

20th International Conference on Machine Learning (ICML 2003) 8/21 ~ 8/24 および 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003) 8/24 ~ 8/27 · ワシントン DC

今回の ICML と KDD の両会議は、「我々は今後どのような方向に進むべきか？」を模索する会議であったと思う。むろん ICML (採択率 33%)、KDD (採択率 15%) とともに高い評価が確立した会議であり、発表はさすがに粒ぞろいである。ただ、個々の研究テーマを着実に発展させた論文が多い中で、この分野を一気に活性化するような新しいテーマは見いだしにくかった。

参加者数で見れば、今回の会議は成功と言えるだろう。ICML と KDD (と COLT) の同地開催が功を奏し、ICML の登録者約 400 人、KDD の登録者約 780 人のうち、約 120 人が両方の会議に参加した (ちなみに ICML と COLT の共通参加者は約 70 人)。以下では各会議の内容を手短に紹介する。

まず ICML のプログラム (全 34 セッション) は、その多くが統計的学習法に関する研究であるが、特に Reinforcement Learning をタイトルに含むセッションが 5 個もあるのが目立つ。また SVM を含むカーネル法に関するセッションも 3 個あり、これらのテーマが ICML コミュニティの関心といえるだろう。また、多少新しい話題として扱われていたのが、関係データからの学習を扱う relational learning で、これは ICML/KDD 共同招待講演の “Statistical Learning from Relational Data” (Daphne Koller) でも取り上げられた。そのほか、Dimension Reduction や Feature Selection などの伝統的なテーマにも、新しい切り口に基づく佳品があった。

論文の内容は理論やアルゴリズムを主題とするものが多く、アプリケーションに関する論文は少なかった。この点は ICML コミュニティにも懸念があるようで、招待講演でも “The Role of Applications in the Science of Machine Learning” (Foster Provost) において「アプリケーションは機械学習の研究を進展させる原動力になる」という呼び掛けがあった。

一方 KDD のプログラム (全 13 セッション) は、アルゴリズムとアプリケーションが適度に混ざった構成である。例えば、「飛行機のチケットをいかに安く買うか」や「飛行機のチケットをいかに高く売るか」など、すぐに使えそうな研究もある。またアプリケーションと同様にアルゴリズムも多彩であるが、やや成熟化の気配も感じられ、新しい方向性は見えにくかった。

むしろデータマイニングに対する危機意識が語られた会議でもあった。そもそものきっかけは、米国国防総省の Total Information Awareness プログラム、すなわちテロ対策を目的とした大規模なデータ収集プログラムにある。これが「データマイニング計画」と呼ばれて政治問題化したために、データマイニングはスパイ技術の一種であり、人々のプライバシーを侵害する危険な技術なのではないか、という論争が高まった。このような動きに対して KDD コミュニティの研究をアピールするため、開催地がワシントンであることを活用し、政府関係者や NSF 担当者を招いたパネルが開かれた。このプライバシーの問題は、研究・ビジネスの双方において、今後無視できない問題となりそうである。

両会議の招待講演のうち、筆者にとって興味深かったのは以下の二つの講演である。最初は、ペンシルバニア大学の Michael Kearns による “Computational Learning Theory: A Reptrospective” (ICML) である。この講演は、計算論的学習理論に関する研究の歴史を、特に本人がその初期から関係してきた boosting の発展の歴史を軸に回想するものである。完成した理論は美しいが、その背景には長い研究の歴史や生みの苦しみがあることを、具体的な登場人物を交えながら楽しそうに語る姿が印象的だった。

もう一つは amazon.com の Andreas Weigend による “Analyzing Customer Behavior at Amazon.com” (KDD) である。これは、amazon.com の chief scientist であり KDD にもかつては参加したことがある講演者が、amazon.com が構築する巨大なデータベースと機械学習との関係とをわかりやすく紹介する講演であった。講演では、まずデータベースに見出せるさまざまなベキ乗則を述べたあと、amazon.com における顧客の行動分析の話に移った。この講演で痛感したことは、amazon.com が今や機械学習の壮大な「実験場」になっているということ。なぜなら、アルゴリズムを実装してウェブで公開すれば、膨大な顧客からの反応を即座に収集できるからである。amazon.com ではこのような実際の顧客データを研究にも活用できるよう、Amazon Cup と題する標準データセットを研究者コミュニティに公開する予定だそうである。公開後は Andreas Weigend 自身のウェブページからリンクが張られるとのことだが、データの質量ともに他を圧倒する実世界データセットとなる可能性があり、要注目である。

[北本 朝展 (国立情報学研究所)]