生を入力する作業を使い実験したところ，こうした限界があることが確かめられた．

3. チュートリアル・ワークショップ

例年と同じ形式のチュートリアル10件に加え，本会議と並列して前述のハンズオンチュートリアルがあった．筆者が聴講したのは，現在取り組んでいる公正性に関するものである．ローンの可否や採用にデータに基づく分類器などが利用されているが，訓練データに性別や人種に基づく差別があると，その分類器も差別的になってしまう．こうしたデータの差別的な決定を抽出したり，差別的判断を補正した分類器を獲得するための技術である．もう一つ聴講したのはLifelong学習というもので，機械学習で新たなタスクを解くときに，過去の経験を利用して効率的に解くものである．これは過去から続くタスクからの転移学習ともみなせるが，過去のタスクの中から関連するものを特定する必要がある．NELL（Never-Ending Language Learner）というプロジェクトでは，2011年からずっと継続的にWebから知識を獲得し続けており，9 000万以上の信念記述を獲得している．

例年はチュートリアルと同日開催のワークショップは，今年は次の日に17件開催された．因果発見といった基盤術をテーマにしたものや，医療，スポーツ，ファッションなど分析対象ごとのものがあった．スポーツに関するワークショップでは，イギリスの著名なサッカーチームであるアーセナルが，データ分析会社を傘下にしており，データに基づく科学的な戦略がスポーツにまで浸透しているのは知らなかった．

その他，いくつかのパネルも企画され，「深層学習は過大評価されてはいないか？」というテーマのパネルは非常に混雑したとのことだった．

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（Resarch）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．研究トラックでは，2014年の1 036件，2015年の819件に対し，今年は784件と減少したのに対し，応用データ科学トラックでは，2014年の197件と，2015年の189件に対し，今年は331件と大幅に増えた．アカデミアと企業という分類から，分析手法と分析応用例という分類に変わったことを反映しているのだろう．研究トラックは口頭発表70件とポスター72件，応用データ科学トラックは66件の採録があった．採択率の推移は，研究トラックは14.6％→19.5％→18.1％，応用トラックは22％→36％→19.9％とおおむね20％弱となっている．分野別の傾向では，ソーシャルネットワーク関連は例年どおり多く，位置情報を扱うものが増えた印象がある．深層学習そのものの研究はほぼ皆無だが，画像やテキスト情報の抽出器として深層学習はよく用いられるようになったようだ．

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商品評価の従来のニセ評価検出を回避する機能をもったニセ評価検出器，学生ベストペーパはグラフ中の三角形の数という基本的な統計量をストリームデータから得る方法であった．応用データ科学トラックのベストペーパは，Yahoo! 検索でのランキング学習など複数の方法を組み合わせた検索結果のランキング手法について，学生ベストペーパは利用者の行動を予測し，モバイルの適切な情報を表示させる個人化アシスタントであった．会議運営への貢献により与えられるサービス賞はWei Wang，今までの業績に対して与えられるInnovation賞はクラスタリングやデータストリームで多くの業績のあるPhilip S. Yuに贈られた．データ分析コンペティションの先駆けであるKDD Cupは，今年度のKDDなどの国際会議に採録される機関の順位を予測するもので，従来のコンペティションのようにテストデータでの予測結果が得られない難しさがあった．

個人的に関心のあった一般発表をいくつかあげておく．

- “Why Should I Trust You?” Explaining the Predictions of Any Classifier：学習した予測器の説明を得るため，説明したい分類器を局所的に単純なルールで近似することで解釈できるようにする．
- Towards Conversational Recommender Systems：推薦の状況を尋ねたり，選択肢を用意したりして能動的な問いかけを行う推薦システム．

- Assessing Human Error Against a Benchmark of Perfection：チェスの終盤で完全に解析されている手順と，人間のプロの手を比較し，熟練者がミスをしやすい条件を分析.
- The Limits of Popularity-Based Recommendations, and the Role of Social Ties：友人間の推薦により全体の購買動向がどう変化するかを，極限の状態で理論的に分析.
- Dope Learning：A Computational Approach to Rap Lyrics Generation：ラップの歌詞を自動生成する．自然な歌詞にするためクラウドソーシングも利用.

5. お わ り に

今回から発表論文は無償公開となり，短い紹介ビデオとともに SIGKDD のサイトに掲載されている．招待講演やパネルのビデオは YouTube で「KDD 2016 video」のチャネルを検索すると閲覧できる．会議関連の Twitter の Tweet は http://togetter.com/li/1004483 にまとめておいたので参考にされたい.

2017 年は，カナダ東海岸のハリファックス市にて，開催機関が 4 日間に戻って 8 月 14 ～ 17 日に開催される．カナダでの開催は，1995 年の第 1 回モントリオール，2002 年の第 8 回エドモントンに続き 3 回目となる．また，2018 年はイギリスのロンドンでの開催とのアナウンスもあった．オープニングでは，データ分析の産業化は，データサイエンティストからドメイン専門家に担い手が拡大し，電子カルテなどに分析技術が取り込まれるようになる次の段階に移るだろうという見方が紹介された．このように，データマイニング・機械学習技術への注目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）〕