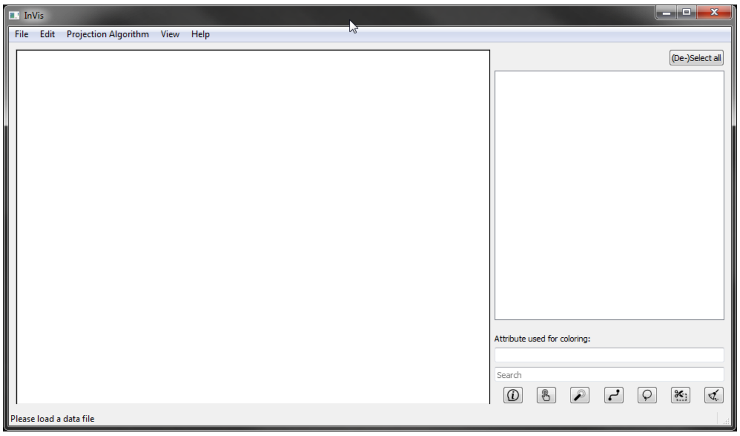
InVis User Manual

InVis is a tool for **In**teractive **Vis**ualization of high dimensional datasets, which can be downloaded from the following location: `http://www-kd.iai.uni-bonn.de/index.php?page=software_details&id=31` At the current state it covers a set of static and interactive algorithms that enable a user to explore two dimensional projections of a dataset. The following Figure 1 shows a screenshot of the graphical user interface of the tool, without any dataset loaded. In the following sections, the menu entries and the user interface will be explained.



Please load a data file

Figure 1: Starting up the InVis tool.

The File Menu

The File menu lets the user load a dataset and export the parameters that generate the currently viewed projection. In addition, basic implementations of four different pattern mining algorithms are available that let the user export the top-k patterns in a format that can be re-imported by the tool.

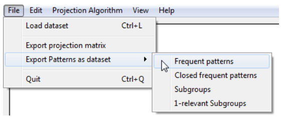


Figure 2: The File menu.

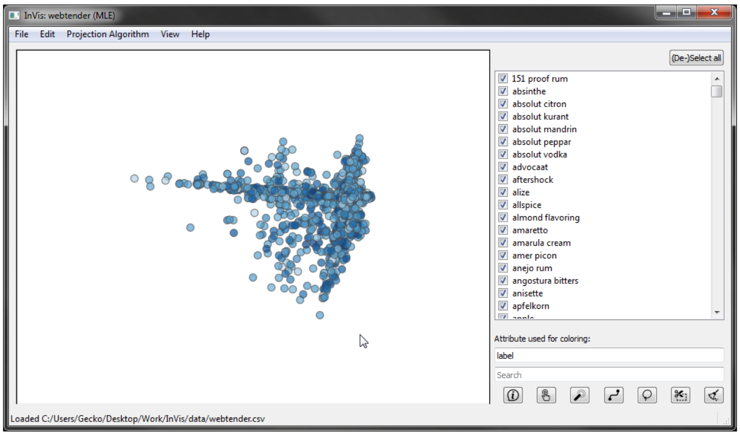
Loading a Dataset

In general, csv, arff and libsvm data-files can be loaded into the tool. Note, that for loading csv files, the parser is quite strict. The data is read row-wise, with the first line being a header and every subsequent row being interpreted as a data-record. The first column is considered to be a string-valued ID or name of each data record. All values of a row have to be separated by commas and only the name entry is allowed to be non-numeric. Also, by default, the last column is considered as label and the numeric entries may not be quoted (e.g. ”12.54”). The following table illustrates the csv dialect that is well understood by the tool.

Example of an accepted csv file
name,a1,a2,a3,label
Name1,1.0,2.3,2.1,100
Name2,3.0,1.3,2.3,80

Example of a not accepted csv file
missing values can be interpreted as zero
”Name1”; 1.0; 2.3; ”2.1”; A
”Name2”; 3.0;   ; ”2.3”; B

Once a dataset is loaded, the tool automatically performs a principal component analysis and renders a visualization of the data, projected into the first two principal directions. The attributes of the dataset are displayed on the right hand side. In case, the user wants to ignore an attribute, he can do so by un-checking the corresponding entry.



☑ 151 proof rum
☑ absinthe
☑ absolut citron
☑ absolut kurant
☑ absolut mandrin
☑ absolut peppar
☑ absolut vodka
☑ advocaat
☑ aftershock
☑ alize
☑ allspice
☑ almond flavoring
☑ amaretto
☑ amarula cream
☑ amer picon
☑ anejo rum
☑ angostura bitters
☑ anisette
☑ apfelkorn

Attribute used for coloring: label

Search

Loaded C:/Users/Gecko/Desktop/Work/InVis/data/webtender.csv

Figure 3: The initial view, after the webtender dataset is loaded.

The Edit Menu

This menu lets the user adjust the area (in pixel) that is considered adjacent to an embedded data record. Especially when using a touch screen, this can be a helpful option. Also the way that the numeric entries of the data records are discretized can be adjusted here. This option can be helpful, when exporting patterns from the dataset (via the Edit menu), or for displaying the ten most frequent item sets within a selected area of the embedding (see next Section). In addition, for the ease of use, when utilizing the experimental feature of must-link and cannot-link constraints, this menu offers to clear all links.

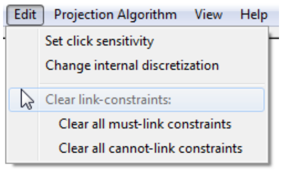

Edit
Projection Algorithm
View
Help
Set click sensitivity
Change internal discretization
Clear link-constraints:
Clear all must-link constraints
Clear all cannot-link constraints

Figure 4: The edit menu.

The Projection Algorithm Menu

Here, the user can select the embedding algorithm that renders the visualization. The menu is sub-divided into static and interactive techniques. The first constitute a set of classic embedding methods, which have proven to be fruitful time and time again. The second set of algorithms allows the user to playfully interact with the layout embedding.

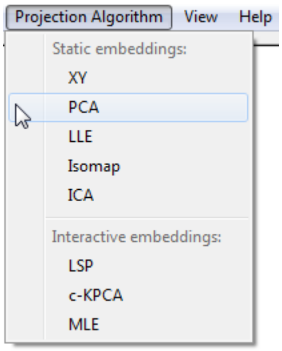


Figure 5: The set of algorithms that can be used by the analyst to project the data into two dimensions.

Static Embeddings

The *static embeddings* let a user have a look at the data via commonly used embedding techniques:

XY (XY scatter plot of the two first selected attributes)

PCA (Princial Component Analysis)

LLE (Locally Linear Embedding)

Isomap (Isometric mapping: basically, multidimensional scaling applied to the knn-graph)

ICA (Independent Component Analysis)

Note that while in a static embedding, the user can still interact with the visualization via searching, highlighting and filtering, etc.. The user can also set and un-set control points, however, since the embeddings are static, he cannot relocate them.

Interactive Embeddings

The *interactive embeddings* let the user actively layout and shape the projection of the data by selecting and re-location individual data records within the embedding as control points. Relocating these control points triggers the underlying embedding algorithm to re-calculate the projection with respect to the user provided feedback. The result is rendered instantly, which yields a life updating visualization. The three interactive embedding techniques implemented in this tool are:

LSP (Least Squared Error Projection)

c-KPCA (Constrained Kernel Principal Component Analysis)

MLE (Most Likely Embedding)

The View Menu

This menu lets the user control the look and feel of the visualized data. The adjustments can be made in the following way: The first four entries let the analyst chose the color scheme in which the data records are highlighted; the default is a blue scale. In addition, the point size for each data record can be set proportional to the considered label value. This can e.g. be of use when studying a dataset of patterns, with the label being their support.

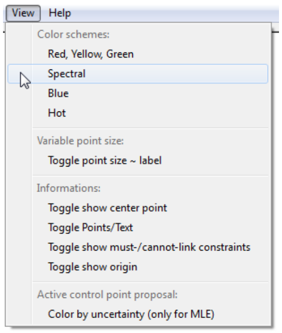


Figure 6: Options that can be adjusted in the view menu.

The third section lets the user toggle the visibility of various elements which may be of help during an analysis session. The last option is an experimental feature which is only available for the MLE algorithm. A side product that can easily be calculated by this probabilistic algorithm, is the confidence about the location of each data record within the embedding. This confidence is then used to colorize the visualization. In one of the following sections, a “magic wand” button is introduced that auto-selects good control points for the MLE method. The magic behind the selection procedure uses exactly this confidence value.

The Help Menu

A survey of the most commonly used interaction methods with the visualization and their keyboard shortcuts. In addition, this document is displayed on pressing F1.

Left-click & drag lets the user re-locate the nearest control-point in a “drag ’n drop” like manner.

Right-click displays information of the clicked data-record (e.g: attribute_name:values>0)

Middle-click lets the user select or de-select a data-record as control-point.

Mouse-wheel lets the user zoom in and out on the mouse pointer.

Ctrl+left-click-lasso-select lets the user select all data records in a region of the embedding.

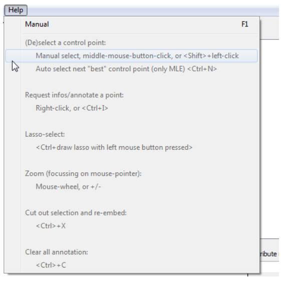

Help
Manual
F1
(De)select a control point:
Manual select, middle-mouse-button-click, or <Shift>+left-click
Auto select next "best" control point (only MLE) <Ctrl+N>
Request infos/annotate a point:
Right-click, or <Ctrl+I>
Lasso-select:
<Ctrl+draw lasso with left mouse button pressed>
Zoom (focussing on mouse-pointer):
Mouse-wheel, or +/-
Cut out selection and re-embed:
<Ctrl>+X
Clear all annotation:
<Ctrl>+C
ribute

Figure 7: A quick reminder of the shortcuts for interaction with the canvas.

Interaction

When using the InVis tool in combination with a touch screen, a keyboard might be disturbing. For this scenario (and for shortcut lazy users) all interaction methods are also available via the graphical user interface. Note that the buttons change their appearance to a colored version once they are active. The buttons meanings are the following:

ⓘ, ⓘ Query information on an individual data record, by clicking on it.

☝, ☝ Selecting a data record as control point.

🔎, 🔎 “Magic wand” control point selection (only available for MLE).

- Introduce must- and cannot-link constraints data record pairs.
- Lasso-select all data records within a region in the embedding.
- Consider only the lasso-selected data records.
- Clear all search annotations and information queries.

When querying information on a single data record, usually the attribute values of that data record are of interest. For a high dimensional dataset this can quickly get out of hand. For this reason, here only the non-zero entries are displayed.

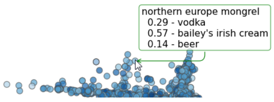


Figure 8: Queried information on a single data record.

The selected control points are highlighted with a pink bold border. This way they are easy to distinguish from the regular data records.

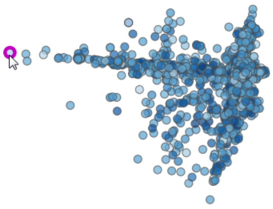


Figure 9: A control point.

When lasso-selecting an area, the enclosed data records are emphasized and a word cloud of the ten most frequent attribute sets is displayed in the bottom left corner. This helps the user to quickly grasp the dominant attributes and attribute combinations within the region of interest.

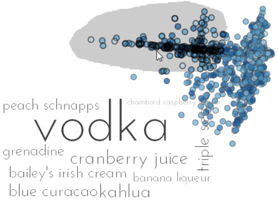
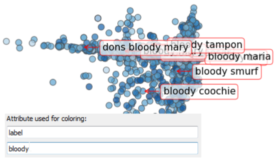


Figure 10: A lasso-selected area and its most influential attribute combinations.

Highlighting and Searching

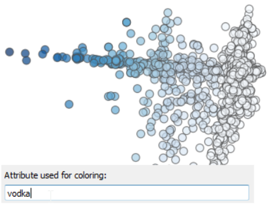
A user can also search for a sub-string that is contained in the data-records name. This can be done by using the free text fields at the bottom right corner of the user interface. The matching results will be highlighted in red. Here, the search term “bloody” reveals the embedding locations of all data records that possess this term as part of their name (e.g. the bloody marry).



Attribute used for coloring:
label
bloody

Figure 11: Searching parts of the data record ID’s.

The second free text field offers the user to enter an attribute name, by which the data points are colored. The value of the attribute determines the shade of the color. For the webtender dataset, the first principal direction coincides heavily with the attribute *vodka*.



Attribute used for coloring:
vodka

Figure 12: Colorizing the data records by an attribute value.