An interactive tool for transparent data preprocessing

Olivier PARISOT, Thomas TAMISIER

Public Research Centre – Gabriel Lippmann, Belvaux, Luxembourg
{parisot,tamisier}@lippmann.lu

We propose a visual tool to assist data scientists for data preprocessing: it interactively shows the transformation impacts and the information loss, while keeping track of the applied preprocessing tasks.

Data analysis is an important topic for several domains in computer science, like data mining, machine learning, data visualization and predictive analytics. In this context, the scientists aim at inventing new techniques and algorithms to handle data and discover meaningful patterns from them.

Usually, data analysis techniques are worked out by using benchmark datasets: for instance, the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) contains a lot of material for different tasks (regressions, prediction, classification, etc.) This repository is widely used, many academic papers in the domain referring to some of its datasets in order to allow meaningful comparisons with the state of the art.

In practice, preprocessing is often necessary to adjust the benchmark data to the specificity of new algorithms or methods [1]. More precisely, data preprocessing is a collection of different data transformation techniques: 'cleansing' (treatment of noise, etc.), 'dimensionality alteration' (filtering of features, etc.) and 'quantity alteration' (sampling of the data records). Moreover, a preprocessing process could drastically affect the original data, and the results of a data analysis could be clearly different depending of the answers to the following questions:

- Are the outliers removed?
- Are the missing values replaced by estimations?
- Are the nominal values replaced by numerical values?
- Have some columns/row been deleted?
- ….

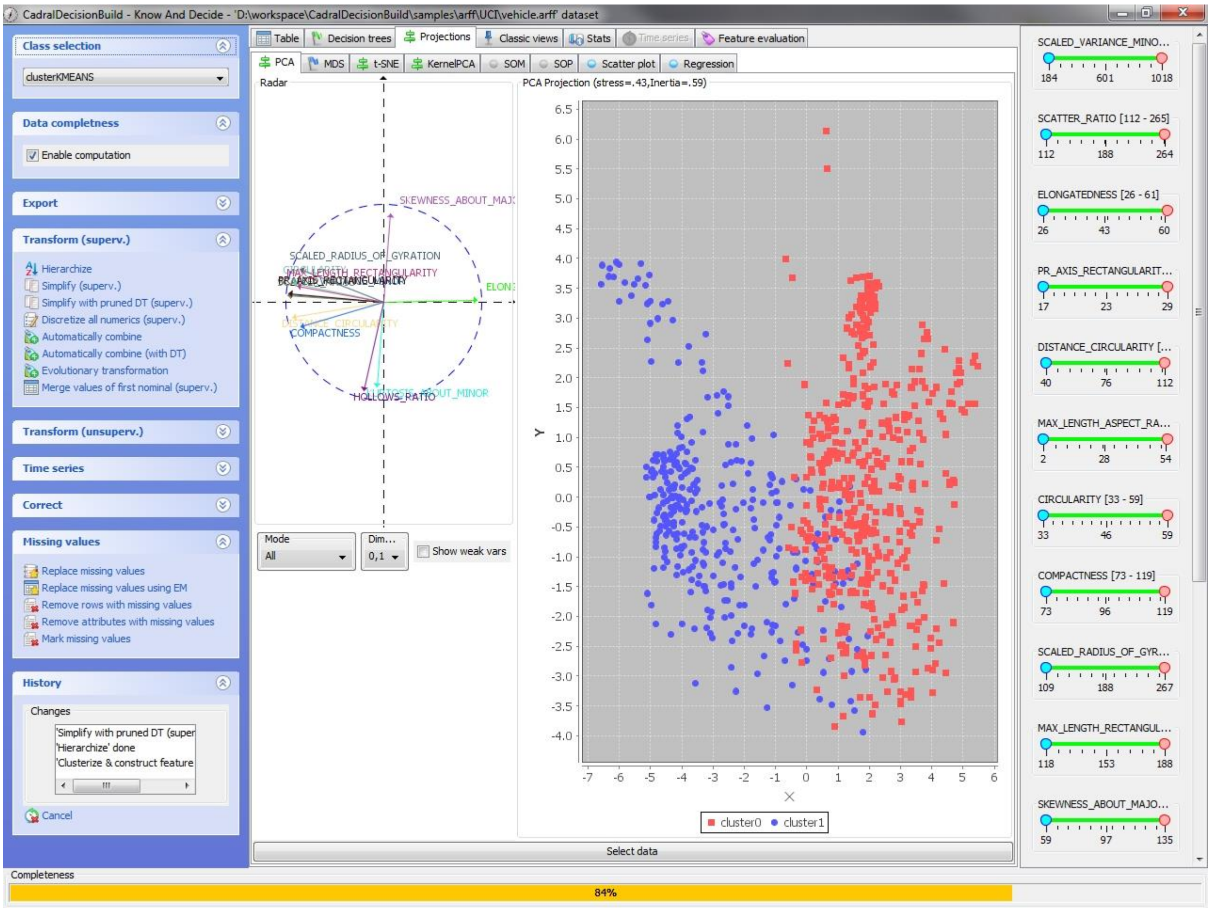
Consequently, a lot of uncertainty remains, related to this preprocessing step because the modifications are not necessarily mentioned, especially in scientific publications about data analysis.

In order to improve the transparency regarding the preprocessing phase, we have developed a JAVA standalone tool that allows transforming the data while keeping traces of the transformations. The tool is developed on top of the WEKA library (http://www.cs.waikato.ac.nz/ml/weka/), and allows to apply the following preprocessing operations: columns/rows deletion, discretization of the numerical features, feature selection, constant feature deletion, missing values imputation, outliers deletion, attributes transformation (numerical fields to nominal fields, nominal fields to binary fields, etc.).

The main features of the tool are the following:

- The tool helps to interactively transform the data: in fact, the user interface provides an exploratory process to successively apply transformations operations, and then check the results using visual views like tables, trees, heat maps, 2D projections, etc. During this process, the user can consult the history of the applied transformations and he can cancel them according to his desires/needs.

- After each data transformation, it is critical to gauge the intensity of data transformation (for example, the discretization of numerical values implies information loss) [2]. To this end, the tool instantly computes the ratio of values that are kept unchanged during the preprocessing steps [3]. This indicator is continuously shown in the user interface of the tool (Figure 1): 100% represents a slight transformation, 0% represents a considerable transformation.
- The tool provides support to generate consistent sequence of data processing: for example, if a data analyst wants to normalize these data, the tool will show him that removing outliers and extreme values should be done before.
- In addition, the tool is able to automatically transform the data for a specific task: as an example, an algorithm has been developed in order to obtain transformed data that lead to simple prediction models [3].



CadralDecisionBuild - Know And Decide - 'D:\workspace\CadralDecisionBuild\samples\arff\UCI\vehicle.arff' dataset

Class selection: clusterKMEANS

Data completness: ☑ Enable computation

History – Changes:
'Simplify with pruned DT (super
'Hierarchize' done
'Clusterize & construct feature

Mode: All
Dim...: 0,1
□ Show weak vars

PCA Projection (stress=.43,Inertia=.59)

Completeness: 84%

Fig. 1: Preprocessing of the 'vehicle' dataset from the UCI repository: the data are represented using a PCA projection, the progress bar shows the data completeness and the applied operations are detailed on the bottom left of the window.

The tool can be applied in several use cases to improve the scientific reusability of data. Firstly, it helps to write reusable scientific papers by providing a full description of the preprocessing steps that have been applied on this dataset; moreover, it will help to avoid the 'cherry picking issue', that biases the results of a lot of data analysis papers. Secondly, it helps the data scientists to inject data assumptions into real datasets (some techniques need data without missing values, other ones need data with numerical values only, etc.). More precisely, it allows the transparent construction of synthetic datasets from well-known data (like the datasets from the UCI Repository, for example) that finally can be used in experiments. As the transformation steps and the information loss indicator are explicitly shown by the tool, they can be described in the further technical/academic papers: it will improve the reproducibility for the article's readers.

A pending issue is to deal with data for which the original source is known but the preprocessing pipeline is unknown due to a lack of documentation. As a future work, we plan to create a reengineering method in order to automatically determine the preprocessing operations that have been applied, given an original dataset and its transformed version.

Links

http://archive.ics.uci.edu/ml/

http://www.cs.waikato.ac.nz/ml/weka/

References

[1] Fazel Famili et al., "Data preprocessing and intelligent data analysis", International Journal on Intelligent Data Analysis, Volume 1, Issues 1–4, 1997.
[2] Shouhong Wang, Wang Hai, "Mining Data Quality In Completeness", *ICIQ* 2007, pp. 295-300.
[3] Olivier Parisot et al., "Data Wrangling: A Decisive Step for Compact Regression Trees". CDVE 2014, pp. 60-63.

Contacts

Olivier PARISOT
Public Research Centre – Gabriel Lippmann, Belvaux, Luxembourg
parisot@lippmann.lu

Thomas TAMISIER
Public Research Centre – Gabriel Lippmann, Belvaux, Luxembourg
tamisier@lippmann.lu