COALDA - User Manual

Stefanie Wiltrud Kessler

April 16, 2010

Contents

This manual assumes that you have properly installed and started the database and the Matlab SOM server before starting this software. For instruction on their installation, please refer to the installation manual.
There are several components of the display as you can see on the screenshot in figure 1. The Map Display is the main component which shows the visualization of the data. It will initially be empty, when you start the software.
To calculate a new SOM with the feature vectors in the database click the `Start` button in the side pane. If you want to load a specific calculation out of the database, you can do this with the menu item `Load` from the `Data` menu. The calculation will appear in the Map Display.

1 Map Display

Here you can see the visualization of the self organizing map calculated out of the data in the database. Different calculations are shown in different tabs. The title of the tab contains the calculation ID of the calculation that is shown. You can close tabs by clicking on the title with the right mouse button.
A double click with the right mouse button adjusts the visualization to the bonds of the display.
You can move the map by clicking with the right mouse button on an empty space (that is where there is no node or edge) and drag.
You can zoom on the display with the mouse wheel, turn it forward to zoom in and backward to zoom out. Zoom is always performed in the center of the display, drag the display to zoom into areas you want. At a certain point it is not possible anymore to zoom in. If at that point you turn the mouse wheel to zoom in while over a node that has associated feature vectors, a new map is calculated for only the feature vectors associated with the node. The result of the zoom is shown in a new tab.
If you click on a node or an edge, information about the item is shown in the components at the bottom (see following sections).
To label all feature vectors associated with a node, perform a double click with the left mouse button. A dialog will open where you can chose the label you want to give all feature vectors of the current node. You will also have to enter a confidence level between 0 (no confidence) and 100 (high confidence). It is not possible to directly label single feature vectors.
On how to recalculate the map with different settings, see section 5. On how to change the colors and nodelabels of the visualzation see section 6.

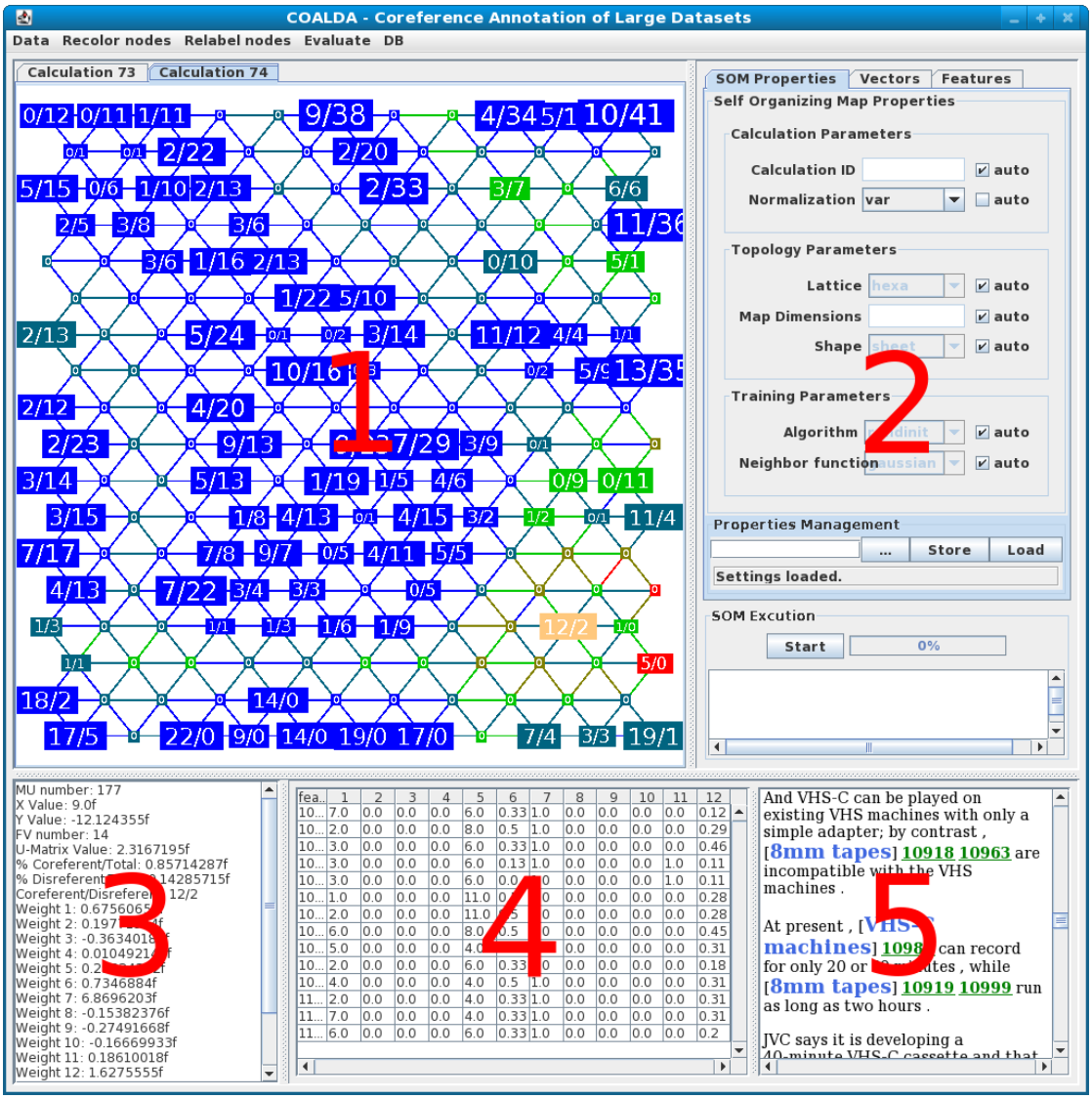

Figure 1: Screenshot of the Application (1 - Map Display, 2 - Side Pane, 3 - Map Unit Information, 4 - Feature Vector Information, 5 - Document Information)

2 Map Unit Information

This text area shows information about the item currently selected. An item can be a node or an edge.
For an edge the following information is shown:

Edge number ID of the edge.

Edge Source ID of the node where this edge starts.

Edge Target ID of the node where this edge ends.

U-Matrix Value Distance of the start node to the end node, calculated in the feature space, a high value (around 2 or 3) significas that the nodes are far away from each other and is colored red, while a low value (near 0) colored blue means the nodes are close together.

For a node (map unit) the following information is shown:

MU number ID of the node

X Value X value of the node in the two-dimensional map space

Y Value Y value of the node in the two-dimensional map space

U-Matrix Value Mean distance to all neighbour nodes calculated in the feature space, a high value (around 2 or 3) significas that the nodes are far away from each other and is colored red, while a low value (near 0) colored blue means the nodes are close together.

% Coreferent/Total Percentage of associated feature vectors of this node that are labeled as coreferent

% Disreferent/Total Percentage of associated feature vectors of this node that are labeled as disreferent

Coreferent/Disreferent Number of coreferent feature vectors associated with this node / Number of disreferent feature vectors associated with this node

Weight i for i in number of features, the weight value of this node for the feature dimension i, what this feature is can be seen in the feature tab in the sidepane, see section 5

FV number Number of feature vectors associated with this node

Assigned Feature Vectors IDs of feature vectors associated with this node

3 Feature Vector Information

This shows a list of the feature vectors associated with the curretly selected node. The first column contains the ID of the feature vector, the following columns contain the values for the featues.
If you click on a feature vector, the two markables that form the link the feature vetor belongs to are shown in the document information (see 4). You can also select multiple feature vectors. If no feature vector is selected, the markables of all feature vectors are shown.
You can make this component bigger by dragging the right border. It will stay that way for the next nodes you select. If you want the component to return to its original size, click the border with the left mouse button and select another node for the change to take effect.

4 Document Information

This shows the text of the document containing the markables that form the link where the currently selected feature vector belongs to. Multiple documents may be shown if multiple feature vectors are selected or the markables of a link come from different documents.
A markable is between square brackets []. The number right after the closing bracket is the ID of the feature vector this markable belongs to. There can be multiple IDs for one markable. If the feature vector is labeled as coreferent, the ID is colored green, if the feature vector is labeled as disreferent, the ID is colored red.

5 Side Pane

In this panel you can make adjustments that affect the settings for recalculating the map. If you want to recalculate with the current settings press the `Start` button. Status messages about the calculation will be shown in the text area below. If everything is successful, a new tab with the result will open in the map display.
Errors will not be shown in the text area. Also if you zoom in, the status of the calculation will not be shown.
The progress of the calculation will also be shown in the progress bar next to the `Start` button. The calculation might take a while to start if you compute new feature vectors.
The side pane has three tabs that will be explained in the following. In the `SOM Properties` tab settings can be made that affect the configuration of the SOM to be calculated. The feature vectors the calculation should work with can be selected in the `Vectors` tab. The `Features` tab shows the features that have been used for the calculation, a subset of them can be chosen for a new calculation.

5.1 SOM Properties

Here you can adjust the properties of the map for recalculation.

Calculation ID Will create (or overwrite) the calculation with that ID, so please be sure the ID doesn't exist before you enter it. Best leave on auto.

Normalization Normalization method used to normalize the data before calculating the map. Use 'var'. Default is none.

Lattice If to calculate a hexagonal or rectangular map, only hexagonal maps are currently supported by the visualization. Default is hexagonal.

Map Dimension Enter format [x:y]. Default is auto-estimate.

Shape If to wrap the map at the ends on both sides (toroid), one side (cycle) or none (sheet). Default is sheet.

Algorithm Choose the algorithm for training the SOM. Default is batch.

Neighbor function Function used to calculate neighbour nodes of a winning node. Default is gaussian.

You can store or load a configuration on your disk with the `Properties Management` in the lower part of this tab.

5.2 Vectors

Here you can select whether the new SOM should use all feature vectors for the next calculation or only those associated with currently selected nodes. To select more than one node press the control key and click the nodes one after another. Selected nodes are highlighted in the display.

5.3 Features

This tab contains a list of the features found in the feature definition file. What is shown is the comment behind the feature declaration or if there is no comment then the feature declaration. You have to enter the path to the file that contains the definitions for the current calculation in the file `coalda.properties` in the property `featureDefinitionFile`. You can select a subset of the feature to recalculate the map by removing the check mark next to the features you don't want to include. This feature inserts new feature vectors in the data base. This process is done before the actual calculation starts and may take some time.
You can remove the additional feature vectors created in the database with the menu item `Clear database` from the `DB` menu (see following section).

6 Menu Bar

The menu `Data` contains functionality for loading and exporting calculations. The functions of the menus `Recolor Nodes` and `Relabel Nodes` affect the visualization of the map. In the menu `Evaluate` a crude quality measure of a calculation can be calculated. Finally, the menu `DB` contains the function to delete feature vectors from the database.

6.1 Data

To use the functions in this menu, you have to know the calculation ID of the calculation you want to load or export.

Load Loads the calculation with the ID you enter in a new tab. If the ID doesn't exist, nothing happens.

Export Exports the calculation with the ID you enter to a text file at the location you enter.

6.2 Recolor Nodes

You can chose the field that is used for coloring the nodes of the map. Edges are always colored by U-matrix value.
The nodes with the lowest values in this field will be colored blue, highest values will be red. Between blue and red are the colors steel blue, green and brown (from low to high). For example for values between 0 and 1, the colors are blue for the nodes with lowest values (0 to 0.2), steel blue for the next lowest (0.2 to 0.4), green for middle values (0.4 to 0.6), brown then (0.6 to 0.8) and red for highest values (0.8 to 1.0).
Selected nodes do always have a different color from the other nodes.
The possible fields that can be used for coloring are:

nodeUmatValue (default) U-matrix value of the node. Blue nodes are closer to their neighbors than red ones.

nodeFVNumber Number of feature vectors associated with that node. Blue are nodes with 0/few feature vectors, red are nodes with a lot of associated feature vectors.

nodeProportionDis Proportion of disreferent feature vectors out of all feature vectors associated with that node. Blue means that nearly all feature vectors are coreferent, red means that nearly all are disreferent.

nodeCorefNumber Number of coreferent feature vectors associated with that node. Blue is 0/few, red is a lot.

nodeDisrefNumber Number of disreferent feature vectors associated with that node. Blue is 0/few, red is a lot.

nodeWeight$i$ Weight of the node for dimension i in feature space. Blue is small, red is high. Values can vary.

6.3 Relabel Nodes

You can chose the field that is used for labeling the nodes of the map. The nodelabel can affect the size of the node.
The fields for labeling are:

nodeCoDisLabel Shows the number of coreferent and disreferent feature vectors associated with that node, the first number is for coreferent and the second for disreferent (coref/disref). Feature vectors that are not labeled do not show up in this nodelabel.

nodeFVNumber Shows the number of feature vectors associated with that node, no matter if they are labeled or not and coreferent or disreferent.

nodeCorefNumber Shows the number of feature vectors associated with this node that are labeled coreferent.

nodeDisrefNumber Shows the number of feature vectors associated with this node that are labeled coreferent.

nodeKey Shows the ID of the node, can be used to unambiguously identify nodes.

6.4 DB

The menu item `Clear database` implements the function to delete the feature vectors from the database, that were created when a subset of the features was used for a calculation. It is recommended to use this function every time you end a session where you created such feature vectors.
In special situation not all feature vectors might be deleted or even too many feature vectors are deleted. Please backup your data.

6.5 Evaluate

In the menu `Evaluate` a crude quality measure of a calculation can be calculated. This can only be done if most or at least a part of the links in the data set have been labeled. To use the function in this menu, you have to know the calculation ID of the calculation you want to evaluate.
For evaluation attempt to label the feature vectors using the gold standard labels. For every node the label of all associated feature vectors is set to be the gold standard label of the majority of its associated feature vectors. So if a node has five coreferent feature vectors and seven disreferent feature vectors, all of the feature vectors would be labeled as disreferent.
After all feature vectors have been labeled, we compare the assigned labels to the gold standard labels. Links that are labeled coreferent (disreferent) and are really coreferent (disreferent) are called true positives tp (true negatives tn). Links that are labeled coreferent (disreferent) and are really disreferent (coreferent) are called false positives fp (false negatives fn).

From these numbers we can compute precision and recall. Precision P indicates how many of the feature vectors that have been labeled coreferent by the software are coreferent in the gold standard. Recall R indicates how many of the feature vectors that are coreferent in the gold standard have been labeled as such by the software.

$$P = tp/(tp + fp)$$

$$R = tp/(tp + fn)$$

To create one measure out of precision and recall, the F1-measure F_1 is computed. The F1-measure is the harmonic mean of precision and recall.

$$F_1 = 2PR/(P + R)$$

Results are printed to the command line.