つぶつぶ表現：カテゴリデータ分析のための視覚的表現手法

白石　宏亮†　　三末　和男†　　田中　二郎†
†筑波大学大学院　システム情報工学研究科　コンピュータサイエンス専攻

1　はじめに

カテゴリデータ（質的データ）とはデータ中の変数が数値ではなく，カテゴリによって区別されるデータである．例えば，アンケートにおける項目（属性）の性別，血液型，職業などはカテゴリによって区別される変数である．カテゴリデータの分析はマーケティングリサーチなど多くの場所で使用されるが，従来方法では，数値で埋め尽くされた表中での作業となり，時間と労力がかかってしまう．

本論文ではカテゴリーデータを直感的に分析する視覚的表現「つぶつぶ表現」について述べる．つぶつぶ表現ではデータのオブジェクトを一つ一つ視覚的に表示し，それらをインタラクティブに操作することで，カテゴリデータの直感的な分析が可能である．

2　カテゴリデータの分析

一般的なカテゴリデータ分析のプロセスについて述べる（図 1）．カテゴリデータの生データはリスト形式のデータである．この生データから着目するいくつかの属性のクロス集計表を作成する．そして，クロス集計表を元にグラフ化を行う．グラフから得られた知見を元に，クロス集計表中の別の部分をグラフ化，または新たなクロス集計表を作成する．このように，クロス集計表作成とグラフ化を繰り返す作業により分析を行っていく．このようなプロセスにおいて，クロス集計表とグラフの要素との対応の把握や，複数のグラフを見比べるといった作業は時間と労力がかかる．

上記の問題を解決し，より多くの情報を視覚的に表現する手法が研究されている．Mosaic Displays[1] では長方形の面積によって度数を表現し，タイル状に並べることで複数の属性の関係を一枚の図中で表現している．Parallel Sets[2] は parallel coordinates を複合した表現でインタラクティブなカテゴリデータ分析を行える．Cattrees[3] では Treemap を用いて表現している．Mosaic Displays と同様に長方形の面積によって度数を表し，階層式に敷き詰めることで表現している．これらの表現の空間効率は良いが，表現方法が特殊なものが多く，直感的な分析が行えないといった問題がある．

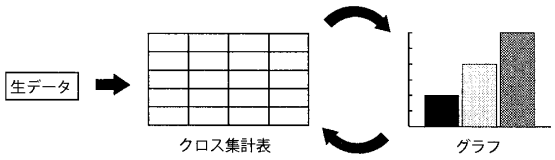

図 1: カテゴリデータ分析のプロセス

3　つぶつぶ表現

つぶつぶ表現とは本論文で述べるカテゴリデータを視覚的に分析するための表現手法である．ここでは例として，図 2 左のような 2 つの属性「性別」,「意見」から成るクロス集計表を考える．つぶつぶ表現では表中の各セル値の度数，すなわちリスト形式のデータにおけるレコード一つ一つを視覚的な要素として表示する（図 2 右）．つぶつぶ表現における一つのつぶを要素と呼ぶ．このように一つ一つ視覚的に表現することで，データをオブジェクトの集合のように直感的にイメージすることができる．

	男性	女性
好き	5	7
普通	6	3
嫌い	3	5

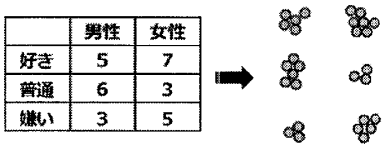

図 2: つぶつぶ表現

3.1　カテゴリの表現

つぶつぶ表現では視覚的に表示された要素の配置によって，その要素の持つ属性のカテゴリを表現する．ゲシュタルトの近接の要因により，人間は位置的に近接している要素を同一の関係であると知覚する．例えば，図 3 では要素は 2 つのグループに分けられていると知覚する．さらに，ラベルとの位置で，左の要素群は男性，右の要素群は女性であると知覚することができる．

3.2　要素のカテゴリ分け

前項では配置によるカテゴリの表現について述べたが，ここでは要素の配置，すなわちカテゴリ分けの方

Granular Representation : A Visual Representation Technique for Analyzing Categorical Data
†Kousuke Shiraishi　†Kazuo Misue　†Jiro Tanaka
†Department of Computer Science, University of Tsukuba

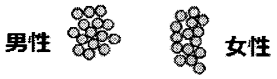

図 3: カテゴリの表現

法ついて述べる．要素のカテゴリ分けはラベルによる操作，またはクラスタによる操作を用いる．ラベルによる操作ではカテゴリのラベルをドラッグすることで，そのカテゴリを持つ要素が引き寄せられる．例えば，図 4(a) では，一つにまとまった要素群の中から「男性」のラベルをドラッグすることで，「男性」のカテゴリを持つ要素が引き寄せられる．

クラスタによる操作では，選択した属性による同一カテゴリを持つ要素同士で近接したまとまりを形成する．例えば，図 2 左のクロス集計表において属性「性別」と「意見」によってクラスタ化を行うと，図 4(b) のように 6 つの要素群が形成される．つまり，2 つの属性によってこれ以上分けられないカテゴリの要素群に分かれる．属性のカテゴリが多い場合，ラベルによる操作では手間がかかる場合があるが，クラスタを用いることでデータの大局的な傾向を概観することができる．

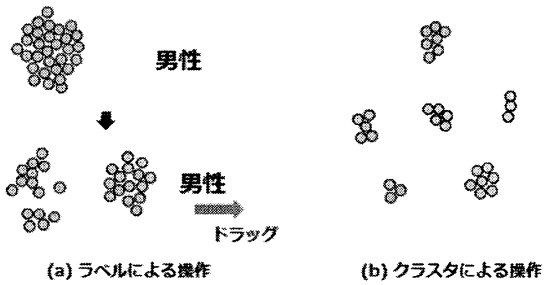

図 4: 要素のカテゴリ分け

4 ツールの開発

つぶつぶ表現はデータを視覚的に表現することで，直感的なカテゴリ分けが可能であるが，割合比較には従来のグラフが優れていると思われる．我々はつぶつぶ表現と棒グラフを統合することで，それぞれの表現を補い合うカテゴリデータの分析ツールの実装を行った．図 5 にツールの概観を示す．本ツールは 2 つの画面から構成される．左画面は設定やグラフ表示を行う画面であり，右画面はつぶつぶ表現を用いてデータが表示される画面である．ユーザは右画面で主に操作し，要

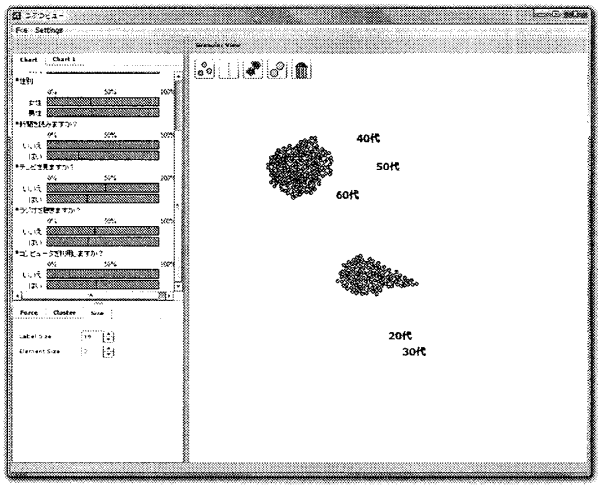

図 5: ツール概観

素のカテゴリ分けを行う．そして，要素を選択することで，その要素の持つ属性のカテゴリが集計され，左画面にグラフで表示される．右画面で要素を細かいカテゴリに分けていくことで，局所的な分析を行うことができる．

5 まとめ

本研究ではカテゴリデータ分析のための視覚的手法「つぶつぶ表現」を開発し，提案手法を用いたカテゴリデータ分析のためのツールを実装した．

つぶつぶ表現はカテゴリデータを視覚的に表現することで，ドリルダウンを直感的に行うことが可能である．また，一つ一つ要素を参照することができるので，個々の観点からの局所的な分析が可能である．

参考文献

[1] Michael Friendly. Mosaic displays for multi-way contingency tables. *American Statistical Association*, Vol. 89, No. 425, pp. 190–200, 1994.

[2] Fabian Bendix, Robert Kosara, and Helwig Hauser. Parallel sets: Visual analysis of categorical data. In *Proceedings of the IEEE Symposium on Information Visualization 2005 (INFOVIS'05)*, pp. 133–140, 2005.

[3] Erica Kolatchm and Beth Weinstein. Cattrees: Dynamic visualization of categorical data using treemaps. (`http://www.cs.umd.edu/class/spring2001/cmsc838b/project/kolatch_weinstein/index.html`), 2001.