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 flat hierarchies

Author:
Jalaj Arjav, Vora

Version 2
November 24, 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 increasingly important especially, the use of tabular data in the field of visual analytics. Tabular data is highly preferred and is a critical data management approach as it has been used by many application domain experts such as scientists, financial practitioners and policy-makers [1, 3]. Most of the existing studies focus on flat simple tabular data rather than hierarchical tabular data [6, 11]. Hierarchical tabular data are used, especially in statistical reports and research papers providing better capability of efficient data management [4]. Many research work focuses on diverse types of tabular data such genomic data [9], secRNA data [8], biome data [16], to find patterns. One of the example where tabular data is generated is visual storytelling taxonomies or narrative visualisation taxonomies. Such taxonomies provide a quantitative analysis of storytelling. Stolper et al. [13] create a taxonomy of journalism stories exploring 20 data-driven story telling techniques broadly categorized into 4 main categories creating a hierarchy; against 45 asynchronous data stories. Although authors visualize which techniques where used but do not highlight how often were these different techniques used.

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.

2 Problem definition and Research Question

Considering above mentioned motivation, this work focuses on:

1. identifying types of various patterns theoretically existing in tabular data
2. investigating effects of flat hierarchical structures in pattern detection
3. examining algorithms to analyse and extract patterns in flat hierarchical tabular data
4. designing and implementing an application to analyse tabular data using the most effective algorithms.

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 constructing a model which defines row/column headings as bi-clustering and hierarchical structures to explore relationships among the

hierarchical row and column labels interactively and effectively [7]. Whereas, Eckelt et. al. proposes TourDino integrated in the Ordino [14], a drug discovery platform for the purpose of identifying new drug targets. TourDino provides a supporting view that helps users, who are not experts in statistics, to verify generated hypotheses and confirm insights through exploration and validation of statistical hypotheses using interactive visualisation on tabular data [5]. Interestingly, Furmanova et. al. provides scalable visualisation of tabular data, providing interactive analysis through hierarchical aggregation of subsets [6].

3.2 Automatic pattern detection in multi-variate tabular data

Related work shows research in the direction of using dimensionality reduction techniques for high dimensional data to reduce it into lower dimensions combined either with supervised learning tasks such as classification or unsupervised learning tasks such as cluster analysis. Steed et al. creates a visual exploration system for multivariate data with heterogeneous type which helps understand inputs to algorithm such as neural network [12]. Xueli Xu et al. used t-SNE algorithm as dimentionality reduction and fed the resulting low dimensional features to commonly used machine learning algorithms for compositional microbiome data [15]. On the other hand, Ju Nam et al. describes the importance of cluster analysis using an interactive tool to control cluster parameters on high dimensional aerosol data [10]. Zhang et al. diagnose errors and find patterns in machine maintenance log data through machine learning assisted visual analytics. Here, authors use data-type dependent dimensionality reduction technique, such as use of contrasting clusters in Principal Component Analysis (ccPCA) for numerical data, contrasting clusters in Multiple Correspondence (ccMCA) for categorical data and Uniform manifold approximation and projection (UMPA) for text data in combination with clustering [17]. Zhou et al. address the spatial clusters of air-quality data using visual analytics tool and exemplify factors responsible for the air-quality using MDS and Hierarchical clustering [18]. Devassy et al. show t-SNE outperforming PCA for forensic document analysis done on hyperspectral imaging data [2].

4 Methodology and Evaluation

This work aims to explore various patterns which can theoretically be extracted from hierarchical tabular data. It also intends to do a literature review on existing algorithms used to find patterns in hierarchical tabular data. After the review, suitable algorithms from the literature will be implemented and an interactive R-Shiny based application will be built. This work will review the literature for visualisation used for detection of patterns. To evaluate, a ground truth with known patterns will be validated against the developed application, validating the usefulness of found patterns in hierarchical tabular data. For usability validation on actual data, the application will be checked by domain

experts from the Chair of Visualization at Otto-von-Guericke University, Magdeburg.

5 Goals and benefits

Descriptive taxonomy generated by Stolper et. al. as the baseline, this work will examine types of patterns created on hierarchical tabular data and create a visual analytics application to semi-automatically find patterns in the hierarchical data. The ways in which these flat relationship can be considered in detecting patterns will be explored. Finding patterns in hierarchical tabular data as well as determining significance of flat hierarchical relationship within the columns will be explored. A tool to effectively find patterns and analyse hierarchical tabular data will be developed for the research community, providing a base for further investigation.

References

[1] Zhe Chen and Michael Cafarella. "Automatic web spreadsheet data extraction". In: *Proceedings of the 3rd International Workshop on Semantic Search over the Web*. 2013, pp. 1–8.

[2] 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.

[3] Wensheng Dou et al. "Expandable group identification in spreadsheets". In: *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 2018, pp. 498–508.

[4] Lun Du et al. "TabularNet: A neural network architecture for understanding semantic structures of tabular data". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 322–331.

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

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

[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] George C Linderman et al. "Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data". In: *Nature methods* 16.3 (2019), pp. 243–245.

[9] Uditha Maduranga et al. “Dimensionality Reduction for Cluster Identification in Metagenomics using Autoencoders”. In: *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. 2020, pp. 113–118. DOI: 10.1109/ICTer51097.2020.9325447.

[10] 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.

[11] Charles Perin, Pierre Dragicevic, and Jean-Daniel Fekete. “Revisiting bertin matrices: New interactions for crafting tabular visualizations”. In: *IEEE transactions on visualization and computer graphics* 20.12 (2014), pp. 2082–2091.

[12] 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.

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

[14] 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.

[15] 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.

[16] Xueli Xu et al. “A t-SNE based classification approach to compositional microbiome data”. In: *Frontiers in Genetics* 11 (2020), p. 620143.

[17] 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.

[18] 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.