Multi-Scale Visual Quality Assessment for Cluster Analysis with Self-Organizing Maps

Jürgen Bernard, Tatiana von Landesberger, Sebastian Bremm and Tobias Schreck

Interactive Graphics Systems Group, Technische Universität Darmstadt, Germany

ABSTRACT

Cluster analysis is an important data mining technique for analyzing large amounts of data, reducing many objects to a limited number of clusters. Cluster visualization techniques aim at supporting the user in better understanding the characteristics and relationships among the found clusters. While promising approaches to visual cluster analysis already exist, these usually fall short of incorporating the *quality* of the obtained clustering results. However, due to the nature of the clustering process, quality plays an important aspect, as for most practical data sets, typically many different clusterings are possible. Being aware of clustering quality is important to judge the expressiveness of a given cluster visualization, or to adjust the clustering process with refined parameters, among others.

In this work, we present an encompassing suite of visual tools for quality assessment of an important visual cluster algorithm, namely, the Self-Organizing Map (SOM) technique. We define, measure, and visualize the notion of SOM cluster quality along a hierarchy of cluster abstractions. The quality abstractions range from simple scalar-valued quality scores up to the structural comparison of a given SOM clustering with output of additional supportive clustering methods. The suite of methods allows the user to assess the SOM quality on the appropriate abstraction level, and arrive at improved clustering results. We implement our tools in an integrated system, apply it on experimental data sets, and show its applicability.

Keywords: Visual Cluster Analysis; Self-Organizing Maps; Cluster Comparison; Quality Visualization and Assessment; Visual Analysis.

1. INTRODUCTION

In many application domains, huge data sets arise, which cannot be effectively analyzed or visualized in their full magnitude. Automatic cluster analysis techniques can help by reducing large data sets to more compact representations based on finding groups in the data. Given a cluster analysis output that represents the distribution of data well, it is much easier for the user to assess the distribution of data elements, and to make sense of it. While cluster analysis output can be studied in numeric or textual form, often visual representations of the cluster results are employed as a user-friendly way to access the results of the clustering process. Well-known techniques include dendrograms,[1] or projection of cluster prototypes to 2D diagram space by techniques such as Principal Component Analysis[2] or Sammon's Mapping.[3]

Clustering is not a purely automatic process, but requires user involvement at various stages of the process. Typically, the user needs to choose a class of clustering approaches (e.g., partitioning-, density-, hierarchic-, or network-oriented) and several parameters (e.g., number of clusters or outlier sensitivity). This causes the output of any given cluster analysis run to be uncertain regarding the quality of the obtained result. Assessing the quality of a given cluster result is important for the user to reflect the degree of validity of the interpretation derived from it, or to decide to rerun an analysis with changed parameter or method choice. However, current cluster visualization approaches typically fall short of visually reflecting the degree of clustering quality.

Further author information: (Send correspondence to Jürgen Bernard)
Jürgen Bernard - Mail: juergen.bernard@gris.informatik.tu-darmstadt.de, Phone: +49 6151 155666
Tatiana von Landesberger - Mail: tatiana.von_landesberger@gris.informatik.tu-darmstadt.de, Phone: +49 6151 155631
Sebastian Bremm - Mail: sebastian.bremm@gris.informatik.tu-darmstadt.de, Phone: +49 6151 155623
Tobias Schreck - Mail: tobias.schreck@gris.informatik.tu-darmstadt.de, Phone: +49 6151 155125

The contribution of this work is to systematically define and incorporate the notion of clustering quality into one popular visual cluster algorithm, namely, the Self-Organizing Map algorithm (SOM).[4] This method is well-known for its practical applicability and robustness with respect to data size and dimensionality, and it is directly suited for cluster visualization, due to its constraint to oragnize clusters on a regular grid. Similar to other clustering approaches, SOM analysis requires setting a considerable number of method parameters. In conjunction with the grid constraint implied in the method, the question of SOM output quality arises. Our basic idea is to define a hierarchy of abstractions for measuring the quality of SOM output, and to define SOM displays incorporating the corresponding measurements. Our hierarchy includes, from coarse to fine, quality measures observed on the global scale per SOM, on local scales per SOM unit, and per data sample. Furthermore, we introduce the novel *correspondence* visualization concept that maps the output of supportive clustering methods to the SOM visualization, effectively allowing to validate the SOM output by comparison with alternative clusterings. Our approach therefore supports multi-perspective quality view on the output of the SOM algorithm. While we implemented an encompassing set of quality views and supportive clustering methods, our approach can easily accomodate further quality views as modules in our system.

The remainder of this paper is structured as follows. In Section 2, we survey related work in cluster analysis and visualization. In Section 3, we derive a hierarchy of quality notions applicable to assess the quality of SOM clusterings. In Section 4, we describe quality measures on the global and per-unit abstraction. In Section 5, we introduce quality notions and a mapping scheme for assessing quality on the data-sample abstraction. In Section 6, we introduce the notion of quality based on supportive clustering comparisons. In Section 7, we apply the implementation of our methods on experimental data sets. Finally, Section 8 concludes and outlines future work in the area.

2. RELATED WORK

In this Section, we review clustering and cluster validity techniques. Subsequently, SOM-based visual cluster analysis methods are presented, and an insight in SOM-based cluster visualization applications is given.

2.1 Clustering and Cluster Validity

Clustering is commonly used to structure huge data sets by grouping objects into cluster such that entities within a cluster posses a high similarity on each other. Clustering is an unsupervised process without requiring previous knowledge about the data. Difficulties in the clustering domain include the selection of the most suitable clustering algorithm and respective parametrization, which is strongly data and application dependent. Up to now, a variety of clustering algorithms have been proposed, taxonomies can be found in.[5,6] Partitioning, density-based and artificial neural network algorithms are among the commonly used clustering techniques. One of the most prominent partitioning clustering algorithms is the k-means algorithm,[7] which divides a dataset into a number of k clusters. As a k-means cluster is represented by the mean of its member points, the algorithm may be negatively affected by outliers.[6] However, partitioning clustering algorithms provide good clustering results if the data set consists of hyperspherical cluster shapes. The DBScan algorithm[8] is a density-based clustering technique. This type of algorithm defines a cluster as a linked network of data points with a user-specified minimum density, thus, it is able to find arbitrarily shaped clusters. In return, data points in low density areas are not covered by the clustering procedure. The Self-organizing Map (SOM) algorithm[4] is a neural network clustering approach. The output of the method is a network of cluster prototype vectors connected to each other. The prototype network approximately preserves the topological structure of the input data space. Thus, SOMs have particular abilities for visual clustering, as the algorithm aligns the input data as topologically ordered groups on the grid of the output map structure.

As clustering is an unsupervised process to group datasets without previous knowledge, an effective evaluation of the results by the user is crucial. Typically, clustering is an iterative refinement process[5,9] in which parameters are modified to optimize clustering results, driven by the question which partition fits best to the given data set. Clusterings can be evaluated by measures for intra-cluster (internal) compactness and inter-cluster (external) separation. According to Halkidi,[10] internal, external and also relative cluster validity indices can be applied to measure the clustering quality. Relative cluster validity indices are suitable to compare clustering results with each other, and are appropriate for iterative refinement strategies. Commonly used cluster indices include

Dunn-like Indices,[11] the Davis Bouldin Index[12] and the Modified Hubert Statistic.[13] The quantization error is one of the most simple and generic cluster quality measures, in the SOM domain often used to measure the vector quantization quality.[4] In case of Self-Organizing Maps, as the algorithm also produces a cluster network structure, quality indices regarding topology are relevant as well. Measuring the SOM topology is a non trivial task, and an exact definition of topology preservation is argumentative. In this paper, we will rely on various topology evaluation measures, which we validated regarding to a three-level classification of topology preservation in.[14] Further general SOM clustering quality measures are surveyed.[14–16]

2.2 SOM Visualization and Applications

An overview of standard SOM-based visualization techniques is given by Vesanto.[17] Most often, a 2D grid network structure is assumed and visualized by mapping certain intra- and inter-cluster properties to visual structures including color, shape, and texture. One of the most important tasks in SOM visualization is supporting the identification of clusters on the grid. The U-Matrix[18] visualizes pairwise distances between SOM prototypes by color-coding, allowing to distinguish similar regions on the SOM map. Also, shape-based and vector-based SOM cluster visualizations have been introduced.[9,19] Further approaches using color to distinguish between clusters exist.[17] In addition, this cluster visualizations enable detecting topographic errors on the map, as SOM prototypes with similar vector attributes are colored with similar color, accordingly. Visualization techniques considering affiliations of the data on the SOM grid are available as well, a basic representative are the density matrices.[17] The S-Map[20] is an improved density-based data visualization techniques. Regarding topological aspects, graph-based visualizations can be applied, e.g., to indicate each units nearest arithmetic neighbor.[21]

The SOM cluster algorithm has been previously used successfully in many different application fields including text analysis,[22,23] multimedia retrieval[24] for geographic information science.[25] In,[26] the authors introduced a system or analysis of space and time dependent data with applications on crime rate and traffic data. The SOM method has also been applied for Image sorting and layout. By this, Image Sorter[27] provides an overview for large image collections. In,[28] visual cluster analysis in 2D time-dependent financial data by means of SOM was introduced.

3. INTERACTIVE SOM-BASED CLUSTER ANALYSIS AND HIERARCHY OF QUALITY NOTIONS

In this Section, we give an overview of our SOM-based visual cluster analysis system. Initially presented in,[28,29] in this work we extend the system by a number of quality and correspondence visualizations. Typically, SOM-based visual cluster analysis treats the SOM algorithm as a given black-box component in the system. However, due to the various possible parameterizations, it may be problematic to easily find appropriate parameterizations and thereby, suitable cluster results.[28] Our system provides visual support to supervise the SOM parameterization, showing the emerging of the SOM results as a function of algorithmic run time. This opens up interactive control of the training steps at user-selectable granularity, useful for arriving at cluster results suited for the user preferences and application needs. Based on specialized renderers accommodating individual data types such as trajectories, image data, and generic high-dimensional data, our system is applicable to various data types. Figure 1 shows an application of our system to trajectory data.

Our system offers a comprehensive tool set for detailed interactive steering of the training process. Having these possibilities, the question of the *quality* of the obtained clustering arises. Such quality assessments are important to guide the interactive parameterization by the user. We next describe a hierarchy of SOM-based quality notions, that will be supported by visual representations in our system. Defining clustering quality is task specific, and typically needs to balance certain aspects, including e.g., projection versus clustering, local versus global quality, or topology preservation versus vector quantization.[30] We distinguish four abstractions at which to evaluate and judge the resulting quality of a Self-Organizing Map output:

(1) Global quality measures. The quality of the SOM is measured by a single numeric quality index such as discussed in Section 2. A visualization of the development of these quality indexes during SOM training is performed by showing line charts (cf. Figure 3).

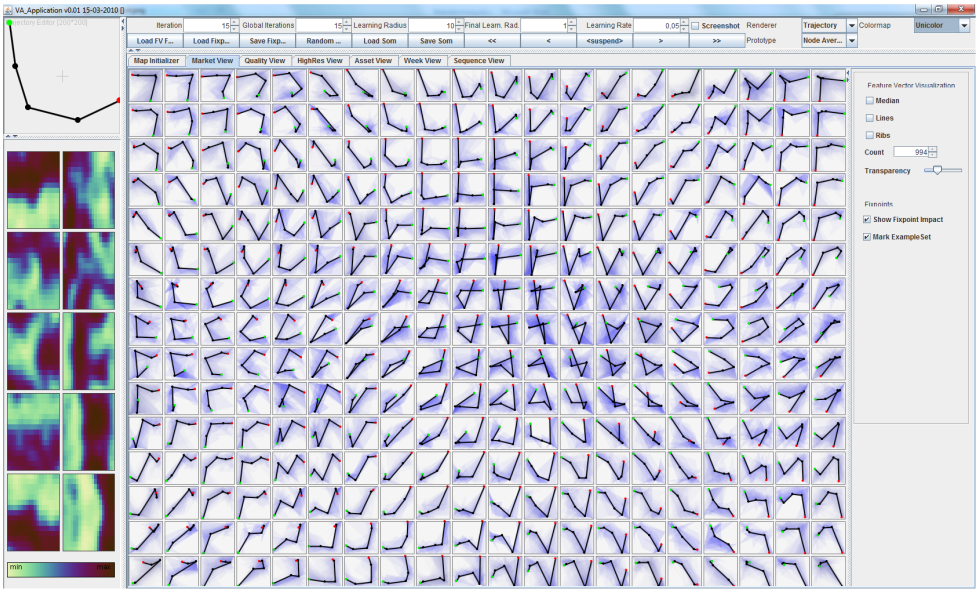

Figure 1. Screenshot of our interactive visual SOM analysis system. Via feature images, representants are shown for each SOM cluster (in this case, trajectory data ist shown). Controls to steer the clustering process are included.

(2) Unit-based quality measures. Quality measures for each prototype of the SOM (called a unit, or node) are considered. The SOM grid can be enhanced with color-coding, shape, or printing scores to support SOM unit-based quality assessment reading. Section 4 details our implemented unit-based quality visualizations.

(3) Data point-based based quality. Quality characteristics are measured for every data entity in the input data set. In Section 5, we will therefore introduce a mapping approach to find data-specific positions within the coarse SOM grid for each data point (our so-called HighResSOM grid) by means of interpolation. Then, the quality of representation of each data point by the SOM clustering can be visualized by using glyphs rendered at each respective position.

(4) Cluster correspondence view. This quality abstraction considers the overall comparison of the SOM-output with supportive clustering algorithms. By means of mapping the *correspondence* between clusters in the SOM on one hand, and the clusters of a given supportive cluster algorithm, the user can perform a global visual validity of the SOM output. Section 6 will detail our approach.

We propose an extended SOM-based analysis workflow, incorporating views to each of these four quality abstractions. Based on input data and user parameterization, an initial SOM is trained and its quality according to notions (1-3) is shown (output analysis). The output can be overlaid by specific correspondence visualizations (4). If the user is satisfied, manual annotation of the found clusters can take place. Alternatively, it is possible to jump back to any previous step, for refinement of the output. Figure 2 illustrates the workflow.

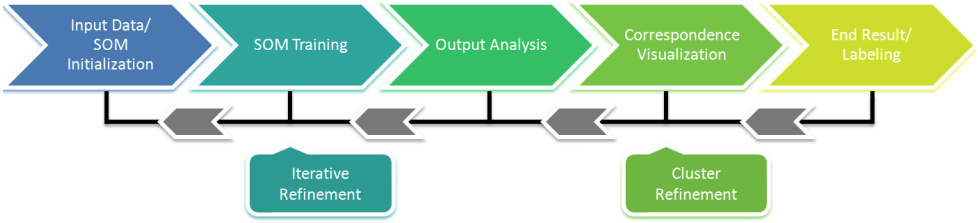

Figure 2. Proposed visual cluster analysis workflow.

4. GLOBAL AND UNIT-BASED SOM QUALITY VISUALIZATIONS

As pointed out, interactive SOM cluster analysis makes it necessary to consider quality assessments. This is important to evaluate the appropriateness of user parameters chosen, or to compare different clustering runs. According to our quality notion hierarchy presented in the last Section, we here present our implementations of views 1 (global quality measures) and 2 (per-unit quality measures), as introduced in the hierarchy in the previous Section.

4.1 Global SOM Quality Assessment

Characterizing SOM quality by a scalar global measure is a simple and straightforward approach that easily allows users to compare different SOM results. We implemented several scalar quality criteria proposed in the literature for monitoring the training process of a SOM in real time, and to compare final results. Both use cases are illustrated in Figure 3 where the development of two important SOM quality measures are shown during several sequential training stages by line charts. Each line describes the quality evolution of a single quality index during one SOM training run. The comparison of multiple lines (observations from more older training stages are faded out) allows an evaluation of the quality behavior and convergence over a number of training steps (called epochs in the SOM terminology[4]). The user can easily change SOM cluster analysis parameters, and observe on-the-fly the effects on the diverse quality criteria. For example, it is thereby easily and effectively possible to balance the topology and average vector quantization yielded by the respective results. At any time during the runtime of the SOM algorithm, the user is able to pause the run and change key parameters.

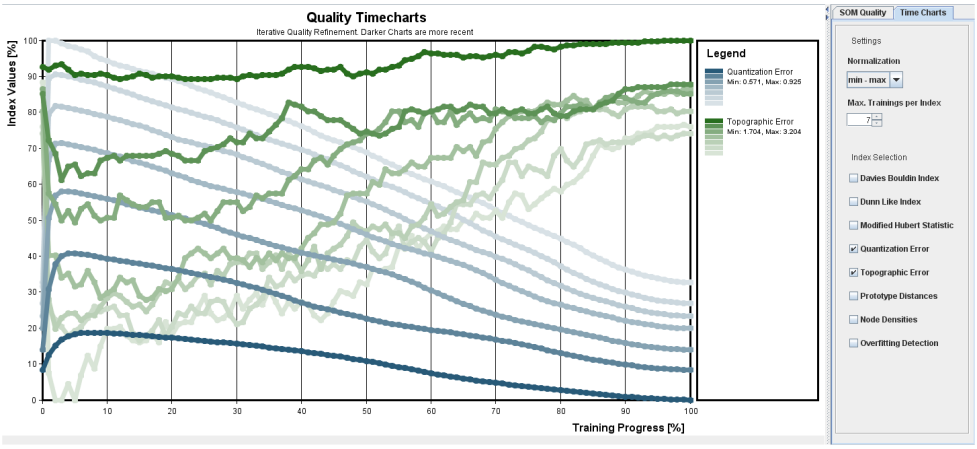

Figure 3. Scalar quality measures such as vector quantization error (here: blue lines) or a topology preservation measure (here: green lines) are shown by real-time line charts.

4.2 Unit-Based SOM Quality Assessment

Our unit-based measurements allow to visually assess quality aspects on the unit level. Recall that each SOM unit represents a cluster prototype, and can represent possibly many data samples. Important local quality visualizations are (1) the inter-unit distances (leading to so-called U-Matrix[18] visualizations), (2) the relative number of data elements mapped per unit (leading to so-called density visualizations),[20,31] and (3) per-unit quantization error (leading to so-called error visualizations). The visualizations are obtained by directly visualizing the measurements as color, and overlaying it over the SOM prototype grid. Additional visual quality measurements that consider the relation between individual units and the surrounding map include (4) RGB similarity colormaps[17] (showing clusters and topological orderings by assigning colors), and (5) a vector fields[19] visualization (showing for each unit, the area and strength of the most similar map areas, by vector direction and length, respectively). Finally, we implemented (6) topographic error connectors (showing similar but non-continuous areas of the map). A topographic error is given if the grid distance between best-matching and second-best matching SOM units of single data element is greater than a predefined threshhold. Quality assessments and SOM clusterings can be

improved by combining multiple visualization tools like an overlay of a colormap and a vector based visualization. Figure 4 shows examples of respective quality views.

We point out that these measurements and visualizations have been previously introduced elsewhere, and we use our own implementations or variants thereof. However, our system for the first time to the best of our knowledge, allows to interactively switch and combine from a large pool of implemented methods, and monitor them in real time during execution of the SOM algorithm. These visualizations make explicit information about the individual SOM-units and their relationships, and can be used to evaluate the overall quality of the SOM, or help to interactively determine the number of clusters. Note that the latter is often difficult in SOM analysis, as the SOM does not explicitly yield the number of clusters. Therefore, a case study will be given in Section 7, where we will show how we can combine the perception of individual visualizations from (1-6) to arrive a concluding assessment of cluster alignments. Furthermore in Section 7, the SOM clustering result will be improved by our new cluster correspondence visualization strategies (Section 5 and 6).

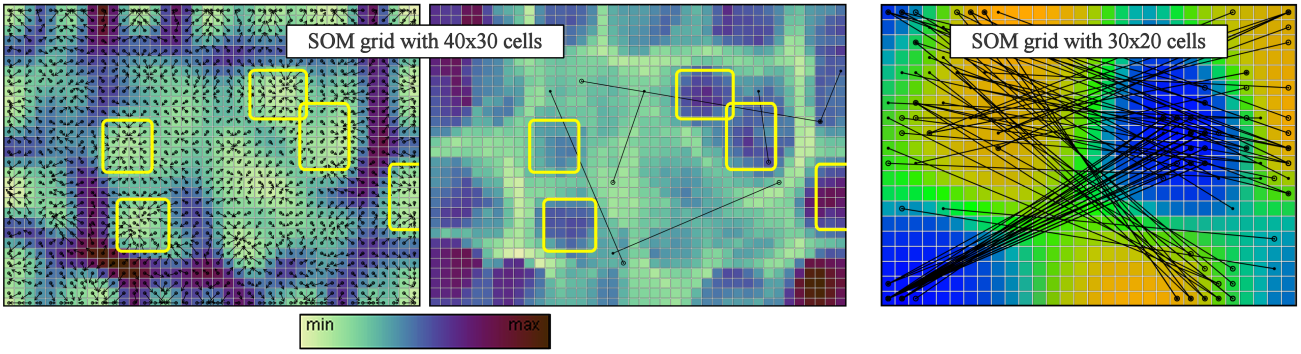

Figure 4. SOM local quality visualizations (numbers in brackets refer to description of respective visualizations in the section text). Left: the U-Matrix (1) visualization highlights SOM units forming clusters (bright colors). Additionally, arrows of the vector fields visualization show the SOM-units cluster affiliation (5). Center: the smoothed density map (S-MAP) (2), that points out SOM units with high density indicating a cluster association (dark colors). As it is indicated by yellow boxes, both colormap visualizations (U-Matrix and S-MAP) predominantly match in the indication of cluster borders, and cluster regions, respectively. The existence of only six topological error connectors (6) overlaid over the display indicates good topology preservation. Right: RGB similarity colormap (4) with topographic error connectors (6). This SOM visualization suggests that two clusters exist (blue and orange), and that this map shows topographic errors (foldings), indicated by topographic error connector (black lines).

5. SOM-BASED DATA CORRESPONDENCE VISUALIZATIONS

To focus on the per-data level for quality visualization, we require a mapping scheme to position individual samples on the SOM grid. We next describe such an approach for mapping data entities (including cluster prototypes from additional supportive algorithms) to the SOM grid. This mapping is the basis for data entity- and additional cluster-centric correspondence visualization.

5.1 Mapping Element Data to the SOM Grid

In typical SOM visualization systems, the granularity of the projection is limited by the SOM grid resolution. By calculating the best matching unit (BMU) for each data item, e.g., the SOM density histogram[17] gives information about the data dispersion along the SOM grid. Depending on the data type to be clustered, the local elements can be overlaid. In case of trajectories, this was done previously by means of opacity bundles (cf. Figure 1 for an example[29]). For arbitrary point-based data, this is not possible. We therefore, and to increase accuracy, we develop a more detailed mapping location of data points over the SOM in a continuous way.

By this, a 2D scatter plot-like SOM-based projection with screen resolution is obtained. Based on it, we can visualize the data points in correspondence with the SOM grid. This visualization naturally increases the precision of the density view, because the injective data projection method allocates individual screen coordinates per data point. Moreover, overlapping problems with respect to highly dense SOM units[17] are avoided. Another

benefit is the potential visualization of clusters in a scatter plot kind. By this, the restriction of a SOM unit being the smallest possible visualization unit is remedied. Our concept relates to hierarchical SOM approaches,[4,32] where the SOM grid consists of multiple layers with different resolutions. In contrast to these SOM-variants, our approach needs no re-training phase but directly works on a given SOM grid. We increase the resolution of the SOM by interpolation, yielding the so-called HighResSOM.[33] SOM units provide a set of support points for spline-based interpolation for calculation of positions for data points. We apply cubic spline interpolation corresponding to Kohonen's suggestion of adequate local interpolation schemes[4] (1), introducing no additional topological disordering. To preserve topology, interpolated prototype values are explicitly restricted to their neighbor sampling points' Voronoi polyhedron (2). Having established a high resolution projection layer, we allocate an 2D scatter coordinate for each data point by calculating the best matching HighResSOM unit. Based on the constraints (1) and (2), this can efficiently be done by first calculating the best matching SOM unit (BMU) for a coarse approximation, followed by a local search on the HighResSOM grid in the corresponding region of the BMU. The result of our approach is an alignment of SOM input and output data in one single visualization (cf. Figure 11).

5.2 HighResSOM Example

After having established the HighResSOM concept, we also have adopted the colormap visualizations from basic SOMs as described in Section 4.2. The evaluation of the approach led to results with rich detail, especially for the U-Matrix visualization and the density maps, as illustrated in Figure 5. This example was constructed based on a synthetic test dataset, where randomized data (Label: A, amount: 75%) was blended with 5 heavily blurred and randomly located clusters (Labels: B,C,D,E,F) amounting to 5% of the overall data size, respectively. The data points were mapped on the HighResSOM grid and labeled, shown on the right image of Figure 5. Also, as well on part of the colormap visualizations, the data point mapping yields an improvement of detail and higher precision can be stated, compared to common data density histogram approaches.

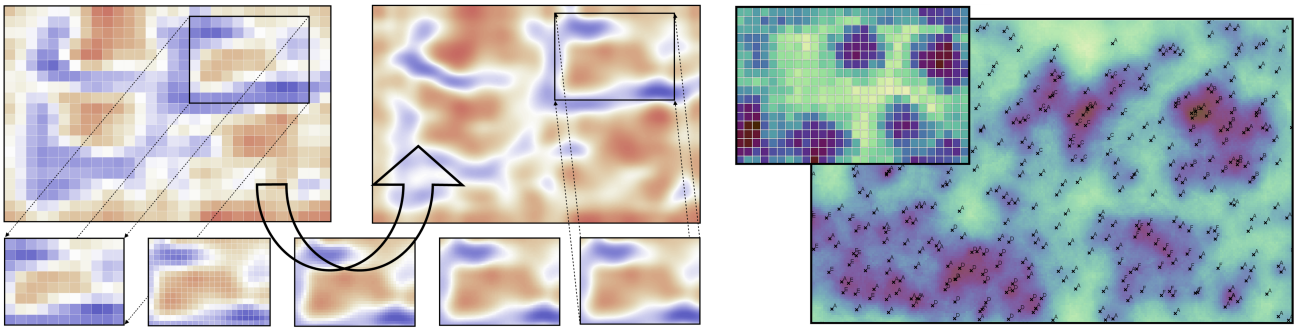

Figure 5. HighResSOM visualizations and data mapping results. Left: U-Matrix visualization with the original SOM (upper left) and the HighResSOM (upper right). As visible in the bottom left images, as the SOM resolution (30x20) is increased (till 480x320), the U-Matrix visualization gets continuously more precise. Right: Visualization of data points in combination with the smoothed density map (S-MAP). The small image pictures the denisty map (S-MAP) of the original SOM, the large image shows the visualization of the density map (S-MAP) with high resolution with the same data basis. Cluster structures are represented more precise, what was a single blur in the original S-MAP can now be explored in detail. In addition, the data points of a synthetic test dataset with 5 clear clusters are presented.

6. SOM-BASED CLUSTER CORRESPONDENCE VISUALIZATIONS

In the last section, we presented a technique to map single data elements to a high-resolution SOM reference grid. Besides the visualization of data samples, we can also leverage this mapping for the correspondence visualization of supportive clustering results. These can stem from other runs of the SOM algorithm, or from completely different clustering algorithms and serve to validate a given SOM clustering. According to this mapping, each cluster prototype of a given reference clustering can be mapped to a distinct position on the SOM grid. We extended our system by integrating the *k-means* and the *DBScan* algorithms as clustering methods for cluster correspondence visualizations. The k-means algorithm adds information about data partitioning, the DBScan algorithm is able to identify distict regions of high data density.

Our basic idea for correspondence visualization is to treat each supportive cluster like a data element, and find its corresponding position in the SOM grid. At that position, we can show the correspondence by (1) drawing a cluster icon, scaling its size to indicate the cluster size, or (2) finding and coloring the nodes matched by each supportive cluster and by distinct colors. In case of partitioning supportive clusterings, we provide a coloring technique that colors each HighResSOM unit with the color of the nearest corresponding supportive cluster, in order to get a colormapping between the HighResSOM and supportive clusterings. Facing the colormapping for density-based supportive clusterings, only HighResSOM units that are density reachable to the data elements of the (potentially arbitrary shaped) cluster structures are colored. If a HighResSOM unit is covered by two or more supportive clusters, the unit color is chosen by the affiliation of the nearest cluster data point. As regards to visual distinction of individual clusters, we assign colors to clusters by equally sampling from the Hue dimension in the HSV color space. Using semi-transparent drawing, it is possible to visually blend the correspondence visualization with further SOM visualizations layers such as the U-matrix. Again, we argue that the visual integration of several views helps to arrive at improved results.

Figure 6 gives an example of our color-coding technique. The image shows a scenario where the wine dataset[34] is used. At first, a SOM with 12x9 units is trained, leveraging certain quality assessment tools, until a SOM with good topological ordering and well-defined cluster properties is obtained. Based on the U-matrix, the density visualization, and the RGB-coloring, three clusters are emerging (cf. Figure 6 top row, clusters are labeled A-C). To validate the identification of these clusters, supportive clusterings are calculated, and visualized on top of the obtained SOM (bottom row in Figure 6).

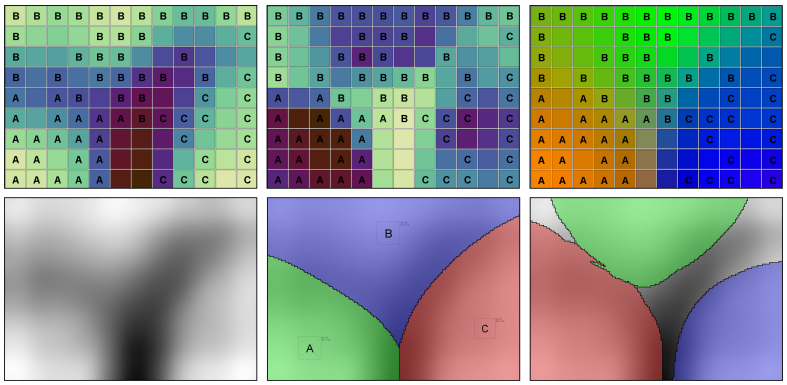

Figure 6. Evaluation of cluster correspondence with the wine dataset.[34] Top row, left to right: based on U-Matrix, density view (S-MAP), and RGB similarity view, three cluster areas are recognized in this SOM (as expected). The additional cluster correspondence views in the bottom row confirm the identification of the three clusters. Specifically, the HighResSOM with U-Matrix (bottom-left), and supportive clusterings based on the k-means (bottom-middle) and DBScan algorithms (bottom-right), are overlaid by color-coding. Note that the HighResSOM U-Matrix (displayed as grayscale colormap) is visualized in all three HighResSOM images. Semi-transparent colormaps of supportive clusterings overlie the U-matrix. Thus, we use the visual attributes color and brightness to assess cluster correspondency between a SOM clustering and supportive clustering results.

7. CASE STUDY

In this section, we demonstrate how our visual cluster analysis system is used to analyze the structure of a data set consisting of several clusters. Our demonstration application workflow includes five successive steps, each with the option to step back to previous phases (cf. Figure 2). By means of examples, we will illustrate applications of individual steps in this workflow.

7.1 Considered Dataset and Initial SOM Training

We use the *UCI ISOLET-5 spoken letter recognition data set*,[34] containing feature vector data of spoken letter samples from 30 speakers, who spelled the A-to-Z alphabet twice, resulting 1560 samples. The data was stored as vectors with 616 dimensions describing phonetic features. As pre-processing and for efficiency reasons, we reduced

the dimensionality to 100 by applying PCA, whereas the variance captured by this reduction amounts to 93% of the unreduced data. Following, we calculated a SOM with a resolution of 20x13 units. We initialized the SOM by a short SOM training with a large neighborhood radius (15 units),[4] and then applied our proposed quality assessment strategies to achieve a preferably high SOM quality, starting with a parametrization according to rules of thumb given in.[4,9] An illustration of the iterative refinement process can be seen in Figure 3, addressing the trade-off between topology preservation and vector quantization. We systematically tried several different parameter sets, arriving at a suitable trade-off.

7.2 Finding Clusters in the SOM Output

Once a SOM output has been obtained, an important task is to identify and distinguish different groups of data (clusters). To this end, in our system we can use two different views: (1) classical SOM visualization views, and (2) correspondence clustering views. In (1) we support the distance-based U-matrix, the density-based S-Matrix, and the Vector Fields visualization. The first three images in Figure 7 show these classic views. From these, the user recognizes an initial grouping of SOM nodes, indicated by red frames manually drawn into the image (fourth image in Figure 7).

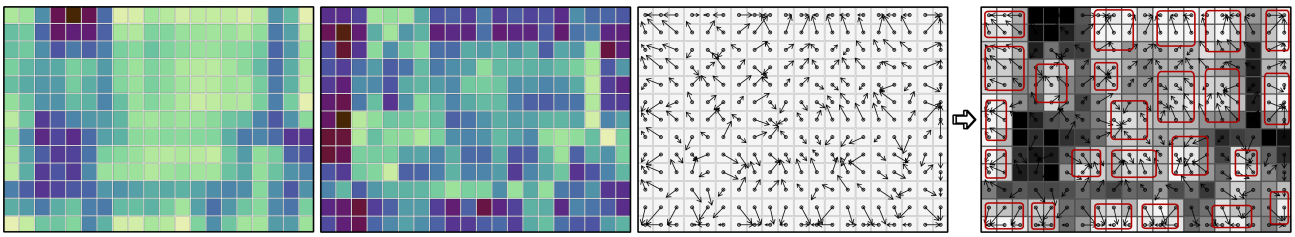

Figure 7. Visual SOM clustering. a) U-Matrix to view unit distances (the darker, the higher the distance). Potential clusters are denoted with bright color values separated by dark values. b) S-Map, distinguishing unit densities (clusters usually exhibit dark color values indicating a high sample density). c) Vector Fields visualization (arrows point at their cluster affiliations). d) Aggregation of all potential cluster information in a grayscale multilayer visualization of the three views a, b(inverted) and c. Red frames indicate the final SOM cluster labeling found interactively by the user.

After this initial investigation, we validate the findings by means of supporting clustering views. We first create two correspondence views for the k-Means and the DBScan cluster algorithm (cf. Figures 8 and 9). Based on these supportive views, the user forms an integrated final grouping, considering all information. Please see the figure captions for details.

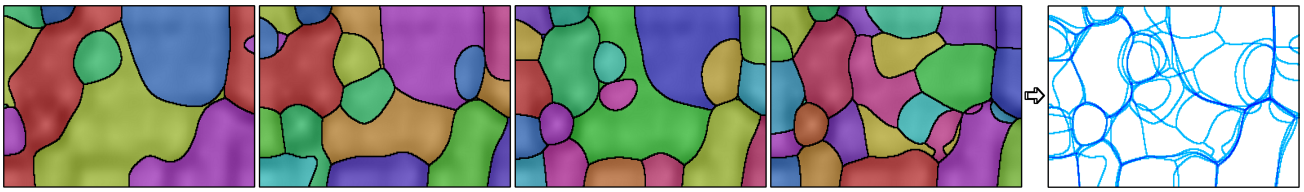

Figure 8. k-Means cluster mappings. The dataset was partitioned 4 times, with different numbers of clusters in each case. Each k-Means correspondence visualization comes along with a cluster color mapping and an indication of the cluster borders (black lines). The U-matrix is blended with the cluster mapping to observe cluster correspondences. k-Means parameters in the Figures are: a) k=5, b) k=10, c) k=20, d) k=30. e) shows a multilayer visualization with the 4 extracted k-Means cluster borders. Note that overlapping cluster borders from different partitionings lead to visually emphasized structures.

Having considered these three individual group structure visualizations, we visually integrate all three views into a display for finding of final grouping by the user. Figure 10 illustrates. Input are the three intermediate results discussed above (small images on the left in Figure 10). The first large image shows the integrating multilayer image. From this, the user interactively outlines (by means of a pencil tool) plausible final groupings at different levels of details. The second large image shows a solution of ten groups, while the last large image shows a solution of more groups, structurally similar to the initial SOM grouping (Figure 7 (right)).

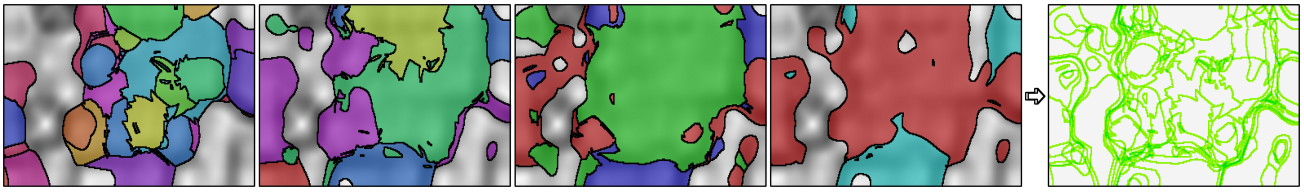

Figure 9. DBScan cluster mappings. The shape of the clusters is different to the map-partitioning results of the k-Means clustering shown in the previous figure. The reason for that effect is explained by the disposition of the DBScan algorithm to pick dense regions of the data set. Due to the possibly arbitrary shape of DBScan clusters in input space, some clusters may be mapped to multiple regions on the HighResSOM. Like in the previous Figure, the U-matrix is blended with the supportive cluster mapping, to support comparison of both clusterings. DBScan parameters s (minimum number of samples) and d (minimum distance) in the Figures are: a) s=5 d=5.00, b) s=10 d=5.50, c) s=15 d=5.75, d) s=20 d=5.50. e) shows a multilayer visualization with the 4 extracted DBScan cluster borders.

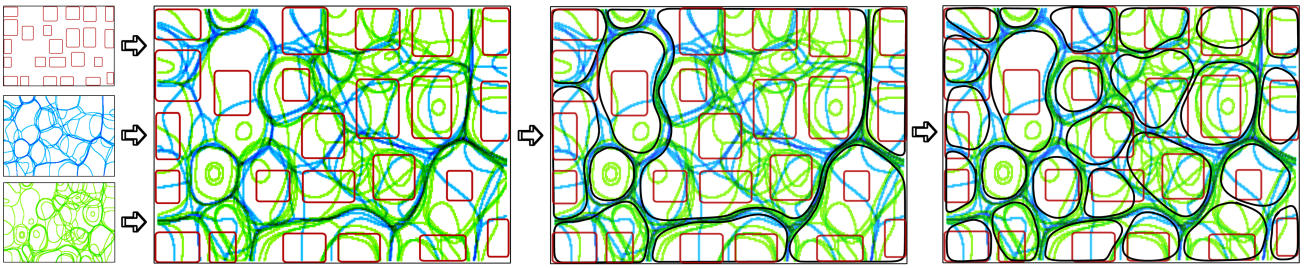

Figure 10. Interactive determination of a final grouping solution based on a multilayer image of intermediate grouping results.

7.3 Evaluation with Class Labels

We also performed an ad-hoc evaluation, by comparing grouping found interactively in the previous section with the class labels of the input ISOLET data sets. These class labels can be regarded as the ground truth grouping information. In Figure 11, we compare the ground truth labeling, with our interactively found groups. The figure shows the SOM grid, where we print for each cell the label of the most frequently occurring data letter sample. We also show, by means of HighResSOM interpolation, all individual letter sample points, indicating the local density of the SOM. We overlay the final grouping found in the interactive cluster analysis process (cf. Figure 10 (most right image)). We also manually added big letter labels denoting the major represented spoken letter samples per found group.

We make two observations from this image. First, the manually found group borders seem to fit with with individual letter groups according to the ground truth classification. Secondly, individual clusters showing more that one letter, contain letters that phonetically sound similar in English (e.g., 'p' and 't', or 'd' and 'e').

We point out that this evaluation is based on the visual cluster analysis that we performed ourselves when experimenting with our system and the ISOLET data set. More objective evaluation should include having a whole user group find and annotate groups, and compare these results in terms of precision and recall in identifying the true clusters (spoken letters, in this case). However, we are convinced that these results indicate that jointly considering multiple cluster visualizations may lead to more robust and useful consolidated clustering results. We point out that based solely on classic SOM visualizations such as the U-Matrix (cf. Figure 7), it would not have been possible to find such a detailed cluster solution. For example, consider the two letter 'w' clusters present in the top-left and bottom-left part of Figure 11. These are not obvious from the SOM views in Figure 7, but are recognized by integration of correspondence cluster views.

8. CONCLUSIONS

Incorporating the notion of quality into visual cluster analysis applications is important due to the explorative nature of the cluster analysis process. Current visual cluster approaches typically fall short of including cluster quality in their visual mappings. In this work, we presented a hierarchy of quality notions with accompanying visual mappings to visually assess the quality of SOM-based clusterings. Our system allows the quality assessment

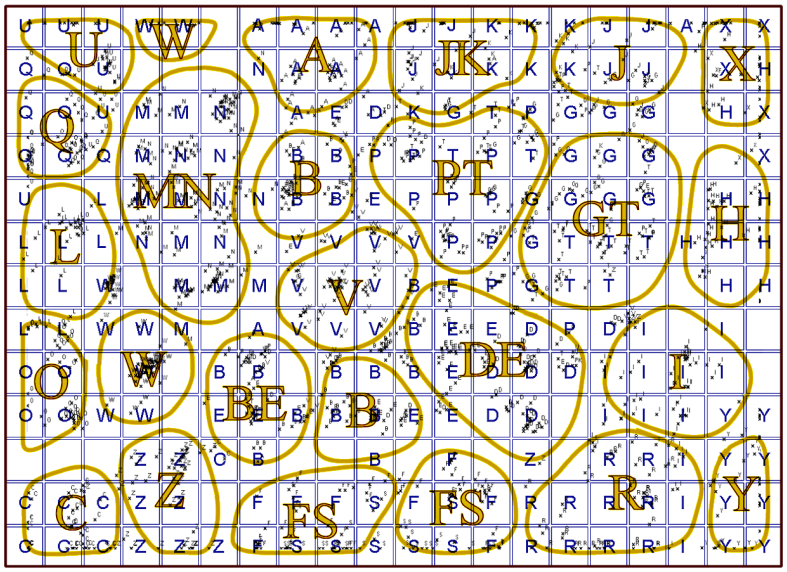

Figure 11. Comparing the interactively found groups with class labels from the test data set (ISOLET).

on varying levels of abstraction, and allows users to validate the results by visual comparison with supportive cluster analysis results. Our approach is useful for improving SOM clustering assessment, and for interactively finding good cluster results. A case study showed the principal applicability of our approach for effective cluster analysis.

Future work includes the extension of our system by further quality notions on all levels, and their combinations. We are specifically interested in enhancing the cluster correspondence views by additional supportive cluster algorithms. Finally, more deep evaluation of the process of converging to best clustering results by means of interactive and quality-aware visual clustering needs be performed.

Acknowledgment

This work was partially supported by the German Research Foundation (DFG) within the project Visual Feature Space Analysis, as part of the Priority Program on Scalable Visual Analytics (SPP 1335).

REFERENCES

[1] Jain, A. and Dubes, R., [*Algorithms for clustering data*] (1988).

[2] Jolliffe, I., [*Principal Components Analysis*], Springer, 3rd ed. (2002).

[3] Sammon Jr, J., "A nonlinear mapping for data structure analysis," *IEEE Transactions on computers* **100**(18), 401–409 (1969).

[4] Kohonen, T., [*Self-Organizing Maps*], Springer, 3rd ed. (2001).

[5] Jain, A., Murty, M., and Flynn, P., "Data clustering: a review," *ACM computing surveys (CSUR)* **31**(3), 264–323 (1999).

[6] Berkhin, P., "A survey of clustering data mining techniques," *Grouping Multidimensional Data* , 25–71 (2006).

[7] MacQueen, J. et al., "Some methods for classification and analysis of multivariate observations," in [*Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*], **1**(281-297), 14, California, USA (1967).

[8] Ester, M., Kriegel, H., Sander, J., and Xu, X., "A density-based algorithm for discovering clusters in large spatial databases with noise," in [*Proc. KDD*], **96**, 226–231 (1996).

[9] Vesanto, J., "Using SOM in data mining," *Licentiates thesis, Helsinki University of Technology, Espoo, Finland* (2000).

[10] Halkidi, M., Batistakis, Y., and Vazirgiannis, M., "Clustering validity checking methods: part II," *ACM SIGMOD Record* **31**(3), 19–27 (2002).

[11] Dunn, J., "Well-separated clusters and optimal fuzzy partitions," *Cybernetics and Systems* **4**(1), 95–104 (1974).

[12] Davies, D. and Bouldin, D., "A cluster separation measure," *trees* **10** (1973).

[13] Hubert, L. and Arabie, P., "Comparing partitions," *Journal of classification* **2**(1), 193–218 (1985).

[14] Goodhill, G., Finch, S., and Sejnowski, T., "Quantifying neighbourhood preservation in topographic mappings," *Institute for Neural Computation Technical Report Series, No. INC-9505* (1995).

[15] Villmann, T., Der, R., Herrmann, M., and Martinetz, T., "Topology preservation in self-organizing feature maps: exactdefinition and measurement," *IEEE Transactions on Neural Networks* **8**(2), 256–266 (1997).

[16] Pölzlbauer, G., "Survey and comparison of quality measures for self-organizing maps," in [*Proceedings of the Fifth Workshop on Data Analysis (WDA04)*], 67–82 (2004).

[17] Vesanto, J., "SOM-based data visualization methods," *Intelligent Data Analysis* **3**(2), 111–126 (1999).

[18] Ultsch, A. and Siemon, H., "Kohonens self organizing feature maps for exploratory data analysis," in [*Proceedings of the International Neural Network Conference (INNC90)*], 305–308 (1990).

[19] Pölzlbauer, G., Dittenbach, M., and Rauber, A., "Advanced visualization of self-organizing maps with vector fields," *Neural Networks* **19**(6-7), 911–922 (2006).

[20] Pampalk, E., Rauber, A., and Merkl, D., "Using smoothed data histograms for cluster visualization in self-organizing maps," *Artificial Neural NetworksICANN 2002* , 81–81 (2002).

[21] Pölzlbauer, G., Rauber, A., and Dittenbach, M., "Advanced visualization techniques for self-organizing maps with graph-based methods," *Advances in Neural Networks–ISNN 2005* , 75–80 (2005).

[22] Nürnberger, A. and Detyniecki, M., "Externally growing self-organizing maps and its application to e-mail database visualization and exploration," *Applied Soft Computing* **6**(4), 357–371 (2006).

[23] Lagus, K., Kaski, S., and Kohonen, T., "Mining massive document collections by the WEBSOM method," *Information Sciences* **163**(1-3), 135–156 (2004).

[24] Youssef, K. and Woo, P., "Efficient music note recognition based on a self-organizing map tree and linear vector quantization," *Soft Computing-A Fusion of Foundations, Methodologies and Applications* **13**(12), 1187–1198 (2009).

[25] Agarwal, P. and Skupin, A., [*Self-organising maps: applications in geographic information science*], Wiley (2008).

[26] Andrienko, G., Andrienko, N., Bremm, S., Schreck, T., von Landesberger, T., Bak, P., and Keim, D., "Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns," *Computer Graphics Forum* **29**(3), 913–922 (2010).

[27] Barthel, K., "Improved Image Retrieval Using Automatic Image Sorting and Semi-automatic Generation of Image Semantics," in [*Ninth International Workshop on Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08*], 227–230 (2008).

[28] Schreck, T., Bernard, J., Von Landesberger, T., and Kohlhammer, J., "Visual cluster analysis of trajectory data with interactive kohonen maps," *Information Visualization* **8**(1), 14–29 (2009).

[29] Schreck, T., Tekušová, T., Kohlhammer, J., and Fellner, D., "Trajectory-based visual analysis of large financial time series data," *ACM SIGKDD Explorations, Special Issue on Visual Analytics* **9**, 30–37 (2007).

[30] Vesanto, J., Sulkava, M., and Hollmén, J., "On the decomposition of the self-organizing map distortion measure," in [*Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*], 11–16, Citeseer.

[31] Ultsch, A., "Maps for the visualization of high-dimensional data spaces," in [*Proc. Workshop on Self organizing Maps*], 225–230 (2003).

[32] Lampinen, J. and Oja, E., "Clustering properties of hierarchical self-organizing maps," *Journal of Mathematical Imaging and Vision* **2**(2), 261–272 (1992).

[33] Bernard, J., von Landesberger, T., Bremm, S., and Schreck, T., "Micro-Macro Views for Visual Trajectory Cluster Analysis," IEEE Information Visualization (2009).

[34] Blake, C. and Merz, C., "UCI repository of machine learning databases," (1998).