Otto-von-Guericke University Magdeburg

Department of Computer Science
Institute of Simulation and Graphics

Master Thesis Proposal

Working Title: Semi-automatic pattern detection in hierarchical tabular data with weak hierarchies

Author:
Jalaj Arjav, Vora

Version 1
November 15, 2022

Advisers:

Supervisor
Prof. Dr.-Ing. Bernhard Preim

Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Supervisor
M. Sc. Benedikt Mayer

Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Contents

1 Motivation

Data is becoming extremely important and especially in the field of narrative visualization. Narrative visualization helps users understand and comprehend complex data easily. Such data-driven storytelling has become important component in many fields including medical data visualization [9, 4] and journalism [11]. This inspires many researchers to create descriptive taxonomies of such form of data as quantitative analysis, classifying existing techniques used to convey story or articles from the field of narrative visualization. One of the examples include [11], where authors create a taxonomy exploring 20 data-driven story telling techniques broadly categorized into 4 techniques for 45 asynchronous data stories. Such descriptive taxonomies have hierarchical structures defining relationships to their super categories. Therefore, it would be great if we had a way to automatically find patterns in such hierarchical tabular data. However, often automatic algorithms depend on certain input parameters. Therefore, this inspires to integrate the execution of the automatic algorithms into an interactive application that allows to customize input parameters and visualize the output in more insightful and effective way [5].

2 Problem definition and Research Question

Related work either focuses in the direction of building visual analytics systems using algorithms such as t-SNE, MDS and PCA as dimentionality reduction technique combining with machine learning algorithms or focuses more in building visual analytics system for diverse datasets such as secRNA analysis dataset, eye-tracking dataset or genomics data. Therefore, it is very interesting to explore automatic pattern detection for descriptive taxonomies with weak or flat hierarchies using algorithm like t-SNE and MDS through building web-based interactive visual analytics tool.

3 Related Work

3.1 Interactive visual exploration and analysis of tabular data

Guozheng Li et al. have developed an interactive visual analysis tool for hierarchical tabular data which constructs an abstract model which defines rowcolumn headings as bi-clustering and hierarchical structures [7, 15]. Whereas, Eckelt et. al. proposes TourDino as a view upon Ordino, providing ~~easy~~ exploration and validation of statistical hypothesis using interactive visualisation on tabluar data [2, 12]. Interestingly, Lex. et. al proposes a visual analytics tool for large scale heterogeneous genomics data, where they find patterns in the data by focusing on exploring relationships among columns with different data-types [6]. Furmanova et. al. provides scalable visualisation of tabular data, providing interactive analysis through hierarchical aggregation of subsets [3].

3.2 Automatic pattern detection in multi-variate tabular data

Related work shows a lot of research in the direction of using Dimensionality Reduction techniques for high dimensional data combing with supervised learning. Chad A Steed et al. creates a visual exploration system for multivariate data with heterogenous type which helps understanding the input to algorithms such as neural network [10]. Xueli Xu et al. used t-SNE algorithm with aitchson distance as dimentionality reduction and fed the low dimensional features to commonly used machine learning algortihms for compositional microbiome data [13]. Whereas, Ju Nam et al. describes the importance of cluster analysis using an interactive tool to control cluster parameters on high dimensional aerosol mass spectra data [8]. Zhang et al. diagnoses errors and patterns in ~~the~~ machine maintainence data by visual analytics tool using dimensionality reduction technique and clustering [14]. Zhou et al address the spatial clusters of air-quality data using visual analytics tool and exemplifies factors responsible for the air-quality using MDS and Heirarchichal clustering [16]. Devassy et al. shows t-SNE outperforming PCA for hyperspectral ink data [1].

Based on these works, I aim to develop visual analytics system for heirarchical tabular taxonomy data using t-SNE and MDS as potential algorithms.

4 Methodology

This thesis aims to do a literature review on existing algorithms used to find patterns in hierarchical tabular data. After the review, thesis will try to work on implementing suitable algorithms from the literature and would build an interactive R-Shiny application. This thesis would also ~~work on reviewing the~~ literature for visualisation ~~used to justify detected patterns and explaining~~ design decisions. To evaluate, a ground truth with known patterns will be validated against the developed application. For usability validation on actual data, the application will be then checked by domain experts from the Chair of Visualization at Otto-von-Guericke, University.

5 Goals and benefits

Based on descriptive taxonomy generated by Stolper et. al. as an example, this thesis will try to create a visual analytics application to semi-automatically find patterns in the hierarchical data. It will also try to explore how this weak relationship ~~affects in detecting patterns.~~ Therefore, I aim to explore how could patterns be found in hierarchical tabular data and what would weak hierarchical relationship of the columns in such forms of data signify. This shall serve as an guidance and helpful tool to for future researchers to effectively find patterns and analyze in hierarchical tabular data and serve as basis for future research in visualisation of patterns with certain techniques in detail.

References

[1] Binu Melit Devassy and Sony George. "Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE". In: *Forensic science international* 311 (2020), p. 110194.

[2] Klaus Eckelt et al. "TourDino: A Support View for Confirming Patterns in Tabular Data." In: *EuroVA@ EuroVis*. 2019, pp. 7–11.

[3] Katarina Furmanova et al. "Taggle: Scalable visualization of tabular data through aggregation". In: *arXiv preprint arXiv:1712.05944* 6 (2017).

[4] Katarína Furmanová et al. "VAPOR: Visual Analytics for the Exploration of Pelvic Organ Variability in Radiotherapy". In: *Computers & Graphics* 91 (2020), pp. 25–38. ISSN: 0097-8493. DOI: https://doi.org/10.1016/j.cag.2020.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0097849320300960.

[5] Daniel Keim et al. "Visual analytics: Definition, process, and challenges". In: *Information visualization*. Springer, 2008, pp. 154–175.

[6] Alexander Lex et al. "StratomeX: visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization". In: *Computer graphics forum*. Vol. 31. 3pt3. Wiley Online Library. 2012, pp. 1175–1184.

[7] Guozheng Li et al. *HiTailor: Interactive Transformation and Visualization for Hierarchical Tabular Data*. 2022. DOI: 10.48550/ARXIV.2208.05821. URL: https://arxiv.org/abs/2208.05821.

[8] Eun Ju Nam et al. "Clustersculptor: A visual analytics tool for high-dimensional data". In: *2007 IEEE Symposium on Visual Analytics Science and Technology*. IEEE. 2007, pp. 75–82.

[9] Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. "Explaining artificial intelligence with visual analytics in healthcare". In: *WIREs Data Mining and Knowledge Discovery* 12.1 (2022), e1427. DOI: https://doi.org/10.1002/widm.1427. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1427. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1427.

[10] Chad A Steed et al. "CrossVis: A visual analytics system for exploring heterogeneous multivariate data with applications to materials and climate sciences". In: *Graphics and Visual Computing* 3 (2020), p. 200013.

[11] Charles D Stolper et al. "Emerging and recurring data-driven storytelling techniques: Analysis of a curated collection of recent stories". In: (2016).

[12] Marc Streit et al. "Ordino: a visual cancer analysis tool for ranking and exploring genes, cell lines and tissue samples". In: *Bioinformatics* 35.17 (2019), pp. 3140–3142.

[13] Xueli Xu et al. "A t-SNE Based Classification Approach to Compositional Microbiome Data". In: *Frontiers in Genetics* 11 (2020). ISSN: 1664-8021. DOI: 10.3389/fgene.2020.620143. URL: https://www.frontiersin.org/articles/10.3389/fgene.2020.620143.

[14] Xiaoyu Zhang et al. “A visual analytics approach for the diagnosis of heterogeneous and multidimensional machine maintenance data”. In: *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. IEEE. 2021, pp. 196–205.

[15] Jian Zhao et al. “Bidots: Visual exploration of weighted biclusters”. In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 195–204.

[16] Zhiguang Zhou et al. “Visual analytics for spatial clusters of air-quality data”. In: *IEEE computer graphics and applications* 37.5 (2017), pp. 98–105.