SOS2025 Experiment Report - Group 21

Martin Kowarik*
TU Wien
Austria

Matthias Frenzl†
TU Wien
Austria

Abstract

This report documents the machine learning experiment for Group 21, following the CRISP-DM process model. We examine the dataset "A hundred Plant Shape" using Self-Organizing Maps (SOM) to identify latent clusters of participants and analyze the sensitivity of the model to various hyperparameter configurations.

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

CRISP-DM, Provenance, Knowledge Graph, Machine Learning, SOM

ACM Reference Format:
Martin Kowarik and Matthias Frenzl. 2026. SOS2025 Experiment Report - Group 21. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Business Understanding

1.1 Data Source and Scenario

We use the **A hundred Plant Shape** provided in local ARFF files.

1.2 Business Objectives

Our primary objectives are:

Analyse the data and train several SOMs according to the assignment

2 Data Understanding

Dataset Description: The dataset contains mixed types of attributes (Integer, Real, Nominal) representing user traits and preferences.

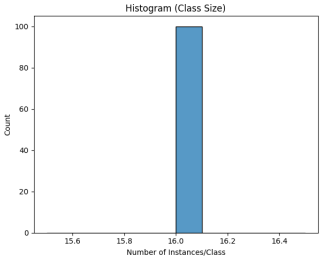
The following features were identified and selected for the SOM:

Number of instances & Attributes:
Number of instances: 1600
Number of attributes: 65

Attributes
float: 64
int: 1
Number of missing values: 0

Value Ranges

	min	max	mean	std	range
V1	0.000168	0.002390	0.000737	0.000270	0.002222
V2	0.000182	0.002247	0.000715	0.000265	0.002065
V3	0.000148	0.002112	0.000690	0.000258	0.001964
V4	0.000104	0.001998	0.000667	0.000252	0.001894
V5	0.000120	0.002151	0.000646	0.000250	0.002031



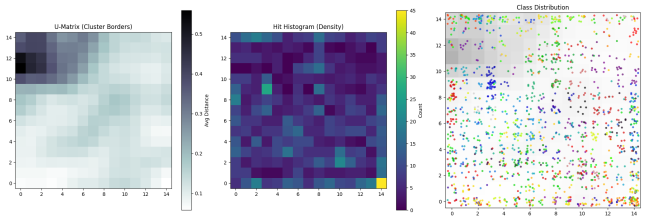
3 Data Preparation

Not needed.

3.1 Data Cleaning

Not needed.

4 Task 1



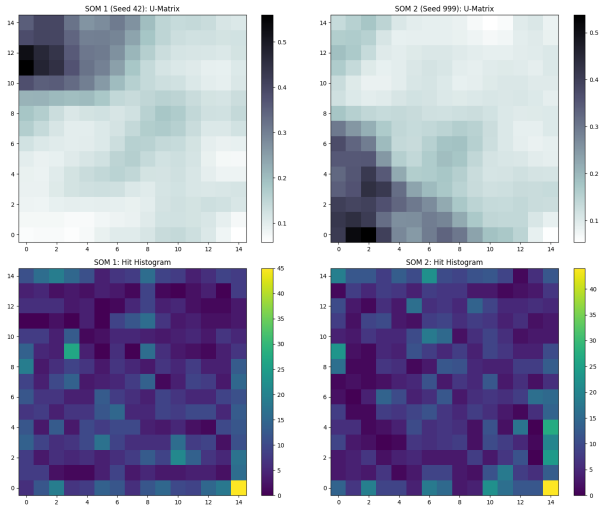
Neurons: Empty=9, Pure=26, Mixed=190
Classes: Coherent (1 neuron)=0, Split (>1 neuron)=100

*Student A, Matr.Nr.: 01634081
†Student B, Matr.Nr.: 00753306

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

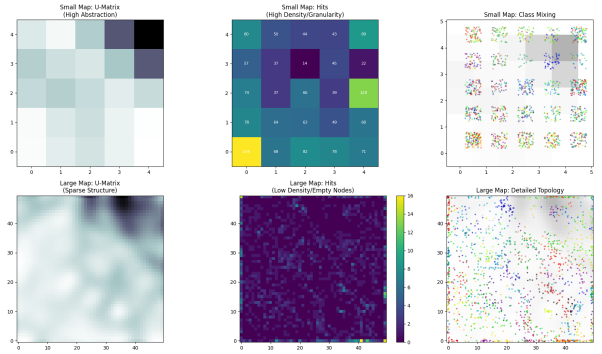
5 Task 2



Comparative Metrics
SOM 1 (Seed 42) | QE: 0.2566 | TE: 0.0150
SOM 2 (Seed 999) | QE: 0.2498 | TE: 0.0144

Cluster Stability Check (Nearest Neighbors)
Class 1 Neighbors -> SOM 1: [np.int64(96), np.int64(95), np.int64(6)], SOM 2: [np.int64(96), np.int64(95), np.int64(6)] | Overlap: 3/3
Class 50 Neighbors -> SOM 1: [np.int64(51), np.int64(81), np.int64(44)], SOM 2: [np.int64(99), np.int64(94), np.int64(74)] | Overlap: 0/3
Class 100 Neighbors -> SOM 1: [np.int64(27), np.int64(20), np.int64(60)], SOM 2: [np.int64(20), np.int64(27), np.int64(60)] | Overlap: 3/3

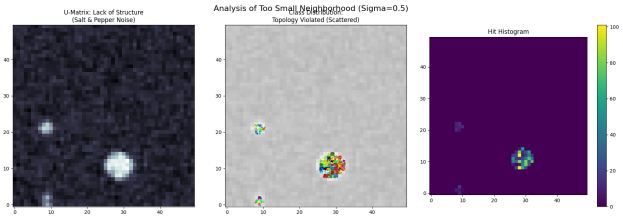
6 Task 3



Granularity & Magnification Analysis
Small Map Empty Neurons: 0/25
Large Map Empty Neurons: 1557/2500

Findings:
- Small Map: Super-clusters are created. High quantization error leads to loss of detail.
- Large Map: The dead (empty) neurons represent the empty space in the manifold.
- Large map: The light areas in U-Matrix (i.e. ridges are much wider -> means that are the separation between clusters.

7 Task 4



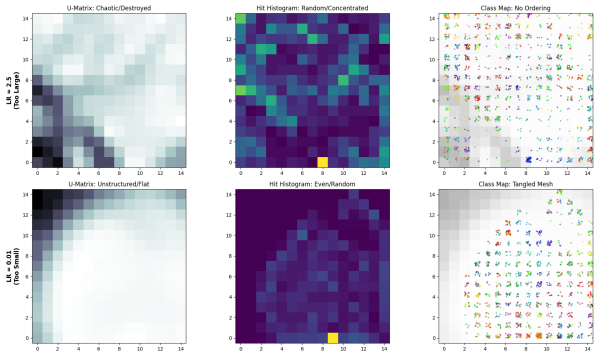
Findings a) Cluster Structure:
- The U-Matrix looks like static noise. There are no smooth valleys or ridges.
- Possible reason why: Adjacent neurons did not learn together. Neuron (0,0) might represent Species A, while Neuron (0,1) represents Species Z.

b) Quantization Error (QE):
- QE is 0.2582.
- Possible reason why: Individual neurons became very good prototypes for specific data points because they stick with their neighbors.

c) Topology Violations
- TE is 0.6956.
- Seems to be a disordered' map

d) Comparison to Correct Map: - Good Map: Classes are grouped. TE should be low.
- Bad Map: Classes are scattered. TE is high.

8 Task 5

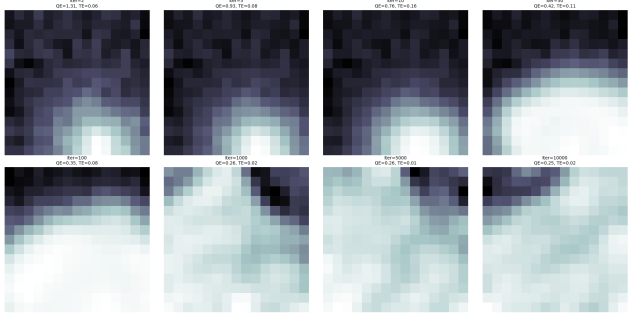


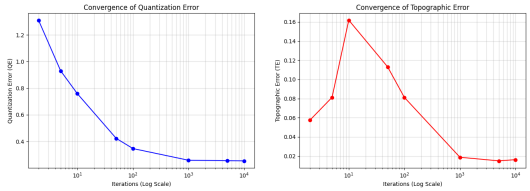
Findings
(I) Too Large Learning Rate (2.5):
a) Cluster Structure: The map looks random noise.
b) QE: Very High. The prototypes keep jumping over the data points.
c) TE: Very High. It seems that the pdates are breaking the neighborhood links.
d) LR is likely too high.

(II) Too Small Learning Rate (0.01):

a) Cluster Structure: Looks like random initialization.
b) QE: High. The neurons havn't moved far enough to make for a proper distribution.
c) TE: High. The map is still not good enough

9 Task 6





Findings

a) Structure Emergence

- Iterations 2-10: The map is random. QE is high, TE is high.
- Iterations 50-100: Global Ordering.
- Structure emerges when number of iterations > Number of Neurons.

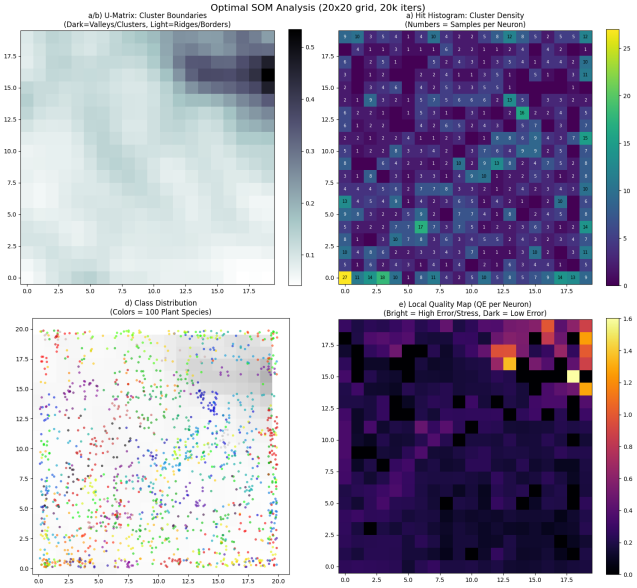
b) Stabilization:

- Iterations 1000: Major clusters become visible. The map is ordered, and TE goes down.
- Iterations 5000-10000: Boundaries become sharper.
- Stability is reached when the learning rate and sigma have decayed.

c) Quality Measures & Stability:

- QE (Quantization Error): Decreases rapidly and later asymptotes.
- TE (Topographic Error): Drops towards zero.
- Seems to overfit.

10 Task 7



Findings

a. Cluster Densities & Cardinalities (Hit Histogram):

- Findings: The Hit Histogram shows a non-uniform distribution. There are hotspots (green), and cold spots (purple) .
- Shapes: The clusters are not circular
- The U-Matrix might show a smooth valley, the Hit Histogram shows, that the data points often are rather at the edges.

b. Hierarchical Relationships (U-Matrix):

- Findings: There are Super-Clusters. There are large dark regions (valleys) and high light walls (i.e., ridges).
- Structure: The large valleys, have smaller ridges. This indicates a Hierarchy.
- Similarity: Adjacent clusters in the U-Matrix are similar.

c. Topological Relations & Violations:

- Global TE is low, so the map is largely unfolded.
- Violations:Yes, there are some

d. Class Distribution:

- Separation: Some classes look like isolated tight knots (=Homogeneous Clusters).
- Overlap: In the center, we see a mix of colors.
- Sub-clusters: Some classes are split obvioulsy into two distinct groups. Could be multi-modal data.

e. Map Quality (QE Map):

- Homogeneity: not much homogeneous.