B-014

データ研磨と NYSOL を用いた解釈性の高いデータマイニングへの道
Toward Interpretable Data Mining with Data Polishing and NYSOL

宇野 毅明 *
Takeaki Uno

データマイニングは、データの中から局所的、部分的な特徴や構造を取り出し、知識発見の助けとする手法の総称である。認識や予測など、ある程度認知的に直接的な機能を有する機械学習手法に対して、データの可視化、抽象化、意味構造の抽出など、自動化に直接利用することよりも人間の認知に直接訴えるものを求めることに主眼が置かれている。代表的な問題としては、アソシエーションルール発見、クラスタリング、パターンマイニングなどがあげられ、それぞれ 10000 を超える関連研究が存在する、ある種花形の分野である。

機械学習では、精度を高めるという最適性の目標があるが、データマイニングでは、網羅性や局所性、独立性など、異なる尺度が主眼となることが多い。そのため、特徴抽出ではなかなか見つけられないような特徴が、たとえばアソシエーションルールの形で見いだされる。スーパーマーケットで、ビールと紙おむつを一緒に買う人が意外に多い、という分析結果が出たという逸話があるが、このように一部の人に限定した特徴を、商品の組合せの形のようなはっきりとした形で提示することは、特徴抽出的な技法では難しく、データマイニングが持つ固有の機能である。そのため、バイオ情報学などをはじめとする多くの分野に応用研究、応用事例が存在する。

その一方で、データマイニング技術の利用には、大きな障壁があるのも事実である。機械学習のような、統一的な問題設定の元で活用することが簡単ではなく、そのため、研究者などプロフェッショナルの利用が主体となっている面がある。また、技術としての行き詰まり感もあり、技術の進展に着目した研究は、年々落ち込んでいる。これは、実利に結びつくような新しいコンセプトをもとにした問題設定がなかなか創出されず分野全体として前に進んでいる感じが乏しくなってきているからであろうと考えている。実利用で次々に発生する、新しいタイプの困難性も、その一員である。昔は「●●が実現できれば世の中は変わる」と言われたようなことがいくつもあり、その目標に向かって研究が行われ、例えばアルゴリズムの性能は 1 万倍以上の高速化を成し遂げるほどに発展した。しかし、高速化を成し遂げた後には、高速な手法が生成する非常に大量の解をどのように処理するか、という新しい課題が発生し、これを自然な形でクリアできていないために行き詰まりが発生しているのだろうと思っている。高速高性能なアルゴリズムを持つデータ解析パッケージなどの実装が少ないことも、一因であろう。たとえパッケージを使ったとしても、長大な時間がかかったり、かけたわりにはたいしたことのない解が出てきたりと、ユーザが満足いく結果が得られないことがあるのである。

著者は、2004 年に、超高速パターンマイニングアルゴリズム LCM を開発した。その実装は国際学会のコンペティションで優勝し、現在のところも世界最速である。当時は、これでデータマイニングの利用は爆発的に増加するだろうと意気揚々とし、実装を早速公開した。着々とダウンロード数は上がるのだが、以前と変わらず商用などより一般的な利用はあまり進まない。ダウンロードしたユーザから質問や要望が来ることが多く、それに応えることで利便性を向上すれば一般での利用も増えるだろう、との考えから機能追加を繰り返し行ったが、結果機能が複雑すぎてよくわからないという質問も受けるようになってしまった。

ユーザからの要望は、典型的には、「新しい制約の追加」(出力するパターンの大きさ、同時に含めてはいけないアイテムのペアなど)、「異なるパターンの扱い」(系列パターンやグラフ)、「文字列を入力させたい」「出てくるパターンが多すぎるのでなんとかしてほしい」「コマンドが複雑」「可視化できないか」などがあり、他にもインストールして実行するまでが手間だ、というものある。データとモデルをしっかり考えないと使えないというところも、ビギナーの参入を難しくしているだろう。

最初のうちは、これらの要望に真摯に答えることが最善と考え、機能追加や系列マイニングなど新しいアルゴリズムの開発などを繰り返した。しかし、そのうちにどうもこれは一般向けには筋が悪そうだという感覚を得るようになった。データフォーマットは、分野やパッケージごとに異なり、様々である。それに逐一対応していくのは難しい。それよりも標準的なフォーマットにのみ対応し、その他の部分は前処理ソフトにまかせたほうが、ユーザもむしろ使いやすい。そういったパッケージソフトが存在する方が、よほど魅力的だ。制約についても、例えば「この大きさ以上のパターンは見つけない」という制約は、アルゴリズムの特性上、枝狩り手法の追加となり、大幅な速度向上を伴うので、機能追加すべきだが、「これ以下のパターンは見つけない」は、機能追加を行ったところで、単に解のスクリーニングをするだけでアル

*国立情報学研究所・情報学プリンシプル研究系

ゴリズムのパフォーマンスは変わらない。つまり、後処理ソフトにまかせて十分なのである。このように、アルゴリズムとして本質的なパフォーマンス向上に関わるものだけに着目し、残りは周辺プログラムにまかせたほうが、柔軟なシステムが設計でき、よってユーザの利便性と自由度も向上するのである。多種のパターンに対するマイニングでも、多くの場合は通常のパターンマイニングに問題変換を行うことが可能であり、そのようなソフトを構築する方がよほど早いのである。

もう少し本質的なことを考えれば、パターンマイニングは、やはり計算機科学のプロ向けのツールなのである。大量の解を網羅性高く得たいのは、主にプロであり、一般ユーザはもっとざっくしりたものがほしい。コミュニティーマイニングなどのクラスタリングでも、ユーザはざっくりとしたわかりやすいクラスタがほしいのであって、網羅性高い完全解を要求しているわけではない。その点を鑑みると、プロ仕様のものにいくら機能追加をしても一般人には響かない。一般人のニーズに焦点をあてた仕組みが必要なのである。

我々が開発したデータ研磨は、パターンマイニングやコミュニティマイニングのもつ「大量の解の生成」を回避し、たがいに類似しない解を適切な量で出す技術である。データを明確化する、というアプローチをとっているため、既存のヒューリスティックや制約追加による方法とは異なり、頑健性や網羅性を担保している。一般的に、パターンマイニングを行うとき、パターン自体に対する興味よりも、パターンを含む項目、パターンを含む項目が作るクラスタに興味がある場合が多い。おむつと缶ビールが一緒に売れることよりも、いったいどういう人がその組合せを買うのだろうか、購買者に興味があるのである。データ研磨は、購買データであれば、ユーザを購買履歴の類似性でクラスタリングして、その特徴をパターンとして出力する。これにより、履歴の類似性からの意味解釈がしやすいクラスタを得ることと同時に、適度な粒度、適度な数のパターン生成を行うことができる。この作業は、顧客を細かいグループにわけてその特徴を見ているという意味で、購買データを抽象化して、見やすくすることに対応している。抽象化によって、「ざっくり見たい」「可視化したい」「見にくいから制約を入れてしぼりたい」「大量にあるから高速化したい」といった要望をすべて斜め方向から解決することにつながっている。

データ研磨と LCM は、現在 Nysol プラットフォームの上にも実装されている。Nysol と連携することにより、ユーザインターフェース、前処理後処理などの問題をすべて解決し、Kizuna による可視化で抽象化されたデータを効果的に閲覧し、ひらめきや理解を得られるようになった。楽に行った作業で効果的な結果が得られれば、機能的な不満は出てこないものである。あとは、ユーザの主体的な工夫によるノウハウの創出で、扱えるデータや課題が増えていくことを期待している。Nysol を使っているエンジニアやデータ解析者などの「セミプロ」の方々と、一般の好奇心とチャレンジ精神でデータに取り組む方々には、強い機能と高い利便性を持つツールの提供により、より創造的なデータ解析への道を開いていけるものと期待している。