男女差を俯瞰するための階層型データ可視化

理学専攻　情報科学コース　2240660　中井 祐希　(指導教員：伊藤 貴之)

1　はじめに

近年，公平性の高い社会や組織を構築することは非常に重要な課題となっており，特定の属性に起因する差別や不平等が問題視されている．ここでいう属性とは例えば，性別，地域，人種，世代などである．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．そこで本論文では，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を提案する．この手法では，データ中の人物群を属性で階層的に分類し，帯グラフを搭載した階層型データ可視化手法を適用する．その際に Earth Mover's Distance[4, 5] による男女間の分布差算出により，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．本論文では，空調の温感に関する評価値の男女差を可視化した事例を報告し，本手法の有効性を議論する．

2　関連研究

Pastor らによる DivExplorer[3] 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 [6] 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．

従来研究での可視化表現は単純な棒グラフ・折れ線グラフ・散布図を中心とした複数のビューで構成されており，ユーザは反復的な操作を繰り返すことでデータのバイアスを探す必要があった．それに対して本研究は，データを俯瞰する最初の一画面でデータ中の特定の部分に潜む偏りを発見させることに重点を置いており，この点において本研究は従来研究と大きく異なる．

3　階層型データとしての偏りの可視化

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．

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

3.2　木構造の生成

属性値 r_{ij} または c_{ik} のうちユーザが選択した複数の属性値を用いてデータを構成する人物を階層的に分類し，木構造を構成する．属性値がカテゴリ型変数の属性値の場合は人物群をカテゴリごとに分類する．属性値が実数型変数の属性値の場合は,「(実数値 x_1) 未満」,「(実数値 x_1) 以上 (実数値 x_2) 以下」，…,「(実数値 x_N) 以上」というように段階（ここで N は段階数を表す）を設定し，人物群を段階ごとに分類する．本手法では属性を 1 つずつ選び，その属性にもとづいて人物を分類し... という処理を何度か反復することで木構造を生成する．生成した木構造の特定のノード配下に可視化の対象値である実数値 e_i の偏りが見られるようであれば，その偏りはユーザが選択した属性値に起因する偏りであることが示唆される．現段階の実装では選択する属性の個数を 3 に固定している．

3.3　EMD による男女の分布間距離の算出

Earth Mover's Distance(EMD)[4][5] 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．EMD の値が大きいほど男女の分布が異なっており，反対に EMD の値が小さいほど男女の分布が類似していることが示唆される．

3.4　「平安京ビュー」を用いた可視化

生成した木構造を「平安京ビュー」[2] によって可視化する．図 1 にその構造を示す．木構造を「平安京ビュー」で可視化した結果において，1 番外側の領域が 1 個目に選んだ属性，その内側の領域が 2 個目に選んだ属性，さらにその内側が 3 個目に選んだ属性と対応する．

「平安京ビュー」では葉ノードを正方形のアイコンで表現したのに対し，本手法では末端の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．具体的には，データを構成する人物群が持つ実数値 e_i をいくつかの階級に分割する．そして，木構造の末端の葉ノード群に相当する人物群に対して各階級に該当する人数を集計し，その集計結果を帯グラフで表示する．現段階の我々の実装では階級数を 7 に固定している．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 の値が指定値より低い場合は男女ともにグレースケールで，EMD の値が指定値より高い場合は男性の色相を青，女性の色相を赤にする．

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．

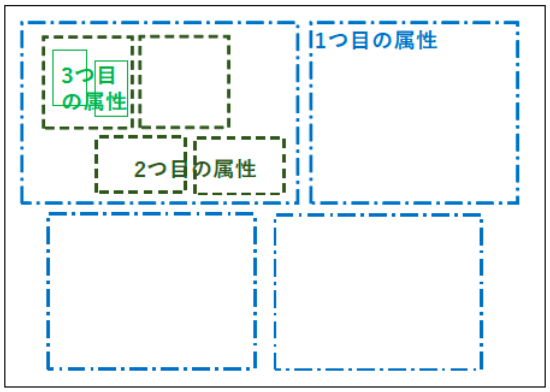

図 1: 「平安京ビュー」による木構造の可視化．

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

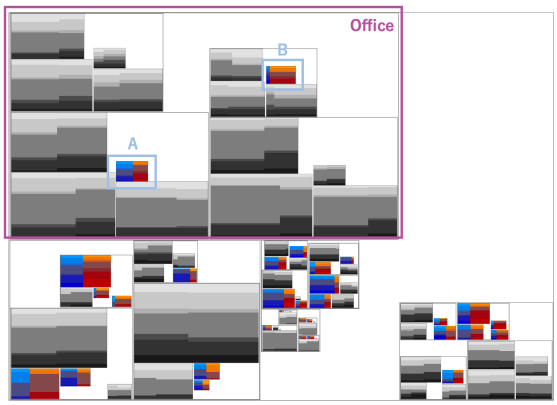

図 2: Strategy, Season, Building の順に人物を分類した例

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 2 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．以上のことから，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

本論文では，多数の人物を対象としたデータ中に潜むデータの分布の男女差の可視化手法を提案した．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] 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.

[4] 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.

[5] 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.

[6] Ami Tochigi, Takayuki Itoh, and Xiting Wang. Visualization of bias of machine learning for content recommendation.