Exploring Big Data Landscapes with a Glyph-based Zoomable User Interface

Dietrich Kammer[1], Mandy Keck[1], Thomas Gründer[1]

Chair of Media Design, Technische Universität Dresden[1]

{firstname.lastname}@tu-dresden.de

Abstract

High-dimensional data sets are hard to explore using common spreadsheet environments. However, data scientists need to develop appropriate clustering and classification algorithms to make sense of big data repositories. Even sophisticated analysis tools often focus on mathematical tasks and offer only basic data visualization with few interactive features. In order to gain more sophisticated insights and test hypotheses with regards to high-dimensional data sets, we developed an interactive zoomable user interface using glyph-based visualizations. The visualization is based on a two-dimensional plot of the data space using multi-dimensional reduction. The resulting Big Data Landscapes are then explored with various controls, filters, and details on demand.

1 Introduction

Data scientists are challenged with the task of creating appropriate clustering and classification algorithms for big data repositories. This is an important task in order to gain insights into data and to offer products and services to end users. Standard spreadsheet environments are not able to provide the required interactive access to the data space. More sophisticated analysis tools and platforms such as Zeppelin[1] or Jupyter[2] are focused on calculus and mathematical tasks. Although the notebooks used by these tools can contain visualizations, interactive and immersive views into the data are limited. To this end, we adopt the metaphor of a Big Data Landscape (cp. Kammer et al., 2018). In this paper, we describe how common data scientist tasks can be solved using our prototype. We also describe concrete usage scenarios.

[1] https://zeppelin.apache.org/, Retrieved on: 20.06.2018
[2] http://jupyter.org/, Retrieved on: 20.06.2018

Veröffentlicht durch die Gesellschaft für Informatik e. V. 2018 in
R. Dachselt, G. Weber (Hrsg.):
Mensch und Computer 2018 – Workshopband, 02.–05. September 2018, Dresden.
Copyright (C) 2018 bei den Autoren. https://doi.org/10.18420/muc2018-demo-0497

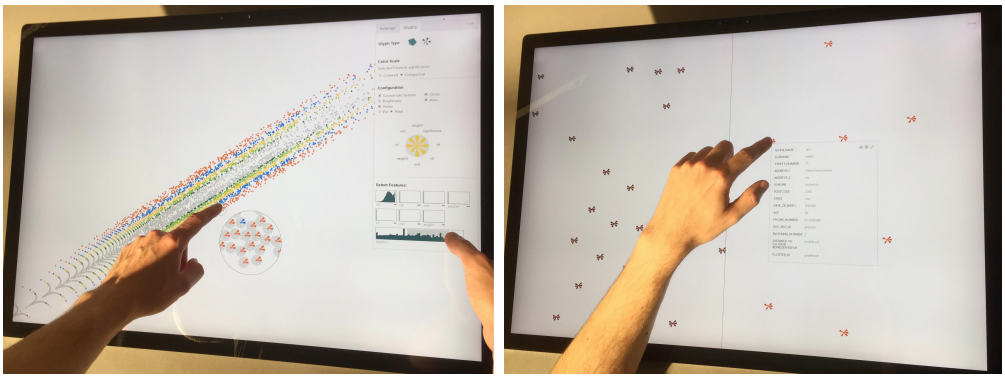

Figure 1: Glyph-based Zoomable User Interface used to analyse clusters using magic lens (left) and splitscreen (right)

2 Related Work

Recently, several systems have been proposed to assist data scientists with their intricate task of optimizing clustering and classification in big data repositories. Heimer et al. focus classifying textual corpora and offer a tool that should be integrated into larger systems (Heimerl et al., 2012). In contrast, our approach is to establish an integrated, immersive view on a data set with a comprehensive set of controls and tools to gain insights into data. Seek-a-view achieves an overview about relevant dimensions in high-dimensional data sets (Krause et al., 2016). However, this system also does not provide a comprehensive overview over the Big Data Landscape. Particular features such as the use of histograms are incorporated in our prototype. Similarly, Choo et al. present an interactive testbed system for dimension reduction and clustering (Choo et al., 2013) without an immersive experience.

3 Glyph-based Zoomable User Interface

Our interactive prototype is focused on a *Big Data Landscape* with minimal additional windows and views as in common dashboard solutions. In contrast, a comprehensive view is explored using zooming and panning as main interaction techniques. A semantic zoom changes the representation of data items from dots to glyphs. Glyphs are used for the detailed inspection of data items, containing a selected set of important features to quickly compare items. Moreover, a tooltip shows concrete data values for an even more detailed analysis.

3.1 Glyph Visualization

Our interface concept is based on the mapping of each data item to a color-coded pixel in a scatter plot that is computed using dimension reduction. The resulting clusters in this Big Data Landscape are first explored by a data analyst, who then selects subsets or small clusters for

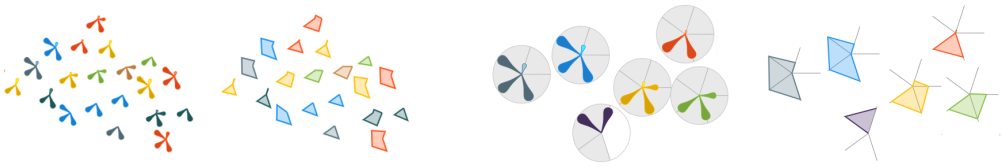

Figure 2: Glyph variations with different zoom levels and flower and star glyphs

detailed inspection. Once the amount of data items is reduced by zooming, another level of detail in the form of glyphs is presented for an in-depth analysis (see 2, left). Glyphs are small independent visual objects that map each data attribute to a graphical attribute, such as size, shape, color, and orientation (Borgo et al., 2013). The major strength of glyphs is that patterns involving more than three dimensions can be more readily perceived and subsets of dimensions can form composite visual features that are easy to recognize. The analyst can choose between star plots and flower glyphs, which have different advantages depending on the data sets (Keck et al., 2017). Moreover, we use more detailed glyphs with a coordinate system at the highest zoom level to observe even small differences between feature values at a glance (see 2, right). The coordinate system also allows the identification of NULL values by hiding the respective axis in case of a missing feature. We propose to use this visualization technique with different levels of detail to analyze the features involved in clustering and classification algorithms.

3.2 Exploration Tasks

Wenskovitch et al. propose distinct tasks for data scientists, which can be facilitated using interactive visualizations (Wenskovitch et al., 2018). In the following, we describe how these tasks can be completed using the features of our prototype:

- **Explore alternate projections**. Each data set in the prototype can be supplied with several dimensionality reduction versions that can be either compared successively or side-by-side in a splitscreen
- **Investigate clusters and features**. The Big Data Landscape intuitively shows clusters of similar data items based on the dimensionality reduction, which can be explored in detail by either zooming into clusters or using a magic lens (see 1)
- **Individual observations**. Tooltips are shown that contain concrete feature values when hovering items either displayed as dots or as glyphs.
- **Determining parameters for algorithms**. Our prototype uses a JSON exchange format for data items with feature values of different calculations. Hence, different calculated versions can be compared. Moreover, we implemented histograms for all features that show the variance with regard to the whole data. These histograms can also be used to filter items in the Big Data Landscape.

- **Selection of pipelines for dimension reduction and clustering**. Also different calculation pipelines can be compared by iterating through the data set versions or showing them side-by-side using the splitscreen (see 1, right).

3.3 Usage Scenarios

In this section, we cover briefly the data sets from very different domains that can be explored using our prototype.

3.3.1 Web Scraping

In order to provide services such as content observation on the web, data scientists need to create models to classify relevant documents using text mining techniques. Information from many heterogeneous resources such as social media or news websites must be downloaded first. By using semi-supervised machine-learning algorithms, appropriate models can be generated. In order to tune the algorithms, data clusters can be explored using our prototype.

3.3.2 Medicinal Data

When analyzing medicinal data, e.g. from tumor patients, clinicians need to investigate hypotheses with regards to the involved patients and their data. For instance, prior diseases, hereditary factors or lifestyle data can be compared to optimize treatments and provide early diagnosis. Moreover, gene sequencing of tumor or healthy cells can lead to new insights into biomedical processes in order to create more suitable treatments. Our prototype can be used to display both patient and gene data with the corresponding features, in order to detect similar and relevant data items.

3.3.3 Product Catalogues

Companies that provide exchange platforms for products and services need to match customer searches to products offered by suppliers in their catalogues. Typically, this poses a large data quality problem, since several suppliers provide equal products with different descriptions and titles. Hence, such aggregated product catalogues often contain large sets of duplicates. Using machine-learning, sophisticated duplicate matching algorithms can be designed based on interactive record linkage. Such approaches can be assessed by exploring clusters of duplicates in our prototype.

4 Conclusions and Future Work

The presented interactive prototype implements a zoomable user interface using glyph-based visualizations[3]. Many tasks for data scientists that require visual support are already covered. A thorough evaluation of the tool with domain experts and data scientists is planned for the future

[3] Video of the prototype: https://www.youtube.com/watch?v=_i09AIlncKQ

in order to ascertain, which features are still missing for our goal of an integrated, immersive tool for the analysis of high-dimensional data.

Acknowledgments

This work has been supported by the European Regional Development Fund and the Free State of Saxony (project no. 100238473). This work has been conducted in cooperation with Mercateo Services GmbH, Chemmedia AG, and deecoob Technology GmbH.

References

Borgo, R., Kehrer, J., Chung, D., Maguire, E., Laramee, R., Hauser, H., … Chen, M. (2013). Glyph-based visualization: Foundations, design guidelines, techniques and applications. *Eurographics State of the Art Reports*. EG STARs, 39–63. Retrieved from https://www.cg.tuwien.ac.at/research/publications/2013/borgo-2013-gly/

Choo, J., Lee, H., Liu, Z., Stasko, J., & Park, H. (2013). An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data. In *Visualization and data analysis 2013* (Vol. 8654, p. 865402). International Society for Optics and Photonics.

Heimerl, F., Koch, S., Bosch, H., & Ertl, T. (2012). Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2839–2848.

Kammer, D., Keck, M., Gründer, T., & Groh, R. (2018). Big data landscapes: Improving the visualization of machine learning-based clustering algorithms. In *Proceedings of the 2018 international conference on advanced visual interfaces* (66:1–66:3). AVI '18. Castiglione della Pescaia, Grosseto, Italy: ACM. doi:10.1145/3206505.3206556

Keck, M., Kammer, D., Gründer, T., Thom, T., Kleinsteuber, M., Maasch, A., & Groh, R. (2017). Towards glyph-based visualizations for big data clustering. In *Proceedings of the 10th international symposium on visual information communication and interaction* (pp. 129–136). VINCI '17. Bangkok, Thailand: ACM. doi:10.1145/3105971.3105979

Krause, J., Dasgupta, A., Fekete, J.-D., & Bertini, E. (2016). SeekAView: an intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. *Large Data Analysis and Visualization (LDAV), IEEE Symposium on*.

Wenskovitch, J., Crandell, I., Ramakrishnan, N., House, L., Leman, S., & North, C. (2018). Towards a systematic combination of dimension reduction and clustering in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 131–141.