Viscovery® SOMine 7.1 – What's New?

New in Viscovery® SOMine 7.1 Data Mining Suite

In **Viscovery® SOMine 7.1,** the following features were added. Additionally, errors discovered in version 7.0.2 have been corrected.

Visual Clusters

■ The presentation of the correlation matrix in the Statistics windows has been changed. A matrix is shown instead of a list with every pair of attributes. The attributes in the rows and the columns of the matrix can be selected independently.

■ The layout of the *Tune Histograms* dialog was changed: the attribute list occupies the full height of the dialog so that more attributes are visible at the same time. Furthermore, the height of the lower pane can now be changed.

■ List entries in the *Remove Records* page of the *Tune Histograms* step can be reordered by dragging them to their new position.

■ Charts can be exported as graphic files.

■ Columns of tables can now be hidden and shown. Right-clicking the list header offers a context menu, where the columns can be switched on and off.

■ More differences in settings between workflow steps are shown. Additionally, differences between steps that are not connected to the same direct ancestor (but to some indirect ancestor) are now shown.

■ In the *Define Nominal Values* dialog, value frequencies are now shown next to the nominal values and aliases that have been defined.

■ The *Duplicate* function has been added to the *Define New Attributes* dialog of the *Define Attributes* step of the *Preprocess* workflow. This provides a convenient way to create copies of formulas.

■ New functions can be used in formulas:

- *find(what, where, start)* and *pos(what, where, count)* return the character position at which a string was found.

- *substr(str, pos1, pos2)* returns the part of the input between the specified positions.

- *mean(n1, ...)*, *sum(n1, ...)*, *count(n1, ...)* take an arbitrary number of arguments and compute the mean, sum, and count of non-missing values. If all values are missing, *mean* returns a missing value, *sum* and *count* return zero.

■ Functions *min(n1, ...)* and *max(n1, ...)* can now be given an arbitrary number of arguments (must have at least one) to compute the minimum and maximum of all non-missing values. The result is only a missing value if all argument values are missing values.

■ Functions *min*, *max*, *mean*, *sum*, and *count* can now split a string at a separating character and operate on the parts (after converting them to numbers).

■ There is now an option in the *Preferences* dialog to set how long popups in map windows and project windows are visible.

■ In the *Define Nominal Attributes* dialog of the *Define Attributes* step of the *Preprocess* workflow, the attribute order to be used in the data mart can now also be defined by dragging one or more highlighted attributes to the new location in the order. A facility to define the attribute order has already been available in the *Define New Attributes* dialog of the same workflow step, but the new facility is more convenient.

Viscovery® SOMine 7.1 – What's New?

- The ability to sort attributes by clicking the column header in the *Define Nominal Attributes* dialog has been removed to avoid accidently destroying a manually conducted attribute order with a single mouse click.
- In the *Select Attributes to Keep* dialog of the *Write Data Mart* step of the *Preprocess* workflow, the ability to arrange the order of attributes has been removed. When attributes are sorted by clicking the column headers in this dialog, the order in the data mart is not changed. The attribute order can be adjusted in the *Define Nominal Attributes* dialog, which occurs earlier in the workflow.
- When an old project file is upgraded, the SOM model files referenced are upgraded to the latest file version in such a way that the old project file (if the user chooses not to overwrite the old file during the upgrade) can still be used with an earlier application version.
- The tabs of the *Preferences* dialog have been rearranged to fit usage patterns better.
- Multi-selection of attributes in the score charts has been enabled.
- Scatter plots now draw the points of the trend line at the mean of values in the interval along the horizontal axis, not at the center of the interval.
- The ability to change the alignment of chart legend texts has been removed. The texts are now always left-aligned.
- The option that automatically determined the size of the font of axis labels of charts has been removed to avoid excessively large axis labels.

SOM Model Editing

- More than one thumbnail of a document (image) can now be shown when the mouse pointer hovers over a map picture. The size, number, and arrangement of thumbnails can be adjusted in the *View* tab of the *Preferences* dialog. The order of thumbnails can be influenced by sorting records in the *Data Records* window.
- The arrangement of windows that are shown when a new SOM is displayed for the first time can now be customized.
- The default font size for labels in map pictures has been increased for better legibility.
- The *Statistics* window of a SOM now analyzes all attributes that are present in the data mart. As a consequence, the data mart is scanned even for the descriptive statistics (this was not required previously, as all descriptive statistics were already present in the model).
- The *Show Columns* dialog has been renamed *Data Mart Attributes* and is invoked from *View | Data Mart Attributes*. The selection and order of attributes defined in this dialog is now applied to both the columns in the *Data Records* window and the attributes listed in the various tabs of the *Statistics* window.
- When attribute lists are sorted in the *Descriptives* or *PCA* tabs of the *Statistics* window, the order is now also reflected in the columns of the *Data Records* window. This order is now preserved in the Arrangements.
- The function that sorts map pictures according to their visual similarity (*Map | Attributes*, *Similarity* button) now arranges primarily the visible map pictures. The hidden map pictures are also analyzed for similarity, but are then ordered after the most similar visible map picture.
- In the previous version, opening the *Clusters* window took a long time when a SOM model had very many attributes (several thousands) and many clusters. The opening time for this window has been reduced considerably.

Viscovery® SOMine 7.1 – What's New?

■ When the first post-processing formula is defined for a SOM model while the *Data Records* window is open, the formula is now visible immediately; previously it was visible only after the *Data Records* window was closed and opened again.

■ The attribute that sorts the data shown in the *Data Records* window of a map is not reset when the list of records is changed or when the window is closed. Additionally, the sort state is recorded in the arrangements and is restored when switching to a different arrangement.

■ During monitoring (*Map | Monitor*), the current record number is now displayed.

■ Previously, node positions were not restored precisely when node data was copied (via *Edit | Copy Labels, Selection, Path*) and then pasted back into the same map. This has been corrected so that the positions are now preserved faithfully. Also, the question "Use replacements?" is no longer asked.

■ Because it is usually of interest to see the *Profile* column of the *Group Profile* window in combination with the *P-value*, it has been moved to the right by two notches.

■ The user interface now uses the term *Nearest Nodes* instead of *Neighborhood* because the latter term has a particular meaning in data analysis and statistics.

Explore and Classify

■ The SOM model that is created in the *Classify* workflow can now be saved under a new name using *File | Save Copy As*. This copy can be loaded and modified as any other SOM model; but such modifications do not affect the workflow from which it originated. This allows a SOM data representation to be created, e.g., for Big Data, by first creating a SOM model from a data sample and then by applying this model to the other data records.

■ The results of *Export Classification* in the *Classify* workflow and *Export Scoring* in the *Score* workflow have always been written back into the application data mart. This happens earlier in the workflow now: in the *Choose Model* step of both the *Classify* workflow and the *Score* workflow. Therefore, classification and scoring results can already be investigated at these workflow steps using the *Statistics* and *Data Records* windows. It is no longer necessary to export the results.

■ The *Choose Model* step of the *Classify* workflow allows attributes to be selected for association to the model. In addition to numeric attributes, it is now possible to select attributes that are defined as nominal in the data mart.

■ The *Classify* workflow creates a new read-only model that includes the data statistics gathered from the application data mart. Previously, the model data statistics were augmented by the application data statistics. Now the **model data statistics are no longer present in this new read-only SOM model**. This avoids inconsistencies in the Statistics window that arose in some cases. The statistics that were formerly presented for attributes marked with "(Applied)" are now available under the original attribute names, and the statistics that were formerly presented for the original attribute names are no longer available.

■ When a classification was evaluated previously, it was necessary that the names of actual clusters recorded with the result data were identical to the names used in the model. This restriction has been lifted. Different cluster names can now be used in the data. However, a nominal definition must be established in the *Define Attributes* step of the *Preprocess* workflow that loads the result data. Here, the primary values must be defined to be the cluster names from the model, and the alias values are the names as they occur in the data.

Viscovery® SOMine 7.1 – What's New?

- Post-processing formulas can now be assigned a type. This type is used when the formula results are exported in the *Export Classification* step of the *Classify* workflow. It is not mandatory to define a type for existing post-processing formulas in existing SOM models; in such cases, the results continue to be converted into a string representation and exported as strings.
- The list of attributes in the *Attribute Values Across Clusters* chart can now be rearranged. The chart now uses more colors.
- Transformations specified for attributes in an application data mart are now applied when the attributes are associated to a model in the *Classify* workflow.

Predict and Score

- At the *Check Application* step of the *Score* workflow, the new *Optimal Gain* chart is available. This chart allows data records to be grouped for a campaign such that the predicted response is maximized.
- The *Prioritize Attributes* dialog of the *Compute Local Model* step allows initialization of priorities from an existing SOM model. Previously, the SOM models to select from were limited to those created in sibling workflow steps that originated from the same *Compute Global Model* step. This restriction has been lifted. Any SOM model can now be selected. Priorities are taken only from those attributes in the SOM model that have the same names as the attributes used for the global model in the preceding *Compute Global Model* step.
- Pointing at the separating line in the *Define Score Groups* chart now shows the exact location of the score group limit.
- When score group separators are dragged, the number of records where the separation takes place is shown in a popup.
- The table listing the global regression result can now be sorted by clicking column headers.
- Using the *Existing map* option (section *Local Models Based on Map* in the *Define Optimization Strategy* dialog of the *Compute Local Model* step), one map training can be saved by using an existing map. Previously, the only maps allowed were created in workflow steps originating from the same *Compute Global Model* step. This restriction has been lifted. Other maps can be used as long as they have the same variable names and preprocessing in those variables that are used in the global regression model.
- Previously, score charts were computed with two different preprocessing options: for attributes that are present in the model, the model's nominal definition and the data mart's value replacements were used. Now both the nominal definition and value replacements are taken from the model.
- The *Coefficients* window of local regression models (*Predict* workflow) shows several pictures of quality parameters. The use of white background color in these pictures has changed somewhat to indicate in a uniform way when the parameter is "not meaningful".
- The positioning of axis labels was not optimal in some charts in the previous version. This has been corrected.
- The minimum and maximum values displayed at various stages throughout the *Predict* workflow previously showed values computed from the entire dataset. However, means, standard deviations, and numbers of missing values were computed from the model data partition only. This has been corrected so that minimum and maximum values are computed only from the model data partition.

Viscovery® SOMine 7.1 – What's New?

- When a SOM with a local model was opened using a name that was saved earlier in a *Compute Local Model* step of the *Predict* workflow, the non-linearity diagnostics and local regression summary were missing. These are now made available. However, for models that were created with an earlier version of Viscovery SOMine, the information is still missing.

Enterprise Data

- Data import from extremely large Microsoft Excel files (*.xlsx) is now possible.
- The *Export Data Mart* procedure now suggests a data mart to export according to the currently selected workflow step, and suggests a file name for the destination file.

Workflow Automation

- The Workflow Automation service can now send notification emails via mail servers that require password authentication. Messages sent by the Workflow Automation to the configured mail server are not encrypted, however.

Compatibility Notes

- The algorithm that partitions data randomly into model data, test data 1, and test data 2 in the *Choose Data Mart* step of the *Predict* workflow has been changed. Previously, batches of 100 records were shuffled and assigned to the partitions with the exact percentage given; the new method assigns each record independently to the partition according to the probability given by the percentages.
- The *File | Export | Map Node Values* function now writes the dichotomous node values of nominal attributes instead of a reconstructed nominal value. Generally, this is more useful. Node values with reconstructed nominal values can be exported using *Edit | Select All* followed by *File | Export | Selection*.
- Functions *min(a, b)* and *max(a, b)* now ignore arguments that are missing values; previously, they returned a missing value when at least one argument was a missing value. This **change is also applied to projects and models that were created with an earlier version** of Viscovery SOMine.
- Take note that the *Compensate Correlations* **option is now off** by default. When projects are loaded, the setting recorded in the project is still heeded.
- Selecting *Reset* on an incomplete (yellow or white) workflow step now **also resets all subsequent steps** (i.e., those to the right of the step).