解説

データ解析コンペティションでの取組みと成果 ー運営のための組織体制作り及び産学連携ー[†]

中川 慶一郎^{*1}・生田目 崇^{*2}

1. データ解析コンペティションについて

本稿で紹介する「データ解析コンペティション」(以下、本コンペティション)は、平成6年(1994年)に日本オペレーションズ・リサーチ学会、マーケティング・サイエンス研究部会(主査：木島正明(現・首都大学東京教授)、幹事：守口剛(現・早稲田大学教授))によってはじめられた、マーケティングおよび経営科学に関するデータ分析を競うことを目的にしたプロジェクトである。

本コンペティションの目的は、「コンペティションを通じて、データ解析及びモデル設計を行うことで、新たな知見を誘発し、学術的及び実務的な研究の発展に寄与すること」である。基本コンセプトは、産学から幅広く参加チームを募集して研究のすそ野を広げることと、共通のデータを利用することで発表者、聴衆両者が分析データの情報を共有し、かつお互いの分析の視点を比較できる点にある。コンペティションでは、データ分析の新規性、ビジネス有効性などを軸に審査をおこなっている。

そのために、本コンペティションの活動をご理解いただいている企業から可能な範囲でデータを提供いただき、参加チームに貸与している。特に近年では、サービスデータを提供いただく機会も増えており、本コンペティションを内外から高く評価いただいている。

1.1 体制づくり

本コンペティションは当初、日本オペレーションズ・リサーチ学会のマーケティング・サイエンス研究部会の活動として始まったが、本活動に賛同する他研究部会も参加するようになり、徐々に規模が拡大してきた。これまでの活動は大きく3つの期間に分けるこ

とができる。

まず、第一期は平成6～10年度の黎明期である。本コンペティションが初めて開催され、徐々に学会内でその活動が認知されてきた。ただし、利用するデータは個人的な伝手を頼ってきた。またこの期間は、POSデータのような物販データが中心であった。

第二期は、平成11～15年度の拡大期である。研究部会の開催に加え、データ概要の説明や募集方法を説明する「発会式」と、研究成果全体を総括する「成果報告会」を新たに設けた。10周年に至るまで、これらの開催については(株)NTTデータが全面的に受け持った。また、この頃から関東以外の参加者も増えてきた。さらに、提供データについてもサービス・データをご提供いただけるようになり、いくつかの共催団体を得るようになった。

第三期は、平成16年度以降であり、いわば広域展開期である。全体を取りまとめる団体の必要性から主催する協議会(経営科学系研究部会連合協議会：座長守口剛・早稲田大学教授)を設立したのははじめ、関東以外での開催、学生部門の新設、分析ツールの貸与など現在の体制がほぼ整った。また、データ提供についても、主催者以外から様々な情報をいただけるようになり、その範囲は格段に広がった。協議会構成団体が現在では以下の6学術団体、2企業からなる。

- 日本オペレーションズ・リサーチ学会マーケティング情報研究部会(平成23年度より先端マーケティング研究部会)
- 日本マーケティング・サイエンス学会ID付POSデータ活用部会
- 日本データベース学会ビジネスインテリジェンス研究グループ
- 早稲田大学マーケティング・コミュニケーション研究所CRM研究会
- 日本マーケティング・サイエンス学会市場予測のための消費者行動分析研究部会
- 日本計算機統計学会データ解析スタディグループ
- (株)NTTデータ技術開発本部
- (株)産業科学研究開発センター

[†] Activity and Outcome of Data Analysis Competition - Establishment of Organization and Joint Industry- University Research Project -

Keiichiro NAKAGAWA and Takashi NAMATAME

^{*1} (株)NTTデータ技術開発本部

R&D Headquarters, NTT DATA Corp.

^{*2} 専修大学商学部

School of Commerce, Senshu University

このように、様々な専門を持つ団体により構成されている。これは、マーケティング・データ解析が学際的な分野であることと大きく関係しており、通常は交流のない専門の異なる参加者が集まって活動できる点が特徴である。

一般的なスケジュールとしては、8月初旬に発会式を開催して参加チームを募集し、参加チームにデータを貸与する。10月から翌年3月までに各研究部会でそれぞれのチームが2回(中間発表と最終発表)を行う。そして、各研究部会から選抜されたチームにより、3月下旬に成果報告会を開催し、最終的な審査を行うというものである。

平成22年度は55チーム、250名の参加を得た。ただし、平成23年3月に起きた東日本大震災の影響のため、成果報告会は延期することになり、最終審査は今後開催する予定である。

2. コンペティションの変遷

このように多数の参加を得るようになった本コンペティションであるが、最初の年度はわずか5チームの開催であった。前述のとおり、年を重ねるに従い参加

者が増加し、規模も拡大してきた。

また、この間提供データの質・量両面での変化もある。まず、データ量の飛躍的な増加がある。初期はフロッピーディスク1枚で十分に足りる量のデータであったが、もっとも規模の大きいデータでは、テキストファイルで約2Gbyteと数千倍のデータ量となった。また、質的にも対象は当初小売業が中心であったが、その後範囲を広げ、金融や食卓、インターネットなど、様々な業界のデータを利用してきた。特筆すべきは、当初は提供データは基幹系業務データが多かったが、サービスに関連するデータを提供いただけるようになったことである。また、データを提供いただいた企業も関東近辺から全国に広がり、地方の小売店からデータ提供を受けたこともある。地方の企業でも全国各地の研究者の知識や技術を活用できる場になっていると考える。これまでの提供データの変遷を表1にまとめる。

チーム数の増加に合わせて学生チームの数も増加した。しかし同時に、データ分析を始めたばかりの学生と、第一線の研究者や企業人のと間でデータ分析のスキルの差も顕著になってきた。そこで、学生部門と一

表1 各年度の概要

開催年度 (平成)	テーマ	提供データ
6	食品購買行動	ストア・スキャン・データ
7	食品・日雑品購買行動	ストア・スキャン・データ
8	日雑品購買行動	ホーム・スキャンデータ
9	旅行行動分析	アンケートによる旅行履歴、意識調査
10	食卓消費マーケティング	食卓メニューデータ
11	金融マーケティング	意思決定、意識アンケート
12	金融マーケティング	金融機関取引サマリ、アンケート
13	流通 CRM	食品 SM ID 付 POS データ
14	流通 CRM	百貨店 ID 付 POS データ
15	10 周年記念 (3 種類提供)	世帯電力消費、属性・意識調査データ ハウスカード スーパー・ドラッグストア POS データ
16	2 種類提供	クレジットカード利用履歴 加工食品 POS
17	アミューズメント・マーケティング	CD 販売店 ID 付 POS データ
18	Web マーケティング	ウェブサイト・アクセス・ログ・データ
19	オークション・データ分析	自動車オークション・データ
20	食卓消費マーケティング	食卓メニューデータ
21	大規模小売店購買行動分析	ファッション大型店 ID 付 POS データ
22	ドラッグストア購買行動分析	複数チェーン・店舗 ID 付 POS データ

般部門を分けることによって、学生は学生同士、一般部門は研究者や社会人と、同じ立場のグループの中でコンペティションを進めた。

さらに、参加者の地理分布にも変化がでてきた。当初、東京だけで開催していた本コンペティションも、回を重ねるにつれて関東地方以外からの参加者も多くなってきた。そこで、関西の研究部会に関西での開催をお願いした。これにより、特に、関西の学生の参加が容易になった。また、計算機統計学会では研究大会中に発表会を開催することで、全国大会への参加に合わせて発表することができる。

これまでに、北は北海道から南は九州まで、全国様々な場所からの参加を得ている。

3. 参加に付随して

3.1 参加者の義務

本コンペティション参加について、参加チームには下記について強くお願いをしている。

誓約書の提出、事前審査 データ提供企業にとっては社外にデータを出すことになるため、データがどういった人に渡され、どのように管理をしているかは大変気になることとなる。また、得られた知見などの知的財産権の帰属や機密保持といったコンプライアンス面での統一的な扱いについても必要となるため、本コンペティションでは、コンペティション参加にあたって、申込時にチームメンバー全員の署名・捺印された所定の誓約書を提出してもらっている。そして、申込のあったメンバー全員について、すべて事前にデータ提供企業に参加の可否について確認をいただいている。データ提供企業と直接の競合関係にある企業からの参加などについては残念ながら参加(データの提供)をお断りするケースもある。

データの管理 データは貸与の形でチームに渡され、コンペティション終了後に回収し事務局で適切に処分する。各チームにはメンバーのみがデータに接触でき、参加メンバー以外が触れないように管理するようにお願いしている。学生の参加も多いため、USBメモリなどの可搬媒体にデータを保存することも想定されるが、暗号化するなどデータ管理の徹底を依頼しており、他人が内容を確認できないようにと周知している。もちろん、法律が規定する個人情報とはもとより提供されていないが、データ管理については情報リテラシの醸成も含めて参加者に義務付けている。

発表の義務化 発表者にはかならず本コンペティションを開催する学術団体のいずれかで研究成果発表をしなければならない。これは、データ提供企業に対しての敬意も含めて、参加者の義務としている。たとえ、

当初想定していたほど分析のための時間が取れないような場合であっても、発表日までの成果を報告してもらっている。

これらは、企業が安心してデータを提供いただけるよう、また、コンペティションを毎年継続して開催するための、協議会としての約束事であり、今後も厳格に運用する。

3.2 分析ツールの貸与

本コンペティションはまさにPC環境の進展とともに歩んできた。筆者らの個人的な感覚では、当初は分析や計算機実験の環境としてはPCよりもWorkstationの方が扱いやすかった。しかし、PCの性能の向上とGUI環境の普及とともに、科学計算のためのソフトウェアやツールもPCへとシフトしてきた。しかし、データ分析やプログラミングを得意とはしていない場合に大量のデータを扱うためには、従来からのように最初からプログラムをするよりも、専用のソフトウェアを使う方が効率が良い。しかし、こうしたソフトウェアは概して高価であるため、持つものと持たざるものの差が優劣の差に直結してしまう。

そこで、協議会を通じて分析ツールを扱う企業に、貸与という形でツールを提供してもらえないかを打診した。これまでにSAS Institute Japan, SPSS(現IBM)、数理システム、日本SGI、ターボデーターラボの各社から賛同を得た。

これらは無料で貸与していただいているが、利用後に使用レポートを提出してもらうなど、貸与の御礼としてフィードバックをお願いしている。

3.3 外部発表

毎年のコンペティションは、3月の成果報告会で終了となるが、研究成果は様々な形で外部に発表されている。ここ数年は、年に50編を超える学会発表や論文投稿が行われている。中でも、日本オペレーションズ・リサーチ学会では、同学会の機関誌で毎年研究成果についての査読付き投稿論文の特集を組ませていただいている。研究者にとっては発行時期が決まっている査読付き論文を掲載する格好の機会となっている。さらに、長年の活動を高く評価していただき、同学会の事例研究賞への推薦を協議会から行うことを許されている。

学生は本コンペティションの成果を直接・間接的に卒業論文や修士論文に使用するというのも多くなってきた。また、博士論文の一部に研究成果が含まれるということも数例出てきている。さらに、企業参加者による、特許公開、出願なども行われている。ただし、

これらに関して、参加者がデータ提供企業側に対価を要求してはならないことも誓約書には明記している。もちろん、貸与したデータをそのまま直接ビジネスに使うことは禁じているが、このように方法論や知見が学術、実務両面に寄与しているのは大変うれしい。

3.4 コンペティションのメリット

参加者のメリット 参加者の第一のメリットは、なによりも最新のリアル・データを入手し、研究に活用できる点である。社会科学系の実学志向の研究においては、実社会との緊密な関わりは重要である。しかし、情報管理に対する意識の高まりなどもあって、近年では以前ほど企業がデータを提供しにくくなっている。受託研究などでデータを入手する方法はあるが、研究成果を公表できない場合もあり、研究者にとっては不便な場合もある。本コンペティションでは学術利用を想定したデータを提供しているため、研究成果の外部発表はむしろできる限り進めて欲しいというスタンスを持っている。

本コンペティションでは、参加者に共通のデータを提供するため、研究発表時にデータの情報がお互いに共有されている。したがって、視点の相違点などを確認しながら発表しあうことができる。また、中間発表と最終発表の発表を義務付けているため、中間発表で他のチームの研究に刺激され、最終発表で見事に昇華されている例も多々見受けられた。

特に、研究者にとっては、学術的には新しい分析の切り口の発見や手法の提案の糸口になり、企業参加者にとっても、新たなスキルや実務上の知識の獲得や、自社リソースの適用、さらには新人育成テーマとして利用するなど様々なメリットがある。

データ提供企業のメリット データ提供企業の関係者の方々も研究発表会には積極的に出席いただいている。50以上のチームが、あらゆる方面から分析をするという点で、今までの社内業務視点とはまったく異なる視点で分析される点では、データ提供企業にとっては新たな視野を広げることができる。成果報告会などに参加いただいた後、分析が大変興味深かったという感想を毎年いただいている。

また、コンペティション終了後にもっと詳しく研究成果を聞きたいという要望があった時には、可能な範囲で当該発表チームを紹介したり、その後共同研究に発展したケースもある。

4. おわりに

平成22年度で17回目を数えた本コンペティションであるが、何より継続した参加者とデータ提供企業あって開催が継続できる。まずはこれまで関係されたすべての方々に御礼を申し上げたい。

こうした取り組みは、海外でもたとえばACM/SIGKDD (The Association for Computing Machinery/Special Interest Group on Knowledge Discovery and Data Mining) が1997年からKDD-Cupを開催しており、世界各地からの参加者を得ている。本コンペティションとの違いは、あくまでも計算機実験的な要素が強く、明確な評価基準により自動的に順位を付けている点にある。また、分析期間もfast trackは5日、slow trackで1か月ときわめて短い。これは、技術者が自身の技術のレベルを確かめるという点では大変優れている。しかし、本コンペティションでは、大きくはマーケティングという枠組みの中で、与えられたデータに対して、どのように分析目標を設定し、その目標にむかってどのように分析を組み立てていくかを考えることが必要であり、本コンペティションは約半年かけて、取り組んでもらっている。このことは、一つの課題にじっくりと分析と議論を繰り返しながら成果を作っていくために必要な時間と考えている。

毎年積み重ねてきた本コンペティションもあと3年で20年を迎える。こうした取り組みを産学連携の一例として参考にしていただければ幸いである。最後になるが、今後とも皆様のご支援・ご参加をお願い申し上げます。

参 考 文 献

- [1] 平成22年度データ解析コンペティション発会式資料
<http://www.isc.senshu-u.ac.jp/thc0640/dac22.html>
- [2] ACM/SIGKDD KDD Cup ウェブサイト <http://www.kdd.org/kddcup/>

(2011年4月8日 受付)

[問い合わせ先]

〒135-8671 東京都江東区豊洲3-3-9

豊洲センタービルアネックス (株)NTTデータ技術開発本部
中川 慶一郎

TEL: 050-5546-2297

E-mail: nakagawaki@nttdata.co.jp

著者紹介

なかがわ けいいちろう
中川 慶一郎 [非会員]

1990年早稲田大学理工学部工業経営学科卒業，2000年早稲田大学院理工学研究科博士課程満期退学，1992年NTTデータ通信(株)(現(株)NTTデータ)入社，2011年より技術開発本部IT活用推進センタBIソリューション担当部長，博士(工学)，マーケティング・エンジニアリング，オペレーションズ・リサーチの応用研究に従事。

なまた め たかし
生田目 崇 [正会員]

1994年東京理科大学工学部第一部経営工学科卒業，99年同大学院工学研究科博士後期課程修了，2002年専修大学商学部専任講師，現在，教授。マーケティング・エンジニアリング，経営科学の研究に従事。