数学月間(SGK)だより

谷 克彦

■2016年4月の米国MAM(Maths Awareness Month)のテーマは,「予測の未来」であった.レイティング,ランキング,世論調査などが取り上げられた.

15年ほど前になるが,私は「パターン識別」[1]の翻訳に参加(第7章:確率論的手法の翻訳を分担)した.原書はスタンフォード大などの大学院テキストに使われ,パターン識別のトピックスを,基礎的原理(線形代数,統計推論)から説明する名著だ.応用は,欠落情報の推定,指紋・生体・顔認証などから,心理学の分野にまで広がっている.私は,非晶質構造の原子配列を解析するために,RMC (Reverse Monte Carlo) modellingに従事したが,これは,系を構成する原子座標をランダムに変化させ,計算した理論値と観測された実験値が一致するように探索を進める手法である.

今年のMAMのエッセイを読むうちに,これらの私の経験はこの分野に無縁ではなさそうに思えた.評価関数を作り探索する手法は,どの応用分野でも同じである.しかし,15年間でこの分野は大変発展したようだ.

現代は,衛星からスマートフォンまで大小のソースから,データが絶えまなく集められている.新しい予測解析法が期待でき,数学,コンピューター・サイエンス,データ科学,統計学を専攻するには,現在は実り多いすばらしい時代である.

「No.1は誰か——レイティング(評価)とランキングの数理」 by Langville[2]

webサイトを渡り歩き,あるサイトで買い物をしたとする.そこに導いた各webサイトの貢献率は如何様なものだろうか? googleの各webサイトのレイティングはどのように計算するのだろうか?

サイト間の遷移確率を成分とする遷移行列を作り,この行列を各サイトの状態に作用させた結果新しい状態になると考える.何度も遷移が繰り返され状態が収束するなら,遷移行列の固有ベクトルを求めれば,各サイトの状態(滞在確率)がわかりランキングが求められる.ここでは線形代数が活躍するし,現在の状態から次の状態が決まるマルコフ連鎖で遷移行列*)が定義される.

さて,Langvilleのエッセイは,3月の狂気(March Madness)で始まる.これは,毎年恒例のNCAAカレッジ・バスケット・トーナメントで,数百万人のファンがこの一月間続くトーナメントの各試合の勝者を当てようとする.ブラケット・チャレンジというのは,インターネットで行う各試合の勝者を当ててポイントを競うことらしい.

学生たちは,数学モデルのみに基づき,ブラケットを埋める方法を学ぶ.2つのモデル(Colley and Masseyモデル)は,チームの評価に線形システムを用い,もう一つのモデル(Eloモデル)は反復更新を用いている.毎年,学生たちはいろいろなデータを加え,モデルが洗練されていく.

Amazonの「これを買った顧客はこれも買う」のような推薦システムや,企業が集めたデータから,顧客の行動を予測する研究には,クラスタリングと最隣接クラス分けのツールを用いている.携帯電話のつながり易さの評価には,学生Tyler Periniが,うまく接続し伝達できる物理過程をマルコフ連鎖でモデル化した.

短いテキスト・サンプル(ツイートやfacebookの「今なにしてる?」程度)の解析から,筆者の性格を知る心理学分野への応用も研究され,今年の大統領選挙戦でも候補者のテキスト分析から面白い傾向を見つけている.

■選挙の開票率が数%なのに当確が出たりする.これはレイティングの予測で,トーナメントの勝ち数の推移から1番を予測するのと同じ原理だ.しかし,1票も開票しないうちに,当確を出す必

*) 各webサイトを頂点とし,頂点間の遷移を矢印で表すと,有向グラフができる.サイト間の遷移確率をこれに書き込むと遷移行列になる.

要があるのか？　意見を聞かれたこともないのに，私の意見が反映されるのか？　納得できず大変不愉快に感じるのは私だけではないだろう．

サンプル集団にバイアスがある．サンプル数が小さい．このような世論調査やビデオリサーチは危険だ．誤った世論調査が出され，人々はそれに引きずられてしまう．よく起こることだ．

「外れた世論調査――それらの限界を理解すべきときだ」 by Brain Tarran[3]

2015年の英国の総選挙で，保守党と労働党の票獲得は，予測された「統計的デッドヒート」が実現せず，保守党が労働党に対し7ポイントの優位で下院の多数を勝ち取った．選挙直後に，外れた世論調査の原因研究が英国世論調査会議BPCと市場調査協会MRSによって立ち上げられ，2016年3月に報告書(120ページの長文)が出た．本稿の引用は，Tarranによる報告書の解説(本節の表題)からである．

「サンプルが母集団を代表するものでなかった」が，世論調査ミスの主原因であるというのだ．世論調査組織が使ったサンプル補集の方法が，労働党有権者を過剰に，保守党有権者を過少に系統的に集め，適用された統計的調整手順も，これらのエラーの低減に効果がなかったという．報告書が勧告する改善提案は，将来起こりうる世論調査ミスのリスクを低減するが，リスクそのものを取り除くものではないことに注意しよう．

世論調査では，今後も非ランダム・サンプリングを使用せざるを得ない．ランダム・サンプリング(確率的サンプリング)は，実行するのに，費用と時間がかかる．しかし，非ランダム・サンプリング(非確率的サンプリング)に比べて明らかに優れている．非確率的サンプリングではサンプルに偏り(バイアス)が生じ易いのだ．回答者がランダムに選択されるなら，母集団のすべてのメンバーに，調査参加者となる一定のチャンスがある．これ自体は，得られたサンプルが，母集団の完全な代表であると保証するものではないが，選択のランダム性は，代表されるグループの外部/内部を調整するためのサンプリング理論の適用を可能にする．また，サンプルへ自己選択される可能性を下げ，回答者の採用過程で，バイアスがかかるリスクを軽減できる．

研究報告書を読んで失望する読者もいるだろうが，失望が畢竟実用主義への道を与え，世論調査の難しさと不確実性を理解することになる．母集団でなくサンプルで解析するのだから限界がある．世論調査は将来起こるかもしれない行動について，有権者のようなよくわからない母集団を調べるので苦しい闘いに直面している．

世論調査の実施方法の高い透明性と，その推定の不確実性レベルを明確に伝える責任がある．それぞれの政党の支持率の信頼区間と前回の公開世論調査に対するそれぞれのシェア変化の統計学的有意差検定を合わせて報告することを報告書は勧告している．

メディアのコメンテータは，世論調査で出た政党支持のわずかな変化を過剰に解釈する傾向があり，証拠が推論をサポートしていない(統計的に有意でない)のに，公衆に党の運命が変わってきたと印象づける．このようなことは避けるべきだ．

●参考文献

1) 『パターン識別』監修：尾上守夫，新技術コミュニケーションズ．
"Pattern Classification", by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, New York(2000).
2) "Who's #1 : The Science of Rating and Ranking", by Amy Langville.
http : //www. mathaware. org/mam/2016/essay/
『No. 1は誰か：レイティング(評価)とランキングの数理』，共立出版より訳書．
3) "Yes, the polls were wrong. But it is time we understood their limitations", by Brian Tarra, 0n 05 April (2016),
https : //www. statslife. org. uk/politics/2752

(たに・かつひこ／SGK世話人)