D-002

多次元データ分析のための可視化推薦システム
A Visualization Recommendation System for Multidimensional Data Analytics

野田 昌太郎
Shotaro Noda

杉浦 健人
Kento Sugiura

石川 佳治
Yoshiharu Ishikawa

1 はじめに

多次元データを分析する際に，グラフを作成して視覚的に情報を理解するのは一般的な手法である．しかし，データのサイズや複雑さが増していく中で，グラフを作成して探索的に分析を行うのは困難な作業となっている．多次元データから作成可能なグラフの数はデータの次元の数に対して指数的に増加しており，その膨大な数のグラフから必要なグラフを探索し，多次元データから知見を獲得するのは容易ではない．

ユーザが多次元データのサブセットを探索する方法として，図 4 のようにデータセット全体から分析を始め，1 つずつ次元を追加して分割してサブセット間の違いを探索する方法が考えられる．例として，年，地域，メーカーの 3 次元を持つ売上データで，縦軸を売上額，横軸を商品カテゴリとなるグラフを探索する場合を考える．このとき，ユーザが探索する空間は図 1 の束構造で記述できる．データセットを年について分割すれば，図 2 左のように年ごとのデータのサブセット（2009 年のデータ，2010 年のデータ...）から生成されるグラフが得られ，メーカーごとで分割すれば図 2 右のようにメーカーごとのデータのサブセットから生成されるグラフが得られる．

既存のサブセット探索を支援する研究として Lee らの手法 [1] が挙げられる．この手法では，分割した結果の評価基準として，そのサブセットから作成したグラフと最も近い親のグラフからの距離をベースとした手法を定義している．例えば，2009 年の関東のグラフを評価する場合，親グラフにあたる 2009 年のグラフ，関東のグラフのうち分布の距離が近くなる方を選択し，選んだ最も近い親のグラフとの距離が遠いものを優先的に推薦する．

しかし，同手法ではグラフを親グラフとの関係のみで定量化しており，隣接する他のグラフの状況を考慮していない．例えば，2009 年関東のグラフを評価する場合，親のグラフとの距離が遠いケースでも，2009 年の関西，2009 年の中部など他の地区のグラフの状況によって得られる情報は異なる．他の地区のグラフがほぼ同一ならば，2009 年の関東のグラフが異常値となる一方，他の地区のグラフの分布が多様な場合は地区の次元が重要な情報となる．

そこで本稿では，隣接するグラフの状況を考慮したデータセットの分割操作の定量化手法について検討する．具体的には，分割結果を隣接するグラフごとにグループ化し，各グループに対してスコア付けを行うことで，同操作の定量化を行った．

名古屋大学大学院情報学研究科 Graduate School of Informatics, Nagoya University

2 問題定義

ここでは，分割結果の定量化に向けて，データセットの分割操作と比較可能性，ユーザにとって有用な比較次元を定義する．

2.1 データセットの分割操作の定義

ユーザが以前の操作で得たサブセット集合 $D = \{D_1, D_2, ..., D_k\}$ の各要素を，分割したい次元 $A = \{a_1, a_2, ..., a_l\}$ を用いて分割することとして定義する．その結果，ユーザは新たなサブセット集合 $D_{new} = \{D_{1,a_1}, D_{1,a_2}.D_{1,a_3}, ..., D_{1,a_l}, D_{2,a_1}, ..., D_{k,a_l}\}$ を得る．

2.2 比較可能性の定義

ユーザが比較可能なグラフは，サブセットを構成する条件式が 1 つ異なるグラフ同士と定義する．例えば，2009 年の関東，2009 年の関西は異なる条件が地域の 1 つのため比較可能，2009 年の関東と 2010 年の関西は異なる条件が年と地域の 2 つとなるので比較不可能とした．

2.3 優先する比較次元

多次元データを複数の次元で分割して比較分析する場合，ユーザが比較に用いる次元は分割に使用している次元集合である．データセット全体から 1 つずつ次元を追加して分割して比較分析を行う場合，最も比較したい次元は最後に追加したものと考えられる．なぜなら，ユーザはその他の次元による分割結果は事前に確認しているためである．

たとえば，2009 年，2010 年，2011 年の商品の売上のグラフを表示していたと仮定する．ここで，新しく地域の次元で分割した場合，ユーザは各地域ごとのグラフ (2009 年の東北，2009 年の関東...2009 年の九州) に注目して比較する，ということである．つまり，図 3 左のように結果をそれぞれの年ごとにグループ分けし，各地域のグラフを比較することが最も優先される．

また，東北，関東... 九州の商品の売上グラフを表示していた場合に，新しく年の次元で分割した場合は，ユーザは各年ごとのグラフ (2009 年の東北，2010 年の東北...) に注目する．そのため図 3 右のように結果をそれぞれの地域ごとにグループ分けし，各年のグラフを比較することが最も優先される．

3 分割結果の定量化

分割結果は，ユーザは比較したい次元に沿った違いを発見することを目的とする，という仮定のもと定量化した．その中でも，最後に分割した次元での比較が最も重要なため，結果セットを最後に分割した次元でグループ分けし，グループ内のスコ

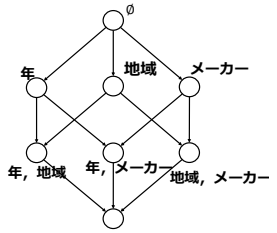

図 1　探索空間となる束構造

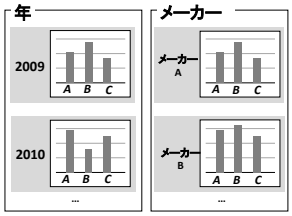

図 2　2 つの分割経路

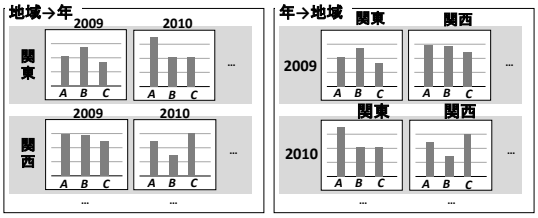

図 3 「年」→「地域」と「地域」→「年」

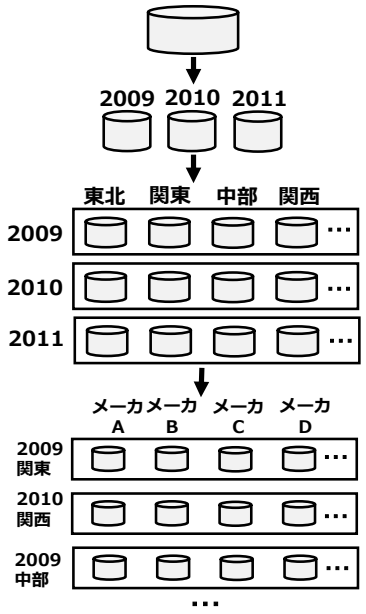

図 4　データセットを分割して比較する流れ

アとグループ間の関係性から分割結果の定量化を図った．

3.1　結果のグループ化

サブセット集合 $D = \{D_1, D_2, ..., D_k\}$ の各要素を，次元 A を用いて分割して得られるサブセット集合 $D_{new} = \{D_{1,a_1}, D_{1,a_2}.D_{1,a_3}, ..., D_{1,a_l}, D_{2,a_1}, ..., D_{k,a_l}\}$ を以下のように，分割前のサブセットごとに分けて管理する．

$$D_{new} = \{\{D_{1,a_1}, D_{1,a_2}.D_{1,a_3}, ..., D_{2,a_l}\}, \{D_{2,a_1}, D_{2,a_2}.D_{2,a_3}, ..., D_{2,a_l}\}, \ldots \{D_{k,a_1}, D_{k,a_2}.D_{k,a_3}, ..., D_{k,a_l}\}\}$$

上の売上データの例で説明すると，年→地域の順で分割したとすると 2009 年の各地域のグラフ，2010 年の各地域のグラフをそれぞれ 1 セットとして管理するということである．

3.2　セット内でのスコア

分布の違いを発見したいという目的から考えると，分割して多くの違いが生まれたセットほど有用なため，各セットの多様性をスコア化する．このスコアとして利用可能な指標は平均距離や平均非類似度など様々なものが考えられる．平均距離を用いて定量化する場合は，各グラフを確率分布に変換した後，以下の式のようにセット内の全ての 2 つのグラフのペアの平均距離を計算する手法がある．

$$Diversity = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} distance(V_i, V_j)$$

3.3　セット間の関係性の考慮

セット間の関係については，まず，考慮するかしないかという選択肢がある．セット間の関係を考えない場合は，単にセット内のスコアが大きいものを優先的に表示していけばよい．

セット間の関係性を考える場合には，結果セット同士を比較可能にするかどうかという観点から分類が可能である．なお，ここでのセット間の比較可能性は，2 つグラフセットの任意の要素に対して，もう一方のグラフセット内に比較可能なグラフが 1 つ存在すること，とした．例えば，図 3 の年→地域の 2009 年と 2010 年のグラフセットでは，年以外の条件が同じグラフが互いに存在しているため比較可能となる．

1 つ目の手法は，任意の 2 つのグラフが比較可能となるように各セットをさらにグループ化するという手法である．例えば，2009 年関東の各メーカーごとの売上グラフセットと，2010 年の関西の各メーカーごとの売上グラフセットを直接比較することはできない．そのため，2009 年の各地区でのメーカーごとの売上グラフセットからスコアの計算を行うこととなる．

2 つ目は，セット間が比較可能でなくとも距離を導出して多様性スコアを導出するという手法である．この手法では，2009 年関東の各メーカーごとの売上グラフセットと，2010 年の関西の各メーカーごとの売上グラフセットは理論的には比較不可能だが，距離を導出して結果セットの多様性のスコアを計算する．

4　まとめと今後の課題

本稿ではデータセットの分割操作を支援するための操作結果の定量化を議論した．今後の課題としては，議論の結果を用いた操作の推薦手法の検討，提案手法の実装が上げられる．

謝辞

本研究の一部は，科研費 16H01722 および 19K21530 による．

参考文献

[1] D. J.-L. Lee, H. Dev, H. Hu, H. Elmeleegy, and A. Parameswaran, “Avoiding drill-down fallacies with Vispilot: Assisted exploration of data subsets,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI ’19, (New York, NY, USA), pp. 186–196, ACM, 2019.