4ZF-02

男女差を俯瞰するための階層型データ可視化: EMDによる男女間の分布差の導入

中井 祐希 †　　伊藤 貴之 ‡
お茶の水女子大学 †　お茶の水女子大学 ‡

1 はじめに

データの偏りは特定の属性を持つ部分集合に見られることが多く，定量的に判断できる問題ばかりではない．そのため，データ中の特定の属性に起因する偏りを発見するには，属性ごとの数値分布の違いを人間が理解する必要がある．この観点から著者らは，多数の人物を対象としたデータから，データの分布の男女差を可視化する手法 [3] を開発している．この手法では，データ中の人物群を属性で階層的に分割し，帯グラフを搭載した階層型データ可視化手法を適用している．本報告ではその拡張手法として，Earth Mover's Distance[5, 6] による男女間の分布差算出を導入することで，注目するべき属性の組み合わせを推薦し，可視化画面において男女差が大きい部分を強調表示する手法を報告する．本報告では空調の温感に関する評価値の可視化例を紹介する．

2 関連研究

Pastor らによる DivExplorer[4] は，モデルが異常な動作をするデータセット中のサブグループを特定し，サブグループとデータセット全体のエラーメトリックの差や個々の属性がサブグループの発散に与えた影響などを可視化する．栃木ら [7] は，機械学習における映画推薦システムデータにおいて鑑賞履歴と推薦結果の差異を表示することで，推薦システムにおける機械学習のバイアスを可視化している．

3 階層型データとしての偏りの可視化

3.1 節，3.2 節の処理は，我々が以前に報告した内容 [3] と同一である．

3.1 データの概要

提案手法では以下のデータを前提とする．A は人物集合によるデータ全体を表し，a_i は i 番目の人物を表し，n はデータ中の人数を表す．

$A = \{a_1, a_2, \ldots, a_n\}$

また，i 番目の人物に相当する a_i は以下の変数を有するものとする．ここで e_i は可視化の対象となる実数値，g_i は i 番目の人物の性別，r_{ij} は j 番目の実数型変数の属性値，c_{ik} は k 番目のカテゴリ型変数の属性値である．

$a_i = \{e_i, g_i, r_i j, \ldots, c_{ik}, \ldots\}$

3.2 木構造の生成

提案手法では，属性値 r_{ij} または c_{ik} のうちユーザが選んだ複数の属性値を用いて，人物群を階層的に分類し，木構造を構成する．

3.3 EMD による分布間距離の算出

EMD[5, 6] は，ある分布をもう一方の分布に移動させるための最小コストとして定義される距離尺度である．本手法では，末端の葉ノード群に相当する人物群のうち，男性と女性の e_i の分布間距離を EMD を用いて算出し，この結果を男女間の分布の非類似度とする．

3.4 「平安京ビュー」を用いた木構造の可視化

「平安京ビュー」[2] は葉ノードを正方形のアイコンで表現したのに対し，提案手法では葉ノード群に相当する人物群が有する e_i の分布を男女別に 2 列の帯グラフで表現する．帯グラフの各領域の色は HSI 表色系を採用し，以下の原則に沿って算出する．

色相 (H): EMD が指定値より低い場合は男女共にグレースケールで，EMD が指定値より高い場合は男性を青，女性を赤にする．

彩度 (S): 平均値に近いほど低く，最大値/最小値に近いほど高くする．

明度 (I): 値が大きいほど高くする．

Hierarchical data visualization of Gender Difference: Introducing Distributional Differences between Men and Women by EMD

† Yuki Nakai, Ochanomizu University
‡ Takayuki Itoh, Ochanomizu University

4　選択する属性の組み合わせの推薦

可視化システムを操作する際に選択するに値する属性の組み合わせをユーザに推薦する．まず，データ中の人物群が有する属性の全ての組み合わせに対して，人物群を階層的に分類して木構造を生成する．生成した木構造の全ての末端の葉ノード群に相当する人物群について，男性と女性の数値分布の分布間距離を EMD を用いて算出する．算出した EMD の合計値が高いものから順に，木構造を生成する際に用いた属性の組み合わせをユーザに提示する．

5　空調温感データでの適用事例

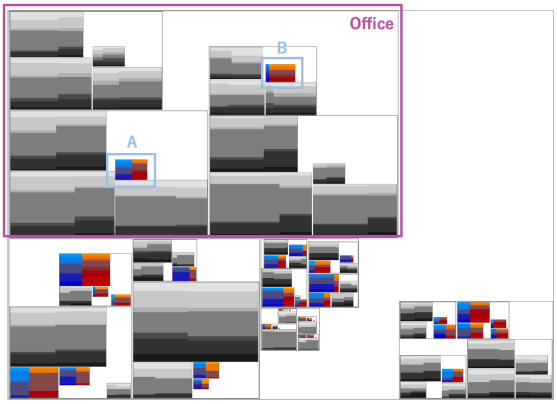

図 1　Strategy, Season, Building の順に人物を分類した例

本報告では空調の温感に関するオープンデータ [1] を適用した事例を示す．このデータから著者らは 32,373 人を対象として以下の属性値を抽出した．

TS　温感に対する 7 段階の評価値.0 がちょうどよい，正値が暑い，負値が寒い．
Sex　生物学的な意味での性別．
Age　年齢
Cloth　服装の厚さの実数値．大きいほど厚い．
Metab　代謝量に関する実数値．
Season　春/夏/秋/冬のカテゴリ値．
Building　オフィス/教室/住居/高齢者施設/その他のカテゴリ値．
Strategy　エアコン/換気/混合のカテゴリ値．

推薦順位が最も高い Age，Cloth，Building の組み合わせを用いて可視化する．図 1 は Building，Cloth，Age の順に属性値を参照して人物を分類した可視化結果である．ここで，左上のオフィスの枠に注目すると，A と B の部分のみ帯グラフがカラースケールで表示されている．A は，服装の厚さが 2 番目に厚く年齢が最も若い人物群の帯グラフであり，男性よりも女性の方が寒いと感じる人が多いことがわかる．B は，服装が最も薄着で年齢が最も若い人物群の帯グラフであり，女性より男性の方が寒いと感じる人が多いことがわかる．以上により，建物がオフィスの場合は，服装の厚さが 2 番目に厚いまたは最も薄着で年齢が最も若いという局所的な部分集合で，温感に男女差が発生していることがわかる．しかし，A と B の部分では男女の分布に偏りがあることは共通しているが，男女間でどのような偏りが生じているかは異なる．このことからも，偏りの判断を計算機に任せるのではなく，人間の解釈を交えることが望ましいと言える．

6　まとめ・今後の課題

本報告では，多数の人物を対象としたデータ中に潜む男女差の可視化手法の拡張として，EMD による男女間の分布差算出を導入した手法を提案した．空調の温感データを題材として可視化結果を示し，その有効性について議論した．

今後の課題として，属性を選択する順番を推薦するシステムの構築や，空調の温感以外の多様なデータへの適用があげられる．

参考文献

[1] Ashrae global thermal comfort database ii. `https://www.kaggle.com/datasets/claytonmiller/ashrae-global-thermal-comfort-database-ii`.
[2] Takayuki Itoh, Yumi Yamaguchi, Yuko Ikehata, and Yasumasa Kajinaga. Hierarchical data visualization using a fast rectangle-packing algorithm. *IEEE Transactions on Visualization and Computer Graphics*, 10(3):302–313, 2004.
[3] Yuki Nakai, Takayuki Itoh, Hidekazu Takahashi, Satoshi Nakashima, and Tetsu Yamamoto. Hierarchical data visualization of gender difference: Application to feeling of temperature. In *27th International Conference on Information Visualisation (IV2023)*, pages 178–183. IEEE, 2023.
[4] Eliana Pastor, Andrew Gavgavian, Elena Baralis, and Luca de Alfaro. How divergent is your data? *Proceedings of the VLDB Endowment*, 14(12):2835–2838, 2021.
[5] Ofir Pele and Michael Werman. A linear time histogram metric for improved sift matching. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, pages 495–508. Springer, 2008.
[6] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.
[7] Ami Tochigi, Takayuki Itoh, and Xiting Wang. Visualization of bias of machine learning for content recommendation.