Visual feature engineering

Susanne Bleisch

Institute of Geomatics, FHNW University of Applied Sciences and Arts Northwestern Switzerland, Muttenz/Basel, Switzerland
susanne.bleisch@fhnw.ch

Abstract

Feature engineering is a key concept in machine learning describing the process of defining the characteristics of an observed phenomenon in a way that makes it usable by an algorithm (e.g., [3]). This process often includes domain knowledge to make the features, as well as the results of the algorithms, meaningful in the respective application area. In data analysis generally, including visual data analysis, the obtained results or insights are often dependent on the employed analysis method as well as the parameters and their dimensions used. A simple but well-known example is the modifiable area unit problem [5]. Depending on the size and form of the spatial units chosen to aggregate the data, different visualizations and potentially interpretations of the information may result. In some cases, the chosen methods or algorithms and their parameters can be argued to be the right ones to support a specific analysis task, in other cases a sensitivity analysis may be helpful in determining the optimal values. Additionally, visual analytics, allowing tight integration of the interaction with the methods and parameters and the visualizations, has the potential to support the evaluation of the right or sensible analysis method and its parameters as well as to provide provenance information for the finally employed approach.

The term visual feature engineering has been used before (e.g., [4]). It usually denotes approaches that use visual methods for feature engineering or is employed where the features used for algorithmic learning or processing are of visual nature. The proposition made here is to more comprehensively think of visual feature engineering as the process of using visualizations and visual analytics approaches for defining the most useful features in the raw data (the definition of feature engineering in machine learning) as well as evaluating the 'right' methods or algorithms and their parameters respectively for an optimal result in a specific setting or application domain.

Figure 1 An example of a representation of walkability index values for road segments assigned to regular grid cells and visualized by square area. For some cells, especially at crossroads, more than one value is available and a decision was required which value should be assigned [1]. The figures show different options for value assignment and the difference between them. Left: smallest possible value assigned; Middle: largest possible value assigned; Right: difference between smallest (Left) and largest (Middle) values in red.

Visualizations have long been used to represent the outcomes of different analysis approaches, i.e. for the MAUP. They allow showing the differences that result, for example, from different parameter value assignments. In a simple example, Figure 1 shows a case where index values calculated for road segments were assigned to regular grid cells [1]. At crossroads often more than one value was available for assignment. Figure 1, right, visualizes the difference between the smallest and largest possible values in red, making explicit the range of choices available. However, for more complex problems, visual analytics approaches which allow for tight integration of

New Directions in Geovisual Analytics: Visualization, Computation, and Evaluation (GVIZ 2018).
Editors: Anthony C. Robinson and Antoni Moore; Article No. 2; pp. 2:1–2:2
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

visualization and the interaction with methods and parameters of the analysis process seem useful. This may especially be so, when the approach shall be used to achieve an optimal result for very specific or highly variable settings, i.e. in personalized visualizations and analyses. For example, we have analyzed quantified-self data from elderly persons. The aim was to analyze the data with the abilities and requirements of the individual in mind so that the intermediate results of a long-term monitoring could continually be used in periodic health-care dialog settings between person and geriatric practitioner to assess motivational or limiting aspects of habits. Analyzing quantified-self data is a challenge by itself (e.g., [2]) but especially so when the results have to be useful for the individual and not for the 'average person'. We learned that comparing the results of different approaches and parameter settings to the individual's contexts and requirements was crucial in achieving sensible results. The application of argued or 'standard' parameter settings in the analysis process is too limiting.

Data and tool availability make the personalization of data analysis processes and methods - and the resulting knowledge - increasingly feasible and important. A concerted effort in visual analytics to explore ways of supporting (personalized) analysis processes, for example through visual feature engineering as proposed above, seems timely.

1998 ACM Subject Classification Applied Computing - Cartography

Keywords and phrases visual analytics, visualization, feature engineering, algorithms, parameters, personalization

Digital Object Identifier 10.4230/LIPIcs.GVIZ.2018.2

References

1 Susanne Bleisch and Daria Hollenstein. Exploring multivariate representations of indices along linear geographic features. In *ICC 2017: Proceedings of the 2017 International Cartographic Conference, Washington DC*, 2017.

2 Tom Fawcett. Mining the quantified self: Personal knowledge discovery as a challenge for data science. *Big Data*, 3(4):249–266, 2015. URL: `http://online.liebertpub.com/doi/full/10.1089/big.2015.0049`.

3 Peter Flach. *Machine Learning - The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, Cambridge, UK, 2012.

4 Daniela Oelke. *Visual document analysis: Towards a semantic analysis of large document collections*. Phd, Universität Konstanz, 2010. URL: `http://kops.uni-konstanz.de/handle/123456789/6078`.

5 S.Openshaw and P.J.Taylor. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In N.Wrigley, editor, *Statistical Applications in the Spatial Sciences*, pages 127–144. Pion, London, 1979.