見過ごされてきた現場の問題
—真に有益なクラスタリングを目指して—

宇野　毅明

クラスタリングは，データ解析の中でも著名な問題であり，学術的に多くの研究があり，かつ応用分野でも盛んに使われている．しかし現実の応用問題では，多くの「見過ごされてきた問題」が存在しており，現場での大きな問題となっている．本稿では，意味解釈のしやすい小さなクラスタを多数見つける問題と，クラスタリングを安定的にする，つまり求解するたびに大きく異なる解が見つかることがないようにする問題を紹介し，シンプルで効率的なアルゴリズムを紹介する．学術界で盛んに研究されたアルゴリズムは実用でもよい性能を発揮し，利便性が高いと思われがちだが，決してそうとは限らないのである．

キーワード：ビッグデータ，クラスタリング，アルゴリズム

1.　はじめに

モデル化．現実社会の問題を数理的な手段で解決するというプリンシプルをもつオペレーションズ・リサーチにおいて，モデル化は中核となる概念である．モデル化自体は物理などの自然科学から社会科学に至るまでさまざまな分野で扱われているものであるが，オペレーションズ・リサーチにおけるモデル化は，独自の質をもっているだろう．現実問題，特に産業界の問題を相手にしているため，コストや納期など契約やコンプライアンスとしての明確な要求項目があり，一方で現場での運用性の高さが求められるなど，ほかの分野には見られない要件がある．そのため，研究者は現場の声に耳を傾けながらよりよい解決方法を模索するのであるが，一方で，ある程度成功が収められた問題においては，以後の研究は学問的動機づけによって展開が進むものもあり，そのようなところでは，原点である現場での運用性の高さがフォーカスから外れてしまうこともある．これは学問として大きな問題であろう．本稿では，データマイニングの代表的な問題であるクラスタリングにおいて，長い間見過ごされてきた現場的な問題を紹介し，それがなぜ起きてしまったのか，その理由を少し考察したい．また，それらの問題に対する筆者らが提案した解決方法もあわせて紹介したい．

うの　たけあき
国立情報学研究所
〒 101–8430　東京都千代田区一ツ橋 2–1–2
uno@nii.ac.jp

2.　マイクロクラスタリング

クラスタリングは，いくつか，あるいは大量の項目からなるデータを入力し，それをいくつかのグループ（クラスタ）に分ける問題である．このとき，解に対するなんらかの評価尺度，つまりコスト関数が与えられるわけではなく，また正解が存在するわけでもなく，「もしデータが（意味的に同質な）いくつかのクラスタに分割できるのであれば，その分け方を答えよ」というあいまいなゴールをもつ問題である．通常は，意味と関係しそうな距離を項目間に定義し，「意味的に似たものはなんらかの距離において似ているだろうから，同じクラスタに所属するようにする」「二つのクラスタの間には項目があまりない境界部分があるであろうから，その境界部分を見つける」といった仮説とアプローチを設定し，それを数理的な目的関数として定義することによって問題を定式化する．距離のモデル化はデータ依存であるためユーザーに任せるとして，データと距離が与えられたときに，どのように「クラスタリングのよさ」を評価するか，つまり目的関数を定め，それを最適化するアルゴリズムを構築する，これがクラスタリング研究の主要な目的となっている．

ゴールが明確でないため，クラスタリング手法の評価方法もまた明確ではない．独自の尺度による評価以外でよく行われる方法は，著名なベンチマーク問題における挙動を示すことである．たとえば 2 次元に分布するデータなど，目で見てわかりやすいもので結果を示すものがある．所属するクラスタごとに色を変えて図示することで，どの領域がどのクラスタに所属しているかがよくわかる．この方法は結果自体も可視化しや

すく，挙動の理解が容易であることが利点ではあるが，高次元のデータでは難しいところが欠点である．一般にクラスタリングが求められるデータは高次元であることがほとんどであり，2 次元や 3 次元で図示することはできない．逆に，図示できるのであれば，データの分かれ方自体を画像として理解することが可能であり，クラスタリングを適用する必要はない．主成分分析などの手法で特徴量を取り出して 2 次元に射影する手法もあるが，そもそも二つの要因ですべてが説明できるデータはあまりなく，あるとしたら，やはりはっきりとした特徴をもつデータなので，クラスタリングが必要なほど全体像を捉えにくいとは思えない．現実データを基にしたベンチマーク問題も存在するが，多くの場合は正解が存在しないため，やはり精度を語ることは難しい．既存の著名な手法との比較を行うことで挙動の違いを見る，ということになる．

このように，現実での有用性を示す効果的な「はかり」がないとき，多くの場合著名な，*k*-means や PLSA，Girvan–Newman などのアルゴリズムにおいて人工的に設定した尺度で評価が行われる．その結果，それを追い求めることが研究の主流となっていく．その結果として，現実で起きている問題意識と研究の目標がだんだんとずれていく．このようなことが，クラスタリングでも起きている．

クラスタリングの利用場面の一つに，新聞記事のクラスタリングによるトピック抽出がある．新聞記事それぞれに対して，その記事が含む単語の集合を作る．この集合はバッグオブワーズ (bag of words) と呼ばれる．この集合の類似性を見ることで，新聞記事の，使われている単語の意味での，類似度が測れる．使っている単語が似ていれば，記事の内容も似ている，つまり両者に相関があると考えられるので，この類似度で得られたクラスタは，似た意味をもつ記事の集合になるだろう．似た意味をもつ記事のクラスタは，なんらかのトピック，プロ野球の結果や飛行機の事故，企業の経常利益の報告などに対応すると考えられるため，クラスタを見つけることで新聞記事が取り上げたトピックを網羅性高く抽出できるだろう，というものである．しかし，実際に 100 万ほどの新聞記事データに既存のクラスタリング手法，*k*-means，Girvan–Newman，graph-cut，などを適用すると，トピックに対応するとは思えないクラスタが大量に生成される．既存手法の詳細や実験の条件などの詳細は筆者らによる論文 [1, 2] を参照されたいのだが，おおざっぱな特徴を述べると，「1 万，10 万記事からなるような巨大なクラスタ」が数個と，「2, 3 個の記事からなる極小のクラスタ」が大量に出てくる．巨大なクラスタには経済から社会，スポーツまで多岐にわたる記事が入っており，とてもそのトピックを見いだすことはできない．一方，極小のクラスタのトピックは明確ではあるが，同じトピックのクラスタがほかに多くあるため，トピックごとにまとめていることになっていない．既存手法の多くでは，生成されるクラスタのサイズがべき乗則に従うために起きると考えられる．多くのデータでこの傾向が見られることから，この現象は，モデルのもつ性質と，アルゴリズムの性質が関係する本質的な性質であると考えられ，より精度の高い最適化を行うことで避けられるとは考えにくい．クラスタリングは解ける問題だと言われているが，記事分類においては，これが現実である．

ここで着目したいのは，トピック抽出においてはクラスタがほしいのではなく，トピックが得たいのだということである．つまり，クラスタは，そのクラスタが対応するトピックを意味解釈して取り出すために使われているのであり，データを分割したいわけではない．意味解釈がしたいのであれば，全体で見たときにある程度境目とおぼしきところにそって切り取られた，数理的に妥当なクラスタであっても，中に意味の異なる記事が大量に含まれていれば，役に立たない．逆に，互いに類似する記事だけで構成されたクラスタがあれば，その意味解釈はしやすい．トピックを抽出するだけが目的であるなら，類似する記事だけを集めたクラスタを，トピックの数だけ提示できればよいはずであり，余った項目は捨ててしまえばよいのだが，そのような設定で行われたクラスタリングの研究は存在しない．

一方，このような，類似する記事の集合を見つける，という目的には，パターンマイニングやコミュニティマイニングが用いられることもある．これらの手法は，類似する物同士を枝で結んだ近接ネットワークを作り（頂点集合が記事の集合，類似度が閾値以上の記事の間に枝を張る），その中から極大クリークや極大 2 部クリークを列挙することで，似ている記事の集合を直接的に見つけ出す．ここで列挙を用いているのは，見つけ損ないをなくすためで，大量のトピックがすべて見つかるようにしている．しかし，100 万もの記事があるようなデータでは，極大クリークの数は数兆以上にもなり，現実的に列挙することは不可能である．例え列挙できたとしても，同一トピックに対応するクリークが大量にあるはずで，大量のクリークを整理すること自体がクラスタリング問題となってしまい，求解は不可能である．

意味解釈は，新聞記事にとどまらずクラスタリングのさまざまな応用において求められる．自動化が目的のときは，意味解釈は求められず，データを完全に分割し，それぞれの要素にラベルを付け，広告などのアクションをする．これは，既存クラスタリング手法の価値観に合致する．また，データの性質を表す指標を算出したい，というような目的でも，クラスタリングの目的関数自体が意味をもつため，都合がよい．しかし，マーケティングにおける顧客分析，位置情報から人流の解析や都市構造の解析をする場合，医療データから人間の体質を分析する場合など，多くの場合はクラスタから得られる知見の意味や質が重要となり，意味解釈の重要性が非常に高い．また，このような場合は，データ中に多くのクラスタが陰に存在し，計算的なアプローチなしには問題解決ができないことが多い．しかし，クラスタを網羅的に見つけると大量の解が出てきてしまう．大きさや禁止ワードなどによって制約を与え，解数を減らそうというアプローチもあるが，しょせん 1 兆が 100 億になってもだめなものはだめであり，大量の類似するクラスタは排除できない．

未だに，クラスタリングに対する現場での大きな課題は，意味解釈がしにくいことである．この難しさにアプローチする方法はいくつもあるだろうが，クラスタリング研究の多くが，分野で定番となった評価値による精度の追求に終始しており，このような方向は着目されてこなかった．さらに言うならば，これらの評価値は意味解釈においてはまったく注目されていない．自動化への応用でも，評価値が実際の利便性やコストの軽減にどれほど役に立っているかは不明なことが多い．それでも，学術的に評価されやすい，評価値の意味での精度向上を目指す研究があまりにも多いことは，現実の問題を直視した研究が行われていないであろうことを暗示している．

本稿の目的はこのような「著名で重要な問題においても，現実を直視しない風潮が強い」ことがしばしばあることを伝えることが目的であるので，ここでクラスタの話をいったん終了してもよいのだが，提言だけに終始するのは無責任であるとも考えられる．その意味で，筆者らのアプローチを少しだけ紹介したい．

筆者らのアルゴリズムは「データ研磨」という．類似する項目からなるクラスタを見つけるには，近接ネットワークはあまりにもノイズが多い．本来クリークであってほしい，クラスタに対応する頂点集合の中には，枝でつながれていないものがあり，その集合はクリークではなく，密な部分グラフになっている．密な部分グラフには大量の極大クリークが含まれ，これがパターンマイニングやコミュニティ発見が大量の解を生成する原因となっている．そこで，密な部分グラフをクリークで置き換えることにより，ネットワーク中のクラスタを「明確に」しよう，というのが，データ研磨のアイディアである．

しかし，密部分グラフの網羅的発見は非常にコストが高い．しかし列挙的なアプローチをあきらめてしまっては，見つけ損ないが大量に発生するリスクを抱えることになる．そこで，異なるアプローチで，密部分グラフに含まれそうな頂点ペアを直接的に発見する．このアプローチは以下の実行可能仮説に基づいている：「頂点 A と頂点 B がある密部分グラフに属するなら，A と B 両方と隣接する頂点が多数ある」．この性質は，リンクディテクション（Facebook などにおける友達と思われる人の紹介）でも使われているものである．この仮説を用いて，与えられた閾値 k に対して新しいグラフを作る．作り方は，すべての二つの頂点の組に対して，「共通の隣接頂点数が k 以上なら枝をはり，そうでなければ枝をはらない」というものである．こうすると，同一のクラスタに所属していそうな頂点の間には枝があり，そうでなければ枝がない，というネットワークができる．このグラフではクラスタの形がより明確になっているだろう．この処理を，ネットワークの変化がなくなるまで続けることで，クリークとクラスタが一対一で対応するであろうネットワークを作る．最後に極大クリークを列挙することで，クラスタを得る．これがデータ研磨の手続きである．

実際にデータ研磨を適用すると，新聞記事はきれいにトピックごとに分類され，会社の取引関係データからは業界の分類が得られ，多数のクラスタをノイズで隠蔽したベンチマークデータからは元のクラスタがきれいに復元される．実際に意味解釈のしやすいクラスタは得られるわけで，つまりはほかにも同様の結果を得る方法はたくさんあるだろうと考えられる．しかし，現実にはそのようなアルゴリズムは少なく，いかに現実問題に目が向けられていなかったかがわかるのである．

ちなみに，データマイニング業界では，昔からパターンマイニングやコミュニティマイニングの解の爆発については議論があった．その中で，解の爆発を解決する一つの方法として，ノイズを考慮したモデルが有効であると言われてきた．近接グラフの中で，クラスタはクリークではなく密な部分グラフになる．よって，極大クリークの代わりに極大な密部分グラフを見つければ，解の爆発は押さえられるだろうというものであ

る．しかし，現実はまったく逆で，極大な密部分グラフの数はむしろ極大クリークよりもはるかに多い．マイニング業界において，現実のデータに対する観察や推察が弱かった，このことは残念である．

3. クラスタリング安定化

現実の課題が直視されていない例の二つ目としてクラスタリングの安定化を紹介したい．多くのクラスタリングアルゴリズムは，その計算において乱数を利用している．典型的には，初期解を乱数で生成し，局所最適解を得るものである．そのため，実行ごとに異なる解が生成されることになる．これらの局所最適解は，似たような評価値をもつため，評価値の面からは質に違いがあるとはみなされない．そのため，研究者の視点からは，異なる解が出てくるのであれば，好きなものを選べばよい，となる．データは本質的に何通りかの分割の仕方を含有し，それが出てきているのであろうという考え方である．

現場では，これが大きな問題となることが多い．たとえば自然科学の研究において，再現性がない，毎回異なる解が得られるような実験は，科学的事実として到底受け入れられない．ほかにも，小売店の顧客行動分析では，その結果によって経営判断が変わり，たとえばいくつかの店舗を閉鎖することもありうる．たくさんある解の中からなぜその解を選んだか，説明責任が生じる．適当に選んだ，では通用しない．いくつもの解を生成し，しっかりと意味解釈をして実情と合う解を選ぶ必要がある．さらには，店舗閉鎖の結果，何人もの従業員が人生の大きな転機を強制的に迎えるかもしれない．不幸な道を歩む人もいるかもしれない．説明が求められなくとも，精神的に大きな重圧である．このほかにも，同じ会社のデータを 1 年後に，1 年分のデータが増えた状態で分析すると，まったく違う解が出てきてしまい，違いの説明に苦労するという話も聞く．同じような解を得るのが困難なのである．研究者の多くは精度が同じならば問題ないと考えがちだが，現実ではその影響を無視できず，重大な問題を引き起こすこともあるのである．

たとえばインターネット検索をしてみると，既存研究ではクラスタリングの安定化がまったく出てこない．robust clustering のような単語を用いるといくつかの研究が見つかるが，いずれもバイオ情報学の研究である．これらの研究では，robust clustering は，「仮説に対応するクラスタが必ず見つかる」という類いの意味であり，仮説に対応しないクラスタについては何も気にしていない．全体的な安定性の意味合いとは大きく異なっている．

ランダムに得られる解を安定させる自然で素直な方法の一つは，複数の解を得てそれらの平均，あるいは中央を取ることである．実はこのような研究はすでに行われており，アンサンブルクラスタリング，あるいはコンセンサスクラスタリングと呼ばれている．提案されている多くのアルゴリズムは，クラスタリング間に距離を定義し，入力したすべてのクラスタリングの中心となるクラスタリングを求めるものが多い．これらは入力サイズの 3 乗以上の時間を必要としており，現実的に扱われる大きなデータには適用できない．しかし，数理的な結果は出しやすいために，多くの研究があるものと考えられる．一方で，現実的な時間で動くアルゴリズムとして，複数のクラスタリングで得られたクラスタをクラスタリングするメタクラスタリングアプローチや，同一クラスタに所属した回数を類似度として距離を再定義してクラスタリングを行うアプローチなどが提案されているが，計算実験を行っている研究が見当たらないなど，研究が深められている形跡は少ない．実際のデータでの挙動や現場での手法の意味などについては注目されてこなかったことがうかがえる．

実際の買い物データで検証してみると，k-means を用いた場合，二つのクラスタリング結果の重なり具合が半分程度であることがわかった．ここでは，重なり具合を，片方のクラスタリングのクラスタに対して，もう一つのクラスタリングのクラスタで最も似ているクラスタとの共通部分がどれくらいであるか，で定義する．計算実験ではその重なりは半分程度であった．これでは，なかなか同じ意味解釈がされるクラスタが生成されるとは言いがたい状況である．実験や，クラスタリング間の類似度などの詳細については，筆者らの研究報告 [3] を参照されたい．一方，データ解析現場で比較的評判のよい PLSA という手法がある．これは，クラスタの特徴を抽出しつつ，大きさのバランスを取りつつクラスタリングを行う手法であるが，これを用いた場合は結果がよく，重なりは 2/3 程度となる．これくらいなら意味解釈である程度類似するものが出てきそうである．

一方，データ研磨のクラスタリング手法を用いて，k-means や PLSA の複数の解に対して上記コンセンサスクラスタリングを適用した場合には，さらによい，重なりが平均的に 8 割程度となる解が得られることがわかった．これは，複数回クラスタリングを実行して

も，よく類似する解が得られることを示しており，意味解釈的にはほぼ同じ意味をもつ解が得られると言ってよいだろう．

クラスタリングにおいて，解の精度は現場的にはあまり参考になるとは言えず，モデル化も難しい．一方で安定化の指標化は簡単であり，かつ現実的に大きな課題となっている．にもかかわらず長い間安定化が着目されてこなかったことは，研究分野の大きな失点であるのではないか．

4. おわりに

以上のように，著名な問題でも，現実と研究の間に目的意識の大きな乖離がありうる例を紹介させていただいた．この例から得られる教訓は，産業界やビジネス，自然科学などの計算技術を利用する立場の人に対しては，「盛んに研究されているものが現実で役に立つものでは限らない」ということと「みんなが困っていることが研究されている，研究者に認知されているとは限らない」ということであろう．技術の詳細までわからなくとも，少なくともその技術のもつ意味や振る舞いは理解する必要があるだろう．研究者に対しては「現場には今まで着目されてない問題が潜んでいる」「学術分野の価値観しか見ていないと研究は現実からどんどん乖離する」というところであろうか．研究者は，論文を書くことが生業であり，それゆえに研究として成果につながりやすい研究課題を選んでしまいがちである．逆に現場の問題は論文になる形で研究するのが難しい．現場の価値観を取り入れた問題を考える，それが難しければ，論文を生産することと，現実社会に貢献することは別の活動として分けて両方やる．このような考え方が研究では大事であろう．

参考文献

[1] 宇野毅明，中原孝信，前川浩基，羽室行信，“データ研磨によるクリーク列挙クラスタリング，”第 146 回情報処理学会アルゴリズム研究会，2014-AL-146 No. 2, 2014.

[2] T. Uno, H. Maegawa, T. Nakahara, Y. Hamuro, R. Yoshinaka and M. Tatsuta, “Micro-clustering by data polishing,” *BigData*, pp. 1012–1018, 2017.

[3] 宇野毅明，岩崎幸子，中原孝信，中元政一，羽室行信，“データ研磨によるコンセンサスクラスタリングの精緻化，”人工知能基本問題研究会，**106**, pp. 43–50, 2018.

