Tiler: Software for Human-Guided Data Exploration

Andreas Henelius(✉), Emilia Oikarinen, and Kai Puolamäki

Department of Computer Science, Aalto University, Helsinki, Finland
`firstname.lastname@aalto.fi`

Abstract. Understanding relations in datasets is important for the successful application of data mining and machine learning methods. This paper describes TILER, a software tool for interactive visual explorative data analysis realising the interactive Human-Guided Data Exploration framework. TILER allows a user to formulate different hypotheses concerning the relations in a dataset. Data samples corresponding to these hypotheses are then compared visually, allowing the user to gain insight into relations in the dataset. The exploration process is iterative and the user gradually builds up his or her understanding of the data.

1 Introduction

An important goal in *Exploratory Data Analysis* (EDA) [9] is to gain insight into different *relations in the data*. Knowledge of relations is essential for successful application of data mining and machine learning methods. Investigating relations can be efficiently performed using interactive visual EDA software, that present the user different *views* of a dataset, thus leveraging the natural human pattern recognition skills to allow the user to discover interesting relations in the data.

Recently, an *iterative data mining paradigm* [1–4] has been presented and also realised in software [6–8] with the emphasis that the user wants to find patterns that are *subjectively interesting* given what she or he currently knows about the data. The system shows the user *maximally informative views*, i.e., views that contrast the most with the user's current knowledge. As the user explores the data and discovers patterns concerning relations in the data, these patterns are fed back into the system and taken into account during further exploration, so that the user is only shown views displaying currently unknown relations in data.

Although the knowledge of the user is taken into account one important problem still remains: by design, the user cannot know beforehand what views of the data that differ the most from her or his present knowledge. Thus, exploration of the most informative views might seem somewhat random to the user and the views shown might be, even though surprising, not necessarily relevant for the task at hand. Initially, the user typically wants to test specific ideas (hypotheses) concerning the relations in the data already. These ideas can also develop during the exploration. It then becomes essential to be able to *focus the exploration process to answer specific questions.* This is realised in our novel EDA paradigm, termed *Human-Guided Data Exploration* (HGDE) [5].

In this paper we present TILER, a software tool for visual EDA that realises the HGDE paradigm for efficient interactive visual EDA. TILER aims to be an easy-to-use tool for exploring relations in datasets by allowing the user to focus the exploration to investigate different hypotheses. TILER is an MIT-licensed R-package available from `https://github.com/aheneliu/tiler`.

2 Human-Guided Data Exploration

We here provide a high-level description of the key concepts in the HGDE framework, for a complete discussion and theoretical details we refer to [5].

The goal of the user is to *discover relations between the attributes in the data* by a comparison of hypotheses, which can be viewed as a comparison of two distributions with the same known marginal distributions. A permutation-based scheme is used to obtain samples from the distributions, i.e., we permute the given data under a set of constraints defined by the hypotheses. The constraints represent the relations which are assumed to be known about the data: one extreme are unconstrained, column-wise permutations (preserving only the marginals) while the other extreme is the fully constrained case where only the identity permutation satisfies the constraints. In general, the constraints are formulated in terms of *tiles*: tuples of the form $t = (R, C)$, where $R \subseteq [N] = \{1, \ldots, N\}$ and $C \subseteq [M]$ are subsets of the rows (items) and columns (attributes) of an $N \times M$ data matrix. A tile constrains permutations so that all items in a tile are permuted together, i.e., there is a single permutation for a tile operating on each $c \in C$, thus preserving the relations inside t.

Hypotheses are represented in terms of tilings (non-overlapping sets of tiles). The hypotheses are that either all the attributes in the original dataset are dependent or they are all independent can be represented with the following two hypothesis tilings: $\mathcal{T}_{\mathcal{H}_1} = \{([N], [M])\}$ () and $\mathcal{T}_{\mathcal{H}_2} = \{(\{n\}, [M]) \mid n \in [N]\}$. A correlation between two variables i and j in a subset of rows R could be studied with the following hypothesis tilings: $\mathcal{T}_{\mathcal{H}_1} = \{(R, \{i, j\})\}$ and $\mathcal{T}_{\mathcal{H}_2} = \{(R, \{i\}), (R, \{j\})\}$. In the general case, the user can focus on specific data items and specific attribute combinations. Focusing allows the user to concentrate on exploring relations in a subset of the data items and attributes, making the interactive exploration more predictable and allowing specific questions to be answered.

In TILER, the user is shown an informative projection of two data samples corresponding to the hypotheses and is tasked with comparing these and drawing conclusions. In an informative projection the two samples differ the most. A *sample* from a distribution corresponding to each hypothesis is obtained by randomly permuting each column in the data, such that the relations between attributes enforced by the tilings are preserved. A tiling hence constrains the permutation of the data. When a user discovers a new pattern, this is added as a constraint (a tile) to both $\mathcal{T}_{\mathcal{H}_1}$ and $\mathcal{T}_{\mathcal{H}_2}$, meaning that the relations expressed by this pattern no longer differs between the two hypotheses. This allows the user to iteratively build up an understanding of the relations in the data.

3 System Design

TILER is developed in R (v. 3.4.4) using SHINY (v. 1.0.5) and runs in a web browser. The tool supports the full HGDE framework and the usage of TILER is described in the video at `http://www.iki.fi/kaip/tiler.html`.

To explore relations between attributes, the user first specifies the hypotheses being compared. The tool implements different *modes* as shortcuts for typical hypotheses. The *explore*-mode (the default) corresponds to iterative data exploration where the two hypotheses to be compared are that (i) all attributes in the original dataset are dependent or (ii) they are all independent. In the *focus*-mode the exploration is focused ono investigating all relations within a particular subset of rows and columns (a focus region). The *compare*-mode implements the general case by allowing the user to specify an arbitrary hypothesis by partitioning the attributes in the focus region into groups.

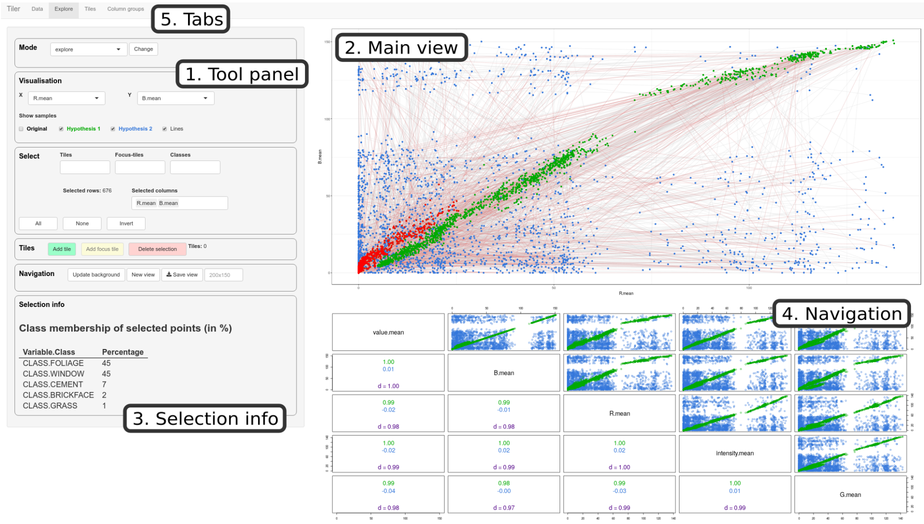
With TILER, the user visually explores a dataset by comparing two data samples corresponding to the two different hypotheses. The exploration is iterative and the user gradually finds new patterns concerning the relations in the data, which are then added as tiles. The main user interface of TILER is shown in Fig. 1 and consists of the following components:

Panel allows the mode (explore, focus, or compare) to be selected and contains tools for selection of points as well as creation of tiles and focus tiles. Points can be selected by brushing in the main view, or by selecting the data from a dropdown menu. Previously added tiles can be selected or deleted. The projection in the main view can be changed and the user can show/hide the original data and the two samples corresponding to the combined effect of the user and hypotheses tilings. The user can also update the distributions after addition of new tiles and then request the next most interesting view.

Main view shows the original data (in black) together with samples (in green and blue) corresponding to the two hypotheses being compared. Points on the same row in the sampled data matrices are connected using lines. These lines indicate how points in the data move around due to the randomisation. Since projection of high-dimensional data to lower dimensions can make interpretation complicated, we have here chosen to use 2D axis-aligned projections. The x and y axis are hence directly interpretable on their original scales. We use correlation as measure of informativeness here, as this is often intuitive and easy to interpret, but other distance measures between the two samples being compared can be used too. This measure is used to show the maximally informative view.

Selection info shows the five largest classes of the selected points (for data with class attributes). This helps the user in understanding what type of points are currently selected and gives insight into the relations in the data.

Navigation is guided by the bottom right corner showing a scatterplot matrix of the five most interesting attributes in the data. The correlations for both samples and their difference using the correlation based measure is shown. The scatterplot helps the user to quickly obtain an overview of the data.



Mode: explore ▾ Change

Visualisation
X: R.mean ▾ Y: B.mean ▾
Show samples
□ Original ☑ Hypothesis 1 ☑ Hypothesis 2 ☑ Lines

Select: Tiles, Focus-tiles, Classes
Selected rows: 676
Selected columns: R.mean B.mean
All None Invert

Tiles: Add tile, Add focus tile, Delete selection, Tiles: 0

Navigation: Update background, New view, Save view, 200x150

Selection info

Class membership of selected points (in %)

Variable.Class	Percentage
CLASS.FOLIAGE	45
CLASS.WINDOW	45
CLASS.CEMENT	7
CLASS.BRICKFACE	2
CLASS.GRASS	1

Fig. 1: The main user interface of the TILER software tool.

Tabs provide functions for loading data, listing tiles, and for defining an attribute grouping in the compare mode.

Acknowledgements. This work has been supported by the Academy of Finland (decisions 288814 and 313513).

References

1. De Bie, T.: Subjective interestingness in exploratory data mining. In: IDA. pp. 19–31 (2013)
2. De Bie, T.: An information theoretic framework for data mining. In: KDD. pp. 564–572 (2011)
3. De Bie, T.: Maximum entropy models and subjective interestingness: an application to tiles in binary databases. Data Min. Knowl. Discov. 23(3), 407–446 (2011)
4. Hanhijärvi, S., Ojala, M., Vuokko, N., Puolamäki, K., Tatti, N., Mannila, H.: Tell me something I don't know: randomization strategies for iterative data mining. In: KDD. pp. 379–388 (2009)
5. Henelius, A., Oikarinen, E., Puolamäki, K.: Human-guided data exploration. arXiv preprint, arXiv:1804.03194 (2018)
6. Kang, B., Puolamäki, K., Lijffijt, J., De Bie, T.: A tool for subjective and interactive visual data exploration. In: ECML-PKDD. pp. 3–7 (2016)
7. Puolamäki, K., Kang, B., Lijffijt, J., De Bie, T.: Interactive visual data exploration with subjective feedback. In: ECML-PKDD. pp. 214–229 (2016)
8. Puolamäki, K., Oikarinen, E., Kang, B., Lijffijt, J., Bie, T.D.: Interactive visual data exploration with subjective feedback: An information-theoretic approach. In: ICDE. pp. 1208–1211 (2018)
9. Tukey, J.W.: Exploratory data analysis. Addison-Wesley (1977)