Gradient visualization of grouped component planes on the SOM lattice

Georg Pölzlbauer[1], Michael Dittenbach[2], Andreas Rauber[1]

1 - Department of Software Technology
Vienna University of Technology
Favoritenstrasse 9-11/188
Vienna, Austria
{poelzlbauer, rauber}@ifs.tuwien.ac.at

2 - eCommerce Competence Center - ec3
Donau-City-Strasse 1
Vienna, Austria
michael.dittenbach@ec3.at

Abstract

The Self-Organizing Map has been successfully applied in numerous industrial applications. An important task in data analysis is finding and visualizing multiple dependencies in data. We propose a method for visualizing the Self-Organizing Map by decomposing the feature dimensions into groups with high correlation or selections by domain experts. Using Gradient Visualization we plot a vector field for each of these groups on top of the map lattice, with arrows pointing towards the nearest cluster center. We provide a real-world example from the domain of petroleum engineering and point out our technique's usefulness in understanding mutual dependencies hidden in the data.

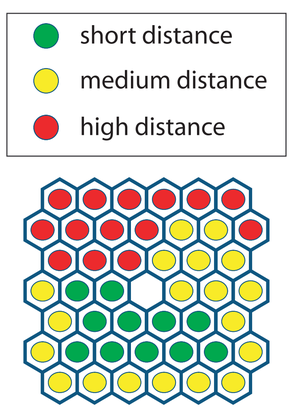
Goals

The goals of our method are to

- find a visualization that takes advantage of apriori knowledge and known variable correlations
- show groups of variables simultaneously
- depict clustering structures at desired levels of detail
- be applicable to real-world scenarios
- decompose the clustering structure into contributing factors
- provide a visualization to be plotted on top of other color coded map lattices to maximize simultaneous information presentation
- complement other visualization techniques such as U-Matrix or graph projection methods

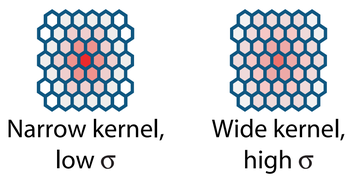
Gradient Visualization Steps

A two-dimensional vector is computed for each map unit. In the first step, the distances from the current prototype vector to the other prototype vectors are computed.



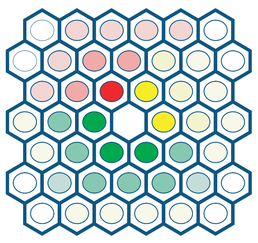
Input space distances

In the next step, the distances are weighted by the neighborhood kernel, such that close units are emphasized. The kernel function usually requires a parameter σ that determines how wide the kernel is.



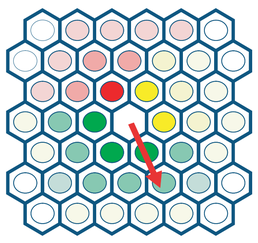
Narrow kernel, low σ

Wide kernel, high σ



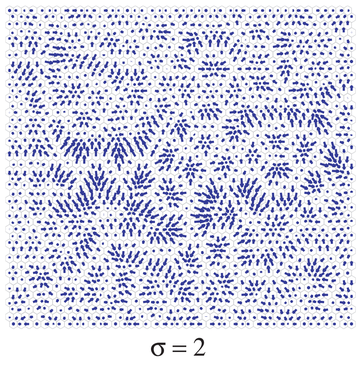
Weighted distances

Finally, the vector is computed by comparing positive and negative orientation of x- and y-axis of the map lattice. The lenght of the vector indicates how strong the prototype is dissimilar from the side it points away from. The most likely cluster center is pointed to. Very short arrows indicate that the prototype vector itself is a cluster center.

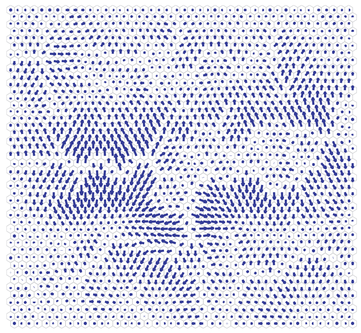


Resulting vector

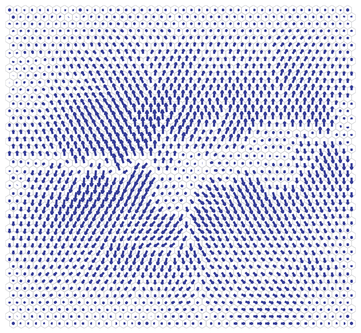
The visualization shows clustering structures at various levels of detail, depending on the choice of σ. Low values show the local cluster boundaries, high values the global structure.



σ = 2



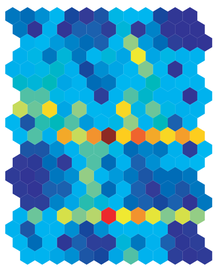
σ = 5



σ = 15

The Data Set

We apply this method in an industrial project analyzing sensor data from gas and oil wells. The Fracture Optimization Data Set consists of 199 samples in 10 dimensions. Each sample represents a gas well. The variables are various measurements, either geological ones ("geo"), parameters set by engineers ("param"), and performance indicators ("output"). The variables are numerical except for "Formation Type" and "Proppant Type", which are categorical.
The SOM trained on this data set consists of 70 map nodes (7 x 10). The U-Matrix is shown on the left. There are mainly two horizontal cluster boundaries.
Inspection of the component planes shows that some of the variables are highly correlated. The most important dimensions ("Produced Gas" and "Stimulation Costs") show that the most desirable position for the wells are in the upper left corner, with low costs and high gas output.



U-Matrix

NetPay geo

Proppant Type param

Produced Gas output

Average Rate param

Proppant in Formation param

Formation Type geo

Stimulation Costs output

Average Pressure geo

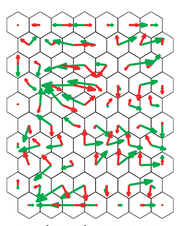
Total Volume param

Pad Fluid Volume param

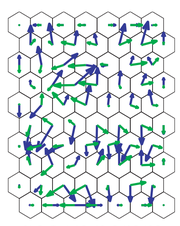
Groups of component planes

To gain additional insight into the clustering structure and the correlation between variables, component planes can be aggregated and visualized with our method. We compute a vector field for two groups and plot them simultaneously with different colors for direct comparison. Groups can be either established by merging variables that belong together in a semantic way (e.g. geological dimensions as described above) or by selecting similar component planes (with high positive or negative correlation).

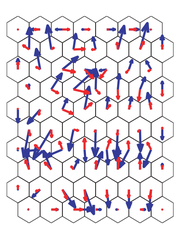
The fracture optimization data set has three groups of variables ("param", "geo" and "output") as shown on the right. The "param" variables are of particular practical importance. The arrows run mostly vertically, thus fine-tuning the engineering parameters will result in shifting a well's position up or down. Moving up is very beneficial in this case, since this corresponds to low costs and high production. The geological variables, determinded by where the well is positioned on the gas field, influences its horizontal position on the SOM lattice. The output parameters' vector field indicates by what amount the well's output will change if its position is changed on the map (e.g. by fine-tuning the engineering parameters).



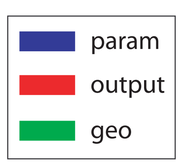
output vs geo



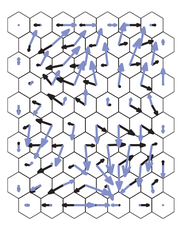
geo vs param



output vs param



Another option for variable groups is to combine highly correlated variables. We combined "NetPay" and "Formation Type" (the two leftmost component planes above) and "Average Rate", "Proppant in Formation", "Pad Fluid Volume" and "Total Volume" (the four rightmost ones). The resulting visualization shows a decomposition of the clustering structure. When compared to the U-Matrix, the two main horizontal boundaries are covered by the latter group.



Presented at WSOM'05 (Workshop on Self-Organizing Maps) in Paris, September 5-8, 2005