5ZB-06

Mapperの視覚化過程に対する定量評価指標の提案とそれに基づく視覚化パラメータの進化計算的最適化

An Optimization of Mapper's Parameters based on Genetic Algorithm using Quantitative Evaluation of Visualization Results

宮永　翔大朗　　延原　肇
Shotaro Miyanaga　Hajime Nobuhara
筑波大学 (University of Tsukuba)

1. はじめに

近年, 人工知能やIoT（Internet of Things）などの分野の急激な台頭とともに, ビッグデータの解析技術の必要性[1]が増している. これに対して, 大規模かつ高次元のデータを抽象化し, データ解析するユーザに分かりやすく視覚化する手法として, TDA（Topological Data Analysis）[2]のMapper[3]が注目されている. 一方で, Mapperは解析時にユーザ側で調整しなければならない距離, フィルタ関数やクラスタリング手法のパラメータ数が多い. またそれらのパラメータによって, 視覚化にどのような影響を与えるのか, 十分体系化されておらず, 出力された視覚化の結果と, 調整するパラメータの関係をユーザ側がイメージしにくい.

本研究では, Mapperのパラメータ調整の試行錯誤をせずとも, データ解析のきっかけとなるような視覚化の結果を出力するように, これらを遺伝的アルゴリズムに基づき自動調整する枠組みを構築する. そのために, Mapperによる適切な視覚化が行われているか否かを, 対象データセットのデータ同士のつながりに着目した独自の評価指標を定義し, この指標に基づき遺伝的アルゴリズムを用いて最適化する.

2. Mapperによる視覚化とパラメータ調整

Mapperは, 対象データを入力し, 距離関数, フィルタ関数, クラスタリング手法, そしてその各々のパラメータを調整し視覚化を行う (Fig. 1). まず, Step. 1の様にデータを入力する. 入力は, M 次元の特徴量を持つ N 個のデータとする. 次に, Step. 2で, 入力データの2点間の距離を定義するための距離関数を設定する. 距離関数には L_p ノルムを用い, $p \in [1, +\infty]$ の値を調整する. 入力データ全ての2点間の距離を求め, $N \times N$ の距離行列を作る. この距離行列にフィルタ関数を用いることで, 低次元に写像することができる. このとき用いるフィルタ関数は5種類あり, それぞれのフィルタ関数内のパラメータも調整する. 低次元に写像して得られた値をフィルタ値と呼ぶ. Step. 3で, このフィルタ値に基づき, データセットを各Intervalに分ける. Mapperでは, クラスタリングはInterval毎に行う. Intervalに分けるとき, 重複を設定する. 重複を設定することで, 視覚化したとき, 重複したノードを持つクラスタ同士に線を引き, クラスタ間のつながりを示すことができる. Intervalの数を I, 重複の割合を O とすると, $I \in \{1, 2, \ldots, 999\}$, $O \in [0, 1)$ を調整する. Step. 4で, クラスリングの手法を7種類から選択し, 各Interval毎にクラスタリングし, 同じノードを含むクラスタ同士で線を引き, つながりを表す.

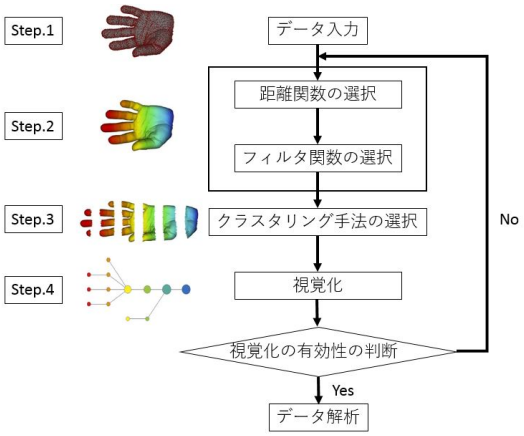

Figure 1: Mapperの処理過程

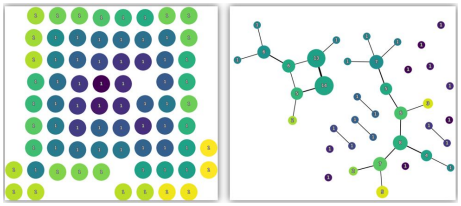

Figure 2: 悪い例 (左) , 良い例 (右)

これらのパラメータを, ユーザが調整し視覚化を行うが, ほとんどの場合, Fig. 2の左図のようなデータ解析には適さない結果となってしまう. Fig. 2の右図のように, ユーザにとって解析しやすいような視覚化が得られるようなパラメータの自動最適化を提案する.

3. Mapperによる視覚化のパラメータ調整のための定量評価指標の提案

視覚化結果を評価する際, Mapperの特徴であるデータのつながりに注目する. データセットを視覚化した際, ユーザにとって解析しやすい結果は, データが何かしらの特徴を持っているとわかる様な出力結果である. つまりクラスタ同士がとつながりを多く持っている状態が望ましい. 本研究では, データはある程度構造を有しているものとして実験を行うので, この仮説を大前提とする. また, 視覚化後のクラスタ内のノード数の平均と, クラスタ全体の分散値も評価に考慮する. 解析する際, クラスタの数が多く出ていると対象のデータセットの特徴が見出しやすい. 以上から, クラスタ同士のつながりが多く, クラスタの平均と分散値が小さいほど良い評価を返す評価指標を設定する. この指標に基づき設定した評価関数を式 (1) に示す. E はクラスタ同士のつながり, C はクラスタ数, A は平均, V は分散値, $a, b \in \mathbb{R}^+$ は係数である.

$$F = -(E - C)^2 - aA - bV \tag{1}$$

このとき, 調整するパラメータの組み合わせ数が膨大になる為, パラメータを GA(Genetic Algorithm)[4] の染色体にコーディングし最適化する. GA を用いた Mapper

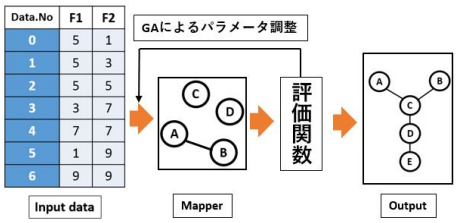

Figure 3: GA による Mapper パラメータ最適化

パラメータの自動最適化の様子を Fig. 3 に示す.

4. 実験

実験には, アヒルの像を 5° ずつずらし 360° から撮影した 72 枚の画像データを用いた. 入力データは, 画像のピクセル値, 12×12 次元を持つ 72 個のデータセットである. 例を Fig. 4 に示す.

Figure 4: アヒルの画像例

Fig. 5 に初期パラメータ, Fig. 6 に最適化後のパラメータで視覚化した結果を示す. Fig. 5 では, クラスタ間のつながりがなく, ユーザにとって解析が難しい結果となっているが, Fig. 6 は各クラスタのノード数が小さく, クラスタ間のつながりも多くユーザにとって解析しやすい視覚化結果となっている. また, Fig. 6 では, 大きく分けて, 左を向いたアヒルと, 右を向いたアヒルで分かれていることがわかる.

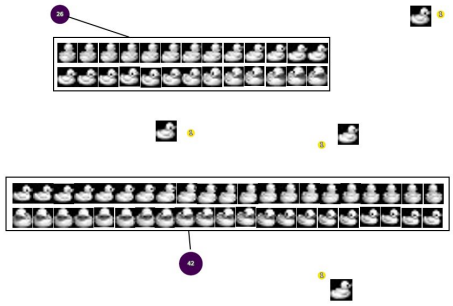

Figure 5: 初期パラメータを用いた視覚化結果 (評価値 $F = -62.2$)

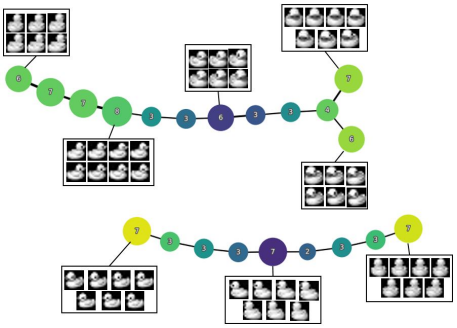

Figure 6: 最適化後パラメータを用いた視覚化結果 (評価値 $F = -8.86$)

5. おわりに

本研究では, Mapperのパラメータを自動で調整する方法を提案した. 自動で調整したパラメータを用いて視覚化した結果は, ユーザにとって解析の手掛かりになるものとなった. 今後は, 時系列データを対象とし実験を行い, クラスタの成長過程から新たな知見を得ることを課題する.

参考文献

1) 鈴木雅彦, and 鈴村嘉右. "データ可視化の必要性と意義: データビジュアライゼーションとは (〈 特集 〉 情報をわかりやすくするデザイン)." 情報の科学と技術 65.11 (2015): 470-475.

2) Lum, P. Y., et al. "Extracting insights from the shape of complex data using topology." Scientific reports 3 (2013): 1236.

3) Singh, Gurjeet, Facundo Mmoli, and Gunnar E. Carlsson. "Topological methods for the analysis of high dimensional data sets and 3d object recognition." SPBG (2007).

4) 伊庭斉志, "C による探索プログラミング-基礎から遺伝的アルゴリズムまで-" オーム社 (2008): pp.204-263