# Analytic Comparison of Self-Organising Maps

Rudolf Mayer[1], Robert Neumayer[2], Doris Baum[3], and Andreas Rauber[1]

[1] Vienna University of Technology, Austria
[2] Norwegian University of Science and Technology, Norway
[3] Fraunhofer Institute for Intelligent Analysis and Information Systems, Germany

**Abstract.** SOMs have proven to be a very powerful tool for data analysis. However, comparing multiple SOMs trained on the same data set using different parameters or initialisations is still a difficult task. In most cases it is performed only via visual inspection or by utilising one of a range of quality measures to compare vector quantisation or topology preservation characteristics of the maps. Yet, comparing SOMs systematically is both necessary as well as a powerful tool to further analyse data: necessary, because it may help to pick the most suitable SOM out of different training runs; a powerful tool because it allows analysing mapping stabilities across a range of parameter settings. In this paper we present an analytic approach to compare multiple SOMs trained on the same data set. Analysis of output space mapping, supported by a set of visualisations, reveals data co-locations and shifts on pairs of SOMs, considering both different neighbourhood sizes at source and target maps. A similar concept of mutual distances and relationships can be analysed at a cluster level. Finally, Comparisons aggregated automatically across several SOMs are strong indicators for strength and stability of mappings.

## 1 Introduction

Self-Organising Maps (SOMs) enjoy high popularity in various data analysis applications. Experimenting with SOMs of different sizes, initialisations or different values for other parameters, is an essential part of this analysis process. In many cases, users want to detect the influence of certain parameters or generally want more details about the relations and differences between input data and resultant clusters across these varying maps. In this paper we thus propose a method to compare two or more SOMs, indicating the differences in how the data was mapped on either of the SOMs. We introduce three quality measures with supporting visualisations for comparing multiple SOMs. Its remainder is structured as follows. Section 2 describes related work in the field of SOM quality measures and comparisons. Section 3 then describes three types of analysis, which are illustrated along with experimental results in Section 4. In Section 5 we draw conclusions and give an outlook on future work.

J.C. Príncipe and R. Miikkulainen (Eds.): WSOM 2009, LNCS 5629, pp. 182–190, 2009.

## 2 Quality Measures for and Comparison of SOMs

A range of measures have been described for assessing the quality of either a SOM's quantisation, projection, or both; an overview is given in [7]. The probably best known *quantisation* measure is the *Quantisation Error*, which sums the distances between the input vectors and their best matching unit (BMU). Among the measures assessing the *projection* quality, the *Topographic Error* increases an error value if the BMU and the second BMU of an input vector are not adjacent to each other on the map. The normalised sum over all local errors is used as a global error value of a given map. The *Topographic Product* [1] measures for each unit whether its $k$ nearest neighbour units coincide, regardless of their order, by assessing the distances of the model vectors in the input and output space. Its result indicates whether the dimensionality of the output space is too large or too small. The *Neighbourhood Preservation* [8] measure is similar to the Topographic Product, but operates on the input data. Additionally [8] introduces *Trustworthiness*, measuring whether the k-nearest neighbours of data vectors in the output space are also close to each other in the input space. It thus gives and indication of the expressiveness and reliability of a given mapping.

Only limited research has been reported for comparing two or more SOMs with each other. An analysis of different distance measures for a supervised version of the SOM and it's application to the classification of rail defects, for example, is studied in [2]. Quality measures for the evaluation of data distribution across maps trained on multi-modal data are explored in [5], where the effect of multiple modalities is shown by the example of song lyrics and acoustic features for audio files. Both types of features are used for the same collection and the resultant map is compared according to spreading features. These help to identify musical genres with respect to their homogeneity in both dimensions. Analysis of different map sizes or other parameter variations are not considered. *Aligned Self-organising Maps* [6] are composed of several SOMs which are trained on the same data with differently weighted features, with the aim of exploring the impact of these differences on the resultant mappings. The maps are aligned as layers in a stack, and a distance measure is defined between stacks for comparison of units across layers. This measure is then used analogous to the distance between units on one layer to preserve the topology across the stack. The Aligned SOMs changes the SOM training algorithm so that each data vector is mapped onto a similar position also in the vertically stacked SOMs. However, this method can not be applied to maps with different sizes.

## 3 Analysing Data Shifts and Co-locations

The following methods allow comparisons of two or more SOMs trained on the same data set. The parameters for the SOM training such as the size of the map, the neighbourhood function, or the learning rate, can differ. Herein lies the strength of these visualisations, namely to compare differences in these parameters or of SOMs trained with identical parameters but different (random)

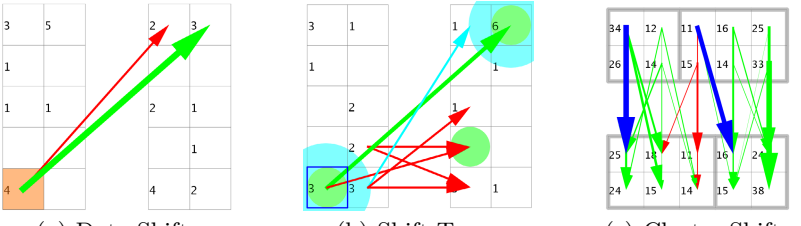
initialisations of the model vectors, with respect to distributions in the output space. All the methods proposed below rely on one *source map* and one or more *target maps* the source is compared to. The resulting description may either be visualised by colour-coding the units, on the *source map*, or actually displaying the detailed components of the resulting measurement using the source and target maps. In order to compare SOMs of radically different sizes, all methods make use of a neighbourhood definition in both the source and target maps.

### 3.1 Data Shifts Analysis

This method analyses and displays changes in the position of co-located data across multiple maps. For a given vector, it shows the position of the other vectors mapped onto the same unit (or within a given source neighbourhood) on a target map. This can be used to find out how stable the mapping is, and how steadily a data vector is put into a neighbourhood of other vectors on different SOMs. Put more abstractly, it measures how much of the data topology on the map really is caused by attributes of the data, and how much of it is simply an effect of different SOM parameters or initialisations, i.e. is caused by differences in parameter settings and training process.

An introductory example for the Data Shifts Visualisation is given in Figure 1(a). The figure shows positions of data vectors between two maps in terms of data and cluster shifts. The SOMs in Figure 1(a) are visualised by the two rectangular grids (each square represents a unit of the SOM and the numbers indicate the number of instances mapped to the respective unit). The arrows show the movement of the four vectors lying on the lower left unit of the left map. Three out of four vectors move to the unit of the right map pointed to by the thick arrow.

The data shifts and their types can be formalised as follows: Let $r_1$ and $r_2$ be the radii of the source and target neighbourhoods, and let $d_1$ and $d_2$ be the distance functions in the output space of the two SOMs. Let $c_s$ be the stable count threshold and $c_o$ be the outlier count threshold, which can be adjusted to ignore shifts concerning only "few" vectors, and to define what "few" means. With $x_i$



(a) Data Shifts (b) Shift Types (c) Cluster Shifts

**Fig. 1.** Positioning of data vectors across different SOMs 1(a). 1(b) shows all types of shifts, and neighbourhood radii. The movement of clusters is shown in 1(c). The arrows denote the movement of data vectors/clusters, respectively.

denoting the data vector in question, its source and target neighbourhoods $U_{1i}$ and $U_{2i}$ contain other data vectors as follows:

$$U_{1i} = \{x_j | d_1(x_j, x_i) \leq r_1\}, \qquad U_{2i} = \{x_j | d_2(x_j, x_i) \leq r_2\} \tag{1}$$

The set of neighbours that are in both neighbourhoods, $S_i$ can easily be found, as well as the set of vectors that are neighbours in the first SOM but not in the second, $O_i$:

$$S_i = U_{1i} \cap U_{2i}, \qquad O_i = U_{1i} \setminus U_{2i} \tag{2}$$

The input vector $x_i$'s data shift is stable for a given absolute threshold if $|S_i| \geq c_s$, or if $\frac{|S_i|}{|U_{1i}|} \geq c_s$ in the case of a relative threshold.

If the data shift is not a stable shift, it is an adjacent shift if there is another data vector $x_s$ whose data shift is stable and it lies within the neighbourhood radii.

$$d_1(x_i, x_s) \leq r_1 \wedge d_2(x_i, x_s) \leq r_2. \tag{3}$$

Finally, if the shift is neither stable nor adjacent, it is an outlier shift if $|O_i| \geq c_o$ in case of absolute, and $\frac{|O_i|}{|U_{1i}|} \geq c_o$ for relative count threshold values.

Figure 1(b) illustrates all types of shifts, i.e. stable, adjacent and outlier shifts, by green, cyan and red arrows, respectively. The circles indicate the neighbourhood for determining the neighbour count (green) and the adjacent shifts (cyan).

### 3.2 Cluster Shifts Analysis

The Cluster Shifts Analysis is conceptionally similar to the Data Shifts Analysis but compares SOMs on a more aggregate level, by comparing clusters in the SOM instead of singular units or neighbourhoods. Thus, we first employ Ward's linkage clustering [4] on the SOM units, to compute the same (user-adjustable) number of clusters for both SOMs. The clusters found in both SOMs are linked to each other, determined by the highest matching number of data points for pairs of clusters on both maps – the more data vectors from cluster $A_i$ in the first SOM are mapped into cluster $B_j$ in the second SOM, the higher the confidence $p_{ij}$ that the two clusters correspond to each other. This can be formalised as follows: let the set $M_{ij}$ contain all data vectors $x$ which are mapped onto the units in $A_i$ and in $B_j$. To compute the confidence $p_{ij}$ that $A_i$ should be assigned to $B_j$, the cardinality of $M_{ij}$ is divided by the cardinality of $A_i$.

$$M_{ij} = \{x | x \in A_i \wedge x \in B_j\}, \qquad p_{ij} = \frac{\mid M_{ij} \mid}{\mid A_i \mid} \tag{4}$$

We then compute all pairwise confidence values between all clusters $C_i$ in the maps. Finally, they are sorted and we repeatedly select the match with the highest values, until all clusters have been assigned exactly once. When the matching is determined, the visualisation can easily be created, analogously to the Visualisation of the Data Shifts. An example is depicted in Figure 1(c), which shows a map trained on synthetic data of two slightly overlapping Gaussian

clusters. The number of clusters to find was set to two. The cluster mappings are indicated by blue arrows, whose thickness corresponds to the confidence of the match. Data vectors which move from a cluster in the first SOM to the matched cluster in the second SOM are considered 'stable' shifts', and indicated with green arrows; the red arrows represent 'outlier' shifts into other clusters.

### 3.3 Multi-SOM Comparison Analysis

While the previous two methods focus on a pair-wise comparison, the Multi-SOM Comparison Analysis can be used to compare multiple SOMs trained on the same data set. Its main focus is one specific SOM, the 'source SOM', to be compared against a number of other maps. More precisely, the visualisation colours each unit in the main SOM according to the average pairwise distance between the unit's mapped data vectors in the other $s$ SOMs. To this end, we find all $k$ possible pairs of the data vectors on $u$, and compute the distances $d_{ij}$ of the pair's positions in the other SOMs. These distances are then summed and averaged over the number of pairs and the number of compared SOMs, respectively. The mean pairwise distance $v_u$ of unit $u$ is thus calculated as follows:

$$v_u = \frac{\sum_{j=1}^{s} \frac{\sum_{i=1}^{k} d_{ji}}{k}}{s} \tag{5}$$

Similarly, the computation of the variance $w_u$ is defined as:

$$w_u = \frac{\sum_{j=1}^{s} \frac{\sum_{i=1}^{k} d_{ji}^2}{k}}{s} - {v_u}^2 \tag{6}$$

where $d_{ji}$ denotes the distance between the vectors of pair $i$ in the output space of SOM $j$.

When applied to the cluster based evaluation, we use the single linkage distance between the respective clusters $r$ and $s$ and their cluster members $x_{ri}$ and $x_{sj}$ as follows:

$$d^{SL}(r,s) = min(d(x_{ri}, x_{sj})) \tag{7}$$

Herein, the distance between two clusters is defined as the minimum distance between any of their respective members. In our case, we use unit coordinates of clusters in the SOMs as the features describing them. As a result of the computations described in this section, we obtain quality measures for single units with respect to the mapping of their data vectors on other SOMs.

## 4 Experiments

We present two sets of experiments, first with an artificial data set tailored to specific challenges in data mining, and then with the Iris benchmark data set.

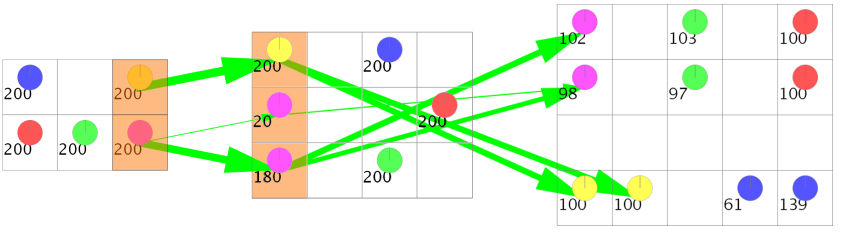



**Fig. 2.** Data shifts of the multi-challenge data set across three SOMs of different sizes trained on the same data

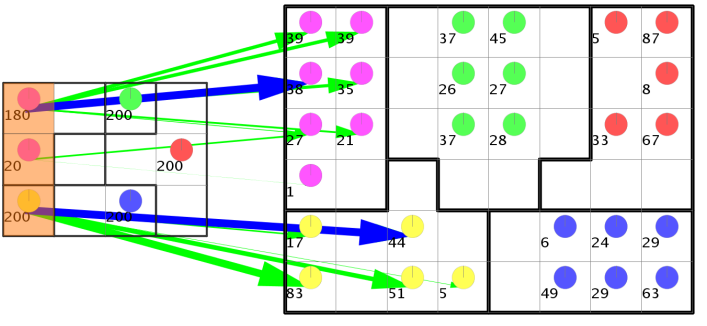


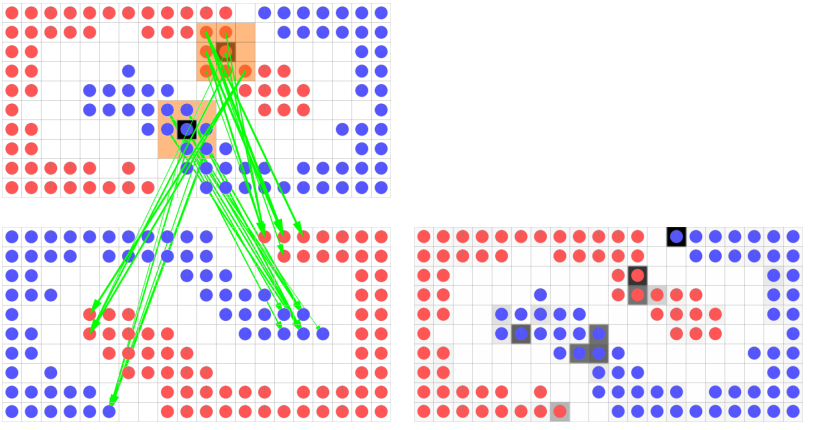

**Fig. 3.** Cluster shifts of the multi-challenge data set for two SOMs of varying sizes

### 4.1 Artificial 'Multi-challenge' Data Set

We created a 10-dimensional synthetic data set, which is used to demonstrate how a data analysis method deals with clusters of different densities and shapes when these different characteristics are present in the same data set [1]. It consists of five sub-sets, four of which live in a three-dimensional space. The subsets themselves are composed of several clusters, thus in total we have 14 distinguishable patterns of data. The first subset consists of one Gaussian cluster, and another cluster formed of three Gaussians, all of which are well separated. The second subset consists of two overlapping, three-dimensional Gaussians, while the third set is similar, but of ten dimensions. The fourth subset is the well-known *chainlink* problem of two intertwined rings. Finally, the fifth subset is sampled along a curve that consists of four lines that are patched together at their endpoints.

Figure 2 illustrates three different map sizes trained on this data set, and shows how the clusters slowly separate into their sub-clusters they are composed of, when the map size increases. In the middle illustration, even with doubling the number of units, only one cluster splits into two sub-clusters; finally, in the right image, all clusters have split on two different units. Figure 3 shows the cluster shifts for three selected clusters from a smaller map with twelve units

[1] The data set is available at `http://www.ifs.tuwien.ac.at/dm/`



(a) Data shifts (b) Variance Comparison

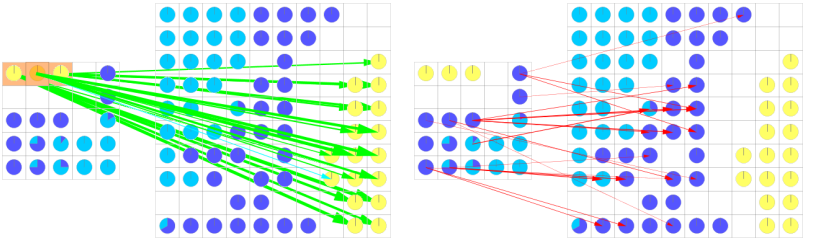
**Fig. 4.** Data Shifts and Variance Analysis on the Chainlink data set

to a bigger map with 48 units. The clusters are identical on both maps, thus with a confidence of 100% each. It is, however, interesting to note that for the cluster arranged in the top-left corner of both maps, the initial separation on the smaller map does not prevail any more on the bigger map. Thus, the initial assumption that could be drawn from the smaller map, namely that the items found on the two units are clearly separable, could be refuted.

Figure 4 illustrates one specific subset, the *Chainlink* problem, for which it is known that it cannot be projected to a two-dimensional space without severely breaching the topology. The two rings are indicated by red and blue colour, respectively. It can be well observed from the visualisation of the Data Shifts in 4(a) that even though the projection looks very similar in both cases, the breaking points in the two rings are actually different in the two maps. Further, the illustration also depicts the mean values of the Multi-SOM comparison, evaluated across eight target SOMs trained with different initialisation and iteration parameters, with two nodes having high pairwise distances, and thus colour black. Figure 4(b) shows the distance variance of the same map. It can be noted that with this measure, we find a higher number of possible breaching points than we were able to detect with the mean pairwise distance only. The intensity of the grey-shade used denotes a higher variance of the distance in the different SOMs, and thus indicates dislocations of vectors, which in this case reveal the topology breach, with the black-filled units marking the points with the highest probability.

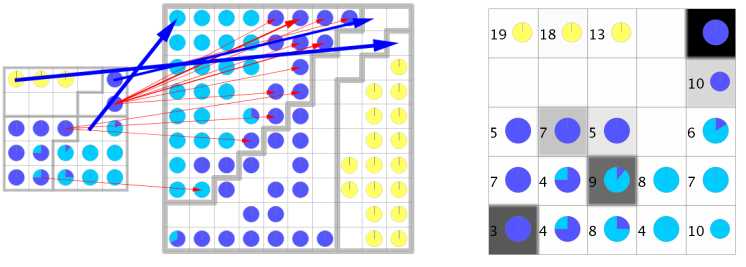
### 4.2 Iris Data Set

Finally, we performed experiments on the benchmark data set Iris [3]. Two maps were trained, with 25 and 100 units, respectively. The three different classes in the data set are marked with yellow (setosa), dark blue (versicolor) and light



(a) Setosa (b) Versicolor

**Fig. 5.** Data Shifts: stable shifts from *Setosa* (a), outlier shifts from *Versicolor* (b)




(a) Cluster Shifts (b) Comparison – Mean

**Fig. 6.** Iris: Cluster Shifts with three clusters and outliers (a), mean comparison (b)

blue (virginica). In Figure 5(a), we can see the data shifts from the setosa class. The topology-preservation ability of the SOM can be easily observed: the vectors from the rightmost setosa unit in the left map are mapped onto the top of the elongated setosa area in the large map, the vectors from the middle unit onto the middle of the elongated area, and the vectors on the leftmost unit are mapped onto the bottom of the elongated area. The borders of the virginica area and the versicolor area, however, are not as cohesive and spread over a wider area than the border between the other two classes. Figure 5(b) shows only the outlier shifts for the data shifts visualisation of the two SOMs. Most of the outlier shifts emerge from units in the versicolor area or the border of the virginica area.

Figure 6(a) shows a Cluster Shifts Visualisation based on three clusters. The setosa cluster is clearly separated from the others, and its mapping has 100% confidence. The other two clusters each represent one of the other two classes in the small map. In the large map the virginica cluster gets assigned quite a few versicolor samples as well. These show up as the outlier shifts drawn in in red. The virginica cluster match confidence is 100%, the versicolor clusters' confidence is only 69%.

Finally, a Multi-SOM comparison was used to find the units in the smaller SOM where the projection onto the two-dimensional SOM-grid is unstable, which is visualised in Figure 6(b). The minimum pairwise distance threshold was set to 2.5, to reduce the impact of the bigger size of the larger SOM – the data

vectors spread over more units in the larger SOM, thus the vectors that lie on one unit in the small SOM will spread over a couple of neighbouring units in the large SOM. This would distort the conclusions we wish to draw from the visualisation, therefore the threshold is used to compensate for the difference in size. The units with the high mean pairwise distances (marked in shades of grey) all are either on the border between the versicolor and virginica classes or within the versicolor class. This points to the relative instability of the projection of the versicolor class onto the SOM-grid: data vectors from the versicolor class are projected differently in both SOMs. Yet again, these results suggest that the setosa class and to some extend the core of the virginica class are well-defined and distinct, while the border between virginica and versicolor and versicolor class itself are a relatively unstable area in a SOM projection. Thus, the results from the three visualisations support and reinforce each other.

## 5 Conclusions and Future Work

In this paper, we presented methods to analytically compare two or more SOMs with each other, and showed the feasibility of the approach on two data sets. Due to space limitations, we could not present experiments on further data sets and had to limit the level of detail in our experiment discussion; more details are availbale at http://www.ifs.tuwien.ac.at/dm/. Future work includes more extensive experiments to provide evidence for certain types of shifts and violations, to eventually automate the process of SOM interpretation, as well as for automatically setting useful threshold and analysis neighbourhood parameters.

## References

1. Bauer, H.-U., Pawelzik, K.R.: Quantifying the neighborhood preservation of self-organizing feature maps. Trans. Neural Networks 3(4), 570–579 (1992)
2. Fessant, F., Aknin, P., Oukhellou, L., Midenet, S.: Comparison of supervised self-organizing maps using euclidian or mahalanobis distance in classification context. In: Mira, J., Prieto, A.G. (eds.) IWANN 2001. LNCS, vol. 2084, pp. 637–644. Springer, Heidelberg (2001)
3. Fisher, R.A.: The use of multiple measurements in taxonomic problems. In: Annual Eugenics, Part II, vol. 7, pp. 179–188 (1936)
4. Ward Jr., J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)
5. Neumayer, R., Rauber, A.: Multi-modal music information retrieval - visualisation and evaluation of clusterings by both audio and lyrics. In: Proc. 8th Conf. Recherche d'Information Assistée par Ordinateur (RIAO 2007) (2007)
6. Pampalk, E.: Aligned self-organizing maps. In: Proc. 4th Workshop on Self-Organizing Maps (WSOM 2003), pp. 185–190 (2003)
7. Pölzlbauer, G.: Survey and comparison of quality measures for self-organizing maps. In: Proc. 5th Workshop on Data Analysis (WDA 2004), pp. 67–82 (2004)
8. Venna, J., Kaski, S.: Neighborhood preservation in nonlinear projection methods: An experimental study. In: Dorffner, G., Bischof, H., Hornik, K. (eds.) ICANN 2001. LNCS, vol. 2130, pp. 485–491. Springer, Heidelberg (2001)