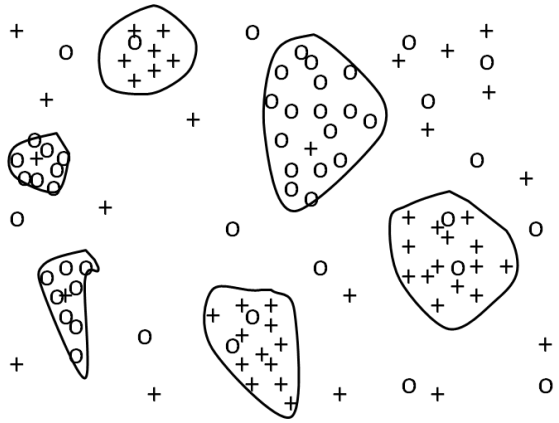
多様性の獲得に向けた次世代マイニング技術

国立情報学研究所・情報学プリンシプル研究系・准教授
宇野 毅明

近年、センサ技術を中心とするIT技術の発達により、多くの分野でビッグデータが利用されるようになった。このようなビッグデータの解析による将来予測やデータの欠損部分の補完を行う研究は、現在のデータ解析分野の中心的な研究である。しかし、データが多様性を持つ場合、図のようにデータは質の異なる多様なモデルの重ね合わせとなるため、全体求解は困難を極める。データが多様である場合、データの強いまとまりをからなる局所構造や領域を網羅的に発見・獲得し、マイノリティや異常事態にも対応できるような予測を行うことが重要である。このような研究を進めていくためには、今までにないアプローチに基づく効率的な手法の開発が急務である。



図：正例(○)と負例(＋)の多様な分布。いくつもの異質なまとまりができる

本研究ではデータの局所性や意味的なつながりを用いて，内包する意味が同じである対象を１つの粒子にまとめ上げることにより効率良く発見し，解の爆発を自動的に回避しながら網羅的に知識獲得するモデルと高速アルゴリズムを提案する．これにより，多様なユーザを高精度でクラスタリングし，グループごとに適したアクションを行う，といった多様性を利用した効果的な解析を可能となる。

このようなアプローチは、類似解の爆発的な発生や網羅性担保の困難さと行った、簡単な方法では回避できない難しいを本質的に内包している。本研究では、項目間の関連性などのデータの構造を用い，データを整形してより粒子を明確化するデータ研磨という新しい手法を提案する．具体的には、グラフのまとまりを見つけるために密な部分構造を列挙する問題に注目し、明らかに本来枝が有るべき所に枝を追加し，明らかに枝があるべきでない場所は枝を消す、という操作を行う．この結果，枝の欠損により１つのまとまりに対応する複数の密構造が生成されることがなくなり，解の爆発を防ぐことができる．データ研磨は，単に計測の誤差から来るノイズを除去するのではなく，データを，意味を変えずにコンピュータや人間に理解しやすいよう積極的に整える新しいパラダイムである．このような操作「マイクロクラスタリング」を行うことで、データの可視化、強い意味を持つクラスタのマイニングなどが非常に効率的にできるようになる。

この手法の実装は、世界最速のパターンマイニングアルゴリズムLCMの構造を利用した類似検索アルゴリズムと、世界最速の極大クリーク列挙アルゴリズムを組合せることにより、非常に大規模なデータにおいてもほんの数分でクラスタを全列挙できるパフォーマンスを持っている。また、実データにおける実験でも解の爆発は起きず、１つ１つのクラスタは意味的に強いつながりを持つグループから構成されていることが確認できる。