

大規模なクラスタリング結果に対する探検型可視化システムの構築

多飯田祐己[†] 井上悦子[†] 吉廣卓哉[†] 中川優[†]

和歌山大学システム工学部[†]

1. はじめに

本研究では、大規模データに対するクラスタリング結果をわかりやすく可視化するために、グラフを用いた探検型可視化システムの構築を行った。クラスタリングは、遺伝子の発現量解析のためのマイクロアレイ実験など、様々な実験結果の分析にもよく用いられる基本的なデータ解析手法である[1]。本稿では、井上ら[2]による試作システムを改良し、グラフとクラスタリング結果（樹状図）の連携を目的とした二つのビューの追加を行ったので、これを報告する。

2. 準備

2.1. 探検型可視化手法

本手法は、クラスタリングによる分析結果をわかりやすく可視化するために、任意の部分領域に対し、表示粒度を変えながらグラフの形式で閲覧するものである。例えばマイクロアレイによる発現量解析のクラスタリング結果に適用する場合、注目したい遺伝子を中心とする部分的な領域を、遺伝子または遺伝子のクラスタをノード、ノード間の発現傾向の類似度をリンクとするグラフで可視化する。対象領域に対して、細部を見る場合にはノードは遺伝子、全体を俯瞰する場合にはノードはクラスタとなる。このような表示粒度の変更に従ってノードの分割・結合を行うことで、グラフ状態を連続的に変化させながらノード間の関係を閲覧できる。

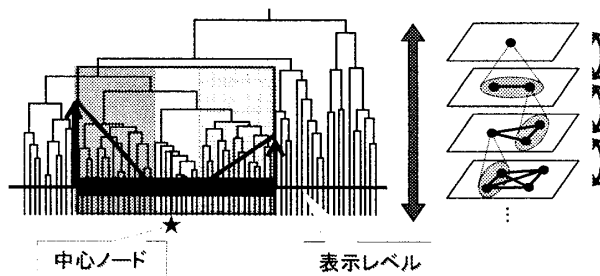


図 1. 探検型可視化手法の概念図

Graph-based Visualization for Interactive Exploration of Large-Scale Clustering Results

Yuki Taida[†], Etsuko Inoue[†], Takuya Yoshihiro[†] and Masaru Nakagawa[†]

[†]Faculty of System Engineering, Wakayama University

図 1 に、本手法の概念図を示す。図 1 左はクラスタリング結果全体の樹状図を表し、可視化の対象領域を矩形で示している。この対象領域をグラフで可視化し、対象領域の移動に合わせてノードの分割・結合を行い、グラフを連続的に変化させる。

本手法の特徴は、クラスタリング結果の対象領域を一定の見易さを保ったグラフで可視化しながら、ユーザ操作による自由な領域の移動に追従してグラフ表示を連続的に変化させられる点である。また、対象領域の移動に伴うグラフの差分情報を、クラスタリング結果のデータ規模によらず一定の計算量で計算できるため、ユーザ操作による対象領域の移動に対してインタラクティブな可視化を実現することができる。

本手法では、大規模なクラスタリング結果であってもグラフの見易さを保つために、グラフの表示ノード数を一定以下に調整する。可視化の対象領域は、注目する要素（中心ノード）と、可視化の粒度を決める樹状図の水平な切り出し位置（表示レベル）で決定される。対象領域の要素数が一定以上であれば、中心ノードからの距離が遠い要素から順にグループ化（親ノードに置換）することで表示ノード数を制御できる。

2.2. 試作ビューアによる評価とその問題点

文献[2]では、提案手法を Java アプレットとして試作し、ノード数一万のクラスタリング結果に適用し評価を行った。その結果、数百程度のデータを扱うのと同程度のストレスを感じない処理速度での操作できることを確認した。

しかし、試作システムのアプレットでは対象領域を可視化したグラフ表示のみであったため、グラフに表示されているノード間の関係は把握できても、クラスタリング結果全体の樹状図のどの領域を可視化しているのかが把握できないという問題があった。

3. グラフと樹状図の連携

本研究では、グラフと樹状図との対応を把握できないという試作システムの問題点を解決するために、二つのビューを追加しシステムを改良した。一つは、現在のグラフで表示している領域がクラスタリング結果全体のどの部分に対

応しているかを表示する「AllTreeView」である。もう一つは、現在のグラフに対応するクラスタリング結果の部分領域を拡大表示するための「SubTreeView」である。SubTreeView は、クラスタリング結果が大規模である場合、グラフの表示領域が樹状図全体に対して対象領域が非常に小さくなり、AllTreeView では読み取ることができないという問題を解消するためのものである。各ビューについて以下で説明する。

AllTreeView では、クラスタリング結果全体の樹状図上に、グラフの表示領域に対応する領域を矩形で表示することにより、グラフと樹状図を対応付ける。また、クラスタリング結果が大規模であっても対象領域の矩形を見つけやすくするために、対象領域と横幅のみが等しい縦長の領域を補助領域として矩形で表示する。これにより、対象領域の矩形が非常に小さい場合でも、補助領域を目安として対象領域のおおよその位置を把握できる。この AllTreeView では、樹状図全体の表示は初期状態から変化しないが、対象領域および補助領域の二つの矩形はグラフの対象領域の移動に追従して更新される。

SubTreeView では、グラフの表示領域に対応する領域のみの部分的な樹状図を表示する。SubTreeView では、グラフの表示ノードと部分的な樹状図の各ノードを一対一対応で表示する。グラフでは表示ノード数を制限するために、中心ノードからの距離が遠いノードをグループ化されており、これらのグループ化されたノードは表示レベルよりも粗い表示粒度の要素となるが、SubTreeView を導入することで、ノードごとの可視化の粒度を正確に把握できる。SubTreeView では、グラフの対象領域の移動に追従して、部分的な樹状図全体が更新される。

4. システム概要

3章で示した二つのビューを実装した可視化システムについて説明する。本システムは、試作システムのグラフ表示部分の Java アプレット（以後、GraphView とする）と、新たに構築した AllTreeView、SubTreeView の三つの Java アプレットを連動させる Java アプリケーションである。クラスタリング結果に対し、指定された中心ノードと表示レベルで決定される対象領域を可視化し、ユーザ操作による対象領域の移動に追従して各 Java アプレットを連続的に変化させる。

図 2 にシステムの画面例を示す。画面左に GraphView、右上段に SubTreeView、右下段に AllTreeView を配置している。SubTreeView および AllTreeView の左端の目盛りは、要素間の結

合距離（類似度）を表す。各アプレットの表示について簡単に説明する。

GraphView は、対象領域に含まれている要素群（ただし、表示ノード数を調整されたもの）をグラフで表示する。各ノードの大きさはそのクラスタに含まれる要素数（マイクロアレイ実験の例であれば遺伝子数）を表し、色の濃淡は中心ノードからの距離を表す。

SubTreeView は、グラフの表示ノードに対応する部分的な樹状図を表示する。各円は GraphView の各ノードに対応する。表示ノード数の調整のためにグループ化されたノードは、表示レベルよりも小さい結合距離となる。これにより、各ノードの正確な粒度を確認できる。

AllTreeView は、クラスタリング結果全体の樹状図と、対象領域および補助領域の矩形を表示する。図 2 の①が対象領域を、②が補助領域を表す矩形である。

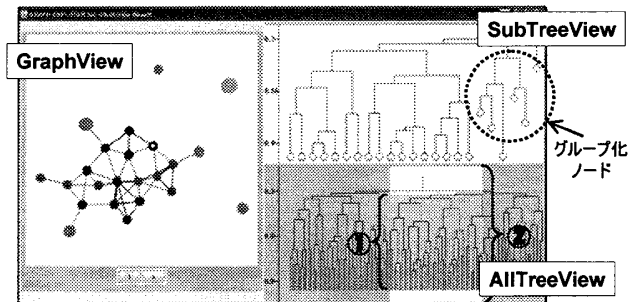


図 2. システムの画面例

5. おわりに

本研究では、探検型可視化システムの試作版に対し、グラフと樹状図を連携させるための二つのビューを追加しシステムを改良した。その結果、グラフと樹状図の対応を把握しながら、クラスタリング結果の部分領域をグラフの形式で閲覧することが可能となった。しかし、実際の分析現場での利用に耐えるシステムとなるには、要素検索などの機能追加やインタフェース改善が必要である。今後は、さらに改良を進め、有用な可視化システムとしての完成を目指す。

なお、本研究は和歌山県地域結集型共同研究事業にて実施した。

参考文献

- [1] 塩島聡, 松本治, 辻本豪三, “わかる!使える! DNA マイクロアレイデータ解析入門,” 羊土社, 2002
- [2] 井上悦子, 吉廣卓哉, 中川優, “大規模クラスタリング結果のグラフによるインタラクティブな可視化手法,” 情報処理学会研究報告 [情報学基礎], vol.2006, no.118, pp.21-28, 2006.