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

Conference Paper *in* Proceedings of SPIE - The International Society for Optical Engineering · January 2011

DOI: 10.1117/12.872545

4 authors:

Jürgen Bernard
University of Zurich
105 PUBLICATIONS **1,889** CITATIONS
SEE PROFILE

Tatiana von Landesberger
Technische Universität Darmstadt
116 PUBLICATIONS **2,185** CITATIONS
SEE PROFILE

Sebastian Bremm
Frankfurt University of Applied Sciences
32 PUBLICATIONS **742** CITATIONS
SEE PROFILE

Tobias Schreck
Graz University of Technology
266 PUBLICATIONS **6,467** CITATIONS
SEE PROFILE

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, by reduction 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 in their full magnitude cannot be effectively analyzed or visualized. Automatic cluster analysis techniques can help to this 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 inlcude 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 per se is not a deterministic, but a rather data- and method-dependent process. Typically, the user needs to choose a class of clustering approaches (e.g., 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 incorporating the degree of clustering quality in the cluster visualization.

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

Further author information: (Send correspondence to Jürgen Bernard)
E-mail: juergen.bernard@gris.informatik.tu-darmstadt.de, Telephone: +49 6151 155666

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 on which to measure quality of a SOM analysis results, and to define SOM displays incorporating these 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 contesting 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 its mean point (prototype), 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 algorithm. This type of algorithm defines a cluster as a linked network of data points with a defined 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] 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,[9] 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,[10] the Davis Bouldin Index[11] and the Modified Hubert Statistic.[12] 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.[13] Further general SOM clustering quality measures are surveyed.[13–17]

2.2 Visualization Techniques for the SOM Algorithm

The Self-Organizing Map algorithm's network output is directly suited for visualization purposes. An overview over common SOM-based cluster visualization techniques is given by Vesanto.[18]

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[19] 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.[20,21] Further approaches using color to distinguish between clusters exist.[18,22,23] 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.[18] Improved density-based data visualization techniques include the S-Map[24] and the P-Matrix.[25] These methods smooth the SOM unit affiliation of data, and are thus able to visualize cluster structures. Facing topological aspects, graph-based visualizations can be applied, e.g., to indicate each units nearest arithmetic neighbor.[26]

2.3 SOM-Based Cluster Visualization Applications

The SOM cluster algorithm has been previously used successfully in many different application fields including text analysis,[27,28] multimedia retrieval[29] for geographic information science.[30] In,[31] 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[32] provides an overview for large image collections. In,[33] 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 platform. Our system, initially presented in[33,34] serves as the basis for a number of extensions we introduce in this work. Recall that the SOM algorithm combines data clustering and projection of the data to two dimensional display space. Typically, conventional SOM-based visual cluster analysis systems treat the SOM algorithm as a given black-box component in the system. However, due to the various possibly parameterizations of the algorithm, it may be problematic to easily find appropriate parameterizations and thereby, cluster results.[33] In contrast, our framework provides visual support to supervise also the SOM parameterization, showing the emerging of the SOM results as a function of algorithmic run time. This opens up possibilities for interactively controlling the training steps at user-selectable granularity. This approach has been shown useful for arriving at cluster results suited for the user preferences and application needs. Our framework initially was derived to support a special kind of data elements to be clustered, *trajectory* data. However, our system is applicable to any data type described by vector data as required as input to the SOM clustering (e.g., in Section 7 we will apply it to spoken letter data). Our system supports a broad variety of abstract data visualizations, based on specialized renderers accommodating individual data types such as trajectories, image data, and generic high-dimensional data (Figure 1) illustrates our software platform. So far, our platform 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. Quality assessments of the obtained clustering results are important to guide the interactive parameterization of the system 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.[35]

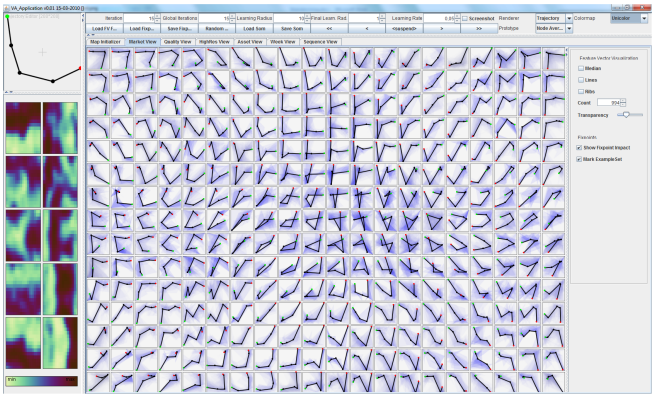

Figure 1. Our SOM-based visual cluster analysis platform. The SOM grid, in this case comprised by visualizations of trajectories patterns, is placed centrally. In the upper left corner, a sketch editor is integrated for interactive SOM initialization, and a component plane view[18] is shown below it. Specific parameterizations and controls to steer the clustering process are arranged on top and right in the window.

We distinguish four abstractions at which to evaluate and judge the resulting quality of a Self-Organizing Map clustering, the first three levels in accordance with.[18]

(1) Global quality network measures. This is the most coarse abstraction of SOM quality assessment. The quality of the SOM is measured by single numeric quality indices such as discussed in Section 2. A visualization of the development of these quality indexes during SOM training is performed by showing line charts.

(2) Unit-based quality measures. At this scaling level, quality measures for each prototype of the SOM network (called a unit, or node) are considered. The SOM grid can be covered with colormaps, shape and specifying text-based visualizations to support SOM unit quality assessment. Section 4 details our implemented unit-based quality visualizations.

(3) Datapoint-based based quality. Here, we measure quality characteristics 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. Finally, 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 next detail the implementations of our quality views on these four levels of abstraction.

4. GLOBAL AND UNIT-BASED SOM QUALITY VISUALIZATIONS

As pointed out previously, interactive SOM cluster analysis makes it necessary to consider quality assessments, 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 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 end results. Both use cases are illustrated in Figure 2 where 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, change key parameters, and continue the process, or to start over again.

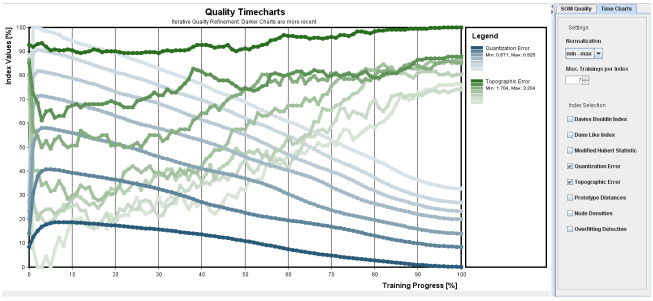

Figure 2. Average vector quantization (blue) versus topology preservation (green) quality scores during a number of epochs in SOM training. Several SOMs have been trained with different parameterizations; darker lines imply more recent trainings. In this example, it is observed that the number of topographic errors increases while the average quantization error rate decreases.

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[19] visualizations), (2) the relative number of data elements mapped per unit (leading to so-called density visualizations),[24,25] 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[18,22,23] (showing clusters and topological orderings by assigning colors), and (5) a vector fields[21] 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. Figures 3 and 4 show examples of respective quality views. Quality assessments and SOM clusterings can be improved by combining multiple visualization tools like an overlay of a colormap and a vector based visualization (cf. Figure 4).

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 to 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. 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 the large pool of implemented methods, and monitor them in real time during execution of the SOM algorithm.

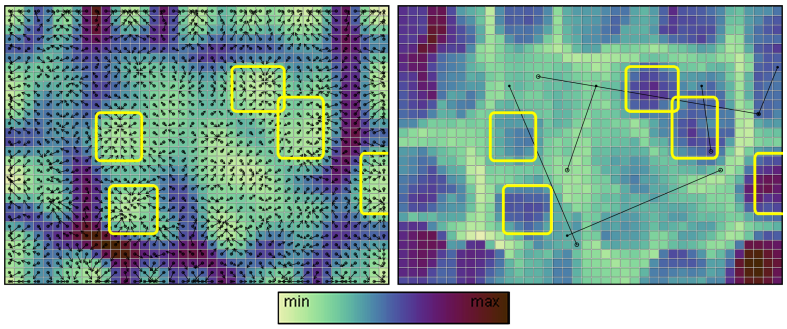

Figure 3. SOM local quality visualizations. Left: the U-Matrix (1) visualization highlights SOM units forming clusters (bright colors). Additionally, arrows of the vector fields visualization show the SOM-units cluster affiliation (5). Right: the smoothed density map (S-MAP) (2), that points out SOM 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. (numbers in brackets refer to description of respective visualizations in the section text.)

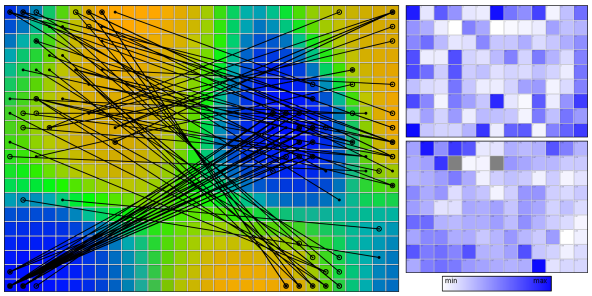

Figure 4. Left: RGB similarity colormap (4) with topographic error connectors (6). The SOM contains two clusters (blue and orange), and the map shows many topographic errors, also called map foldings, indicated by the diverse color distributions and the topographic error connectors visualization (black lines). Both visualization techniques identify a rather low topological SOM quality. Right (top and bottom): density histogram for per-unit density (2) (top) and quantization error histogram for average quantization errors (3) (bottom). Additionally, numeric quality values are visualized on each unit, using a text-based mapping technique. Both histograms give explicit information about individual SOM units, without any structural evidence.

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 develop 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[18] 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[34]). For arbitrary point-based data, this is not possible. We therefore, and to increase accuracy, develop a more detailed mapping location of datapoints over the SOM in a continuous way.

We therefore need a SOM-based mapping technique that is able to allocate each datapoint to an representative screen coordinate able to create high resolution visualizations for the granularity of datapoints. By this, a 2D scatterplot-like SOM-based projection is obtained. Based on it, we can visualize the datapoints in correspondence with the SOM grid. This visualization naturally increases the precision of the density view, because for each point, there is one position available in the display. Moreover, overlapping problems with respect to highly dense SOM units[18] are avoided. Another benefit is the potential visualization of clusters in a scatterplot kind. By this, the restriction of a SOM unit being the smallest possible visualization unit is remedied. Our concept relates to hierarchical SOM approaches,[4,36] 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.[37] SOM units provide a set of support points for Spline-based interpolation for calculation of positions for datapoints. 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 explicitely restricted to their neighbor sampling points' Voronoi polyhedron (2). Having established a high resolution projection layer, we allocate an exact 2D scatter coordinate for each datapoint 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.

5.2 HighResSOM Example

After having established the HighResSOM concept, we also adopted the colormap visualizations from basic SOMs as described in Section 4.2. The main requirement for our high resolution implementation was a realistic preservation of the colormap structures as well as an improvement in precision. 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 datapoints were mapped on the HighResSOM grid and labeled, like 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.

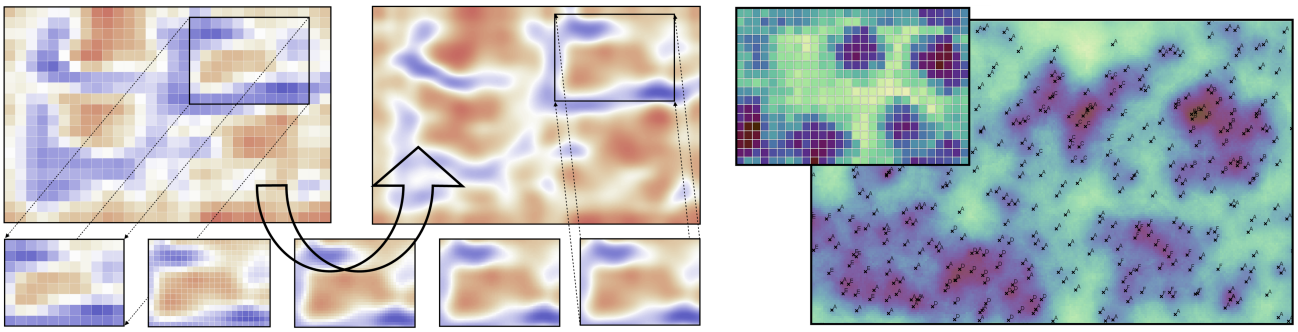

Figure 5. HighResSOM visualizations and data mapping results. Left: U-Matrix visualization with the original SOM (upper left) and the HighResSOM (upper right). As visible in the bottom left images, as the SOM resolution (30x20) is increased (till 480x320), the U-Matrix visualization gets continuously more precise. Right: Visualization of datapoints 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 datapoints 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 as 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. (2) was inspired by the RGB similarity colormap implementation in section 4.2 and the work of Deboeck[23] and Vesanto.[20] 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. In specific, we use transparency to integrate the color-based correspondence visualization with further, underlying SOM visualizations 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[38] 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. 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 the figure).

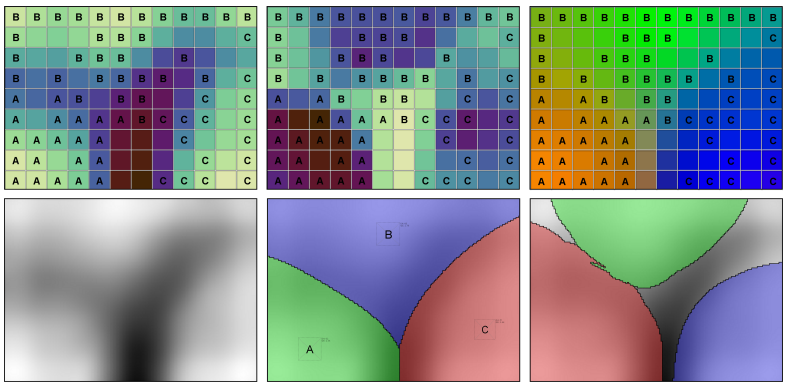

Figure 6. Evaluation of cluster correspondence with the wine dataset.[38] Top row, left to right: based on U-Matrix, density view (S-MAP), and RGB similarity view, three cluster areas are recognized in this SOM (as expected). The additional cluster correspondence views in the bottom row confirm the identification 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 STUDIES

In this section, we demonstrate how our visual cluster analysis system is used to comprehend the structure of a data set consisting of several clusters. Our application workflow includes 5 successive steps, each with the option to step back to previous phases, as illustrated in Figure 7. In contrast to common SOM-based visual cluster analysis tools, our approach addresses an iterative refinement step to improve SOM quality at training. Furthermore, we will make use of our proposed cluster correspondence visualization and refine SOM clustering

results, firstly defined in the SOM clustering phase. Furthermore, we will emphasize the advantage of applying both supportive clustering algorithms: a partitional and a density-based representative.

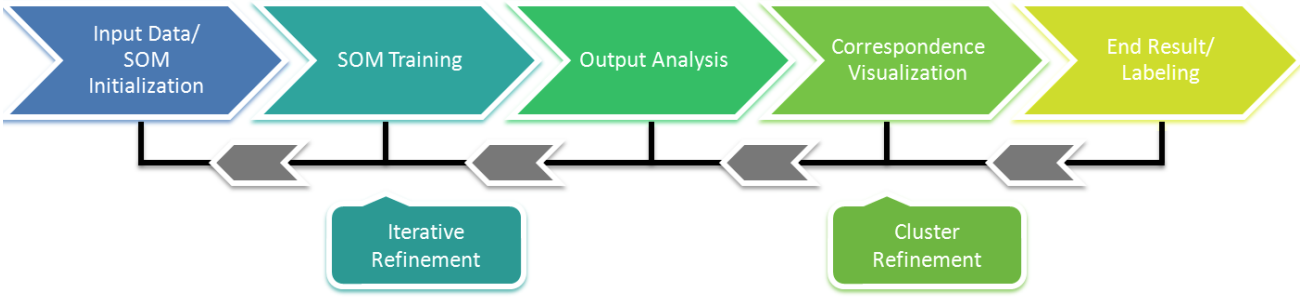

Figure 7. The visual cluster analysis workflow. In contrast to common SOM-based visual clustering approaches, we provide interactive SOM training and iterative SOM cluster refinement strategies, based on correspondence visualizations of supportive clusterings

7.1 Considered Dataset

We use the ISOLET dataset[39] containing feature vector data of spoken letter samples from 150 speakers, who each spelled the A-to-Z alphabet twice. The data was stored as vectors with 616 dimensions describing phonetic features. For evaluation, we took a random subset of 30 speakers, resulting 1560 data samples. 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 huge 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 the rules of thumb.[4,20] An impression of the iterative refinement process can be seen in Figure 2), adressing the trade-off between topology preservation and vector quantization. We aimed for good topology preservation by the reason that near clusters in the input space should also be located closely to each other on the SOM grid. As we had cluster analysis goals, good quality index values regarding vector quantization were crusial as well. As a consequence, we refined the SOM training with a parametrization of well balanced topology preservation and vector quantization, determined by our quality linechart visualization in Figure 2).

7.2 Clustering

This section describes the SOM clustering and cluster refinement process. In contrast to blackbox clustering approaches, our approach is not limited to full automatic cluster calculations, but rather describes an interactive and user-dependent cluster analysis platform with visual support. We initially investigated the SOM cluster structure by applying the most common SOM-based cluster visualization techniques, namely the distance-based U-matrix, the density-based S-Matrix and the nearest neighbor oriented Vector Fields visualization. After performing the SOM clustering (a condensed SOM-cluster visualization can be seen in the last image of Figure 8 d), we examined the allocation of supporting cluster results on the SOM grid, in order to verify our SOM-cluster definition. It has pointed out, that the additional results of the k-means algorithm and the DBScan algorithm (partitioning and density-based) contributed to consolidate a final clustering result, as it can be seen in Figures 9 and 10. Supportive cluster results were evaluated with the Modifed Hubert Statistic, in order to find those results that provided best results, facing the initialization and parametrization problem.[4,5] Please follow Figures 8 - 10 and their captions for details.

7.3 Evaluation of clustering results and classification

After finishing the iterative clustering phase, a final visual evaluation of the result follows. In return, a prototype visualization of the SOM units was established, showing most dominant letters. Additionally, the renderer was expanded to our new data mapping visualization technique, to mark the ISOLET data points according to their

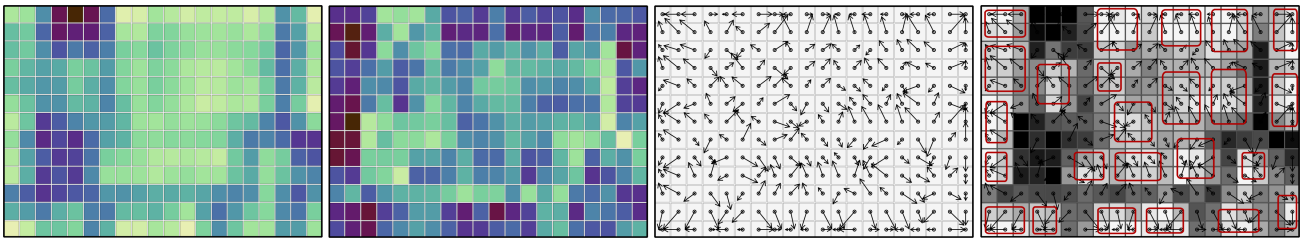

Figure 8. Visual SOM clustering. a) U-Matrix to view unit distances (the darker, the higher the distance. Potential clusters are denoted with bright color values). b) S-Map, destinguishing unit densities (clusters usually exhibit dark color values indicating a high sample density). c) Nearest neighbor oriented 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 ovals indicate the final SOM cluster labeling found by the analyst.

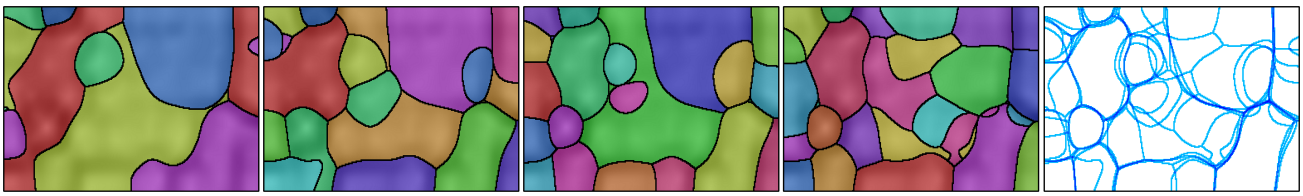

Figure 9. K-means cluster mappings. The dataset was partitioned 4 times, with different numbers of clusters in each case. Each k-means correspondence visualization comes along with a cluster color mapping and a indication of the cluster borders (black lines). The U-matrix is blended with the cluster mapping to observe cluster correspondences a) k=5, b) k=10, c) k=20, d) k=30, e) multilayer visualization with the 4 extracted k-means cluster borders.

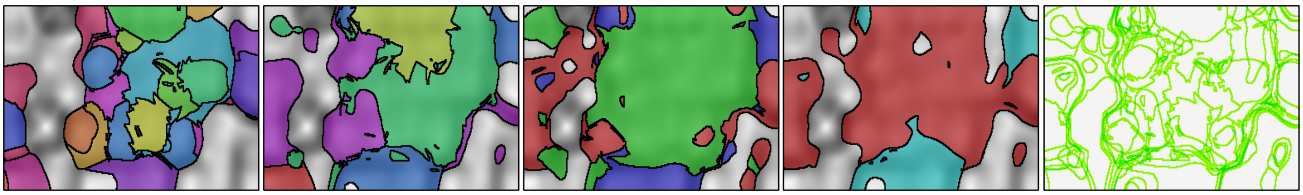

Figure 10. DBScan cluster mappings. Again, each correspondence visualization was enhanced with the cluster color mapping and black outlines of the cluster borders. It can be seen, that the shape of the clusters looks quite different to the map-partitioning appearance of the k-means clustering result. The reason for that effect is explained by the disposition of the DBScan algorithm to pick dense regions of the data set. Due to their arbitrary shape and possibly wide expansion, some clusters even are mapped to multiple regions on the HighResSOM. Combining these two different mapping results, we get a broad cluster mapping feedback from supportive cluster algorithms. The U-matrix was again blended with the cluster mapping to observe cluster correspondences, s denotes the minimum number of samples and d stands for the maximum distance 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) multilayer visualization with the 4 extracted DBScan cluster borders

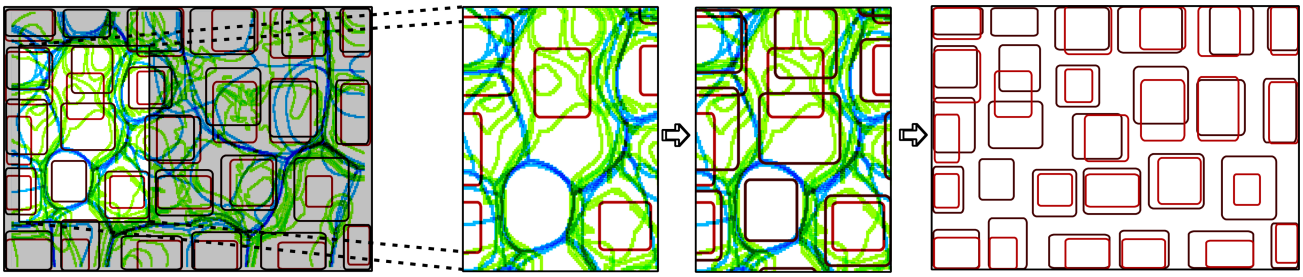

Figure 11. Refinement of the SOM clustering. Left: the results of the SOM clustering and the two supporting clusterings are illustrated in a multilayer image (with a graphics tool) Most of the SOM cluster set fits well to most of the supporting cluster borders. Center: however, some of the red SOM cluster zones do not match with the results of the support clusterings, as the enlarged image shows. Some SOM clusters are imprecise, two SOM clusters are even missing. A cluster refinement was processed (black) by condensation of supportive cluster results. Right: the final clustering (black), compared with the temporal SOM clustering (red).

cluster assignment. The visualizations can be seen in the first two images in Figure 12, where the initial SOM clustering (red) and the final clustering (black) were projected on the image pane. A first conclusion suggests, that the final clustering fits better to the data than the initial one, as it covers more of the potentially clusters and has better defined cluster borders. A second concern to be stated is the promising basis of our prototype visualizations to visually labeling our clusters. Taking both, the unit-based as well as the data-based visualization into account, we hardly had any problems to allocate letters to the clusters. Only in 7 of 27 clusters, we decided to choose a double denotation, like shown in the right image of Figure 12. Finally, we want to introduce some possible interpretations of this ISOLET analysis results shown in Figure 13. First of all, we want to point out our confidence with the SOM training and the clustering result, as we discovered an extensive phonetic consensus. In most cases a 'phonetic topology preservation' of neighbored SOM units and however, a satisfying cluster separation can be stated. On a supercluster-level, the letters (B,C,D,E,G,P,T,V) with relative similar pronunciation, were all allocated at a central SOM area. Further broad clusters were (A,K,J), (F,S), (M,N) and (Q,U). These clusters contained heterogeneous units, respective data points, for which reason the labeling was partially ambiguous. All other clusters could visually be labeled beyond doubt. The two most outlying map regions were covered with W clusters, probably because W contains three syllables and is phonetically quite unique in the alphabet.

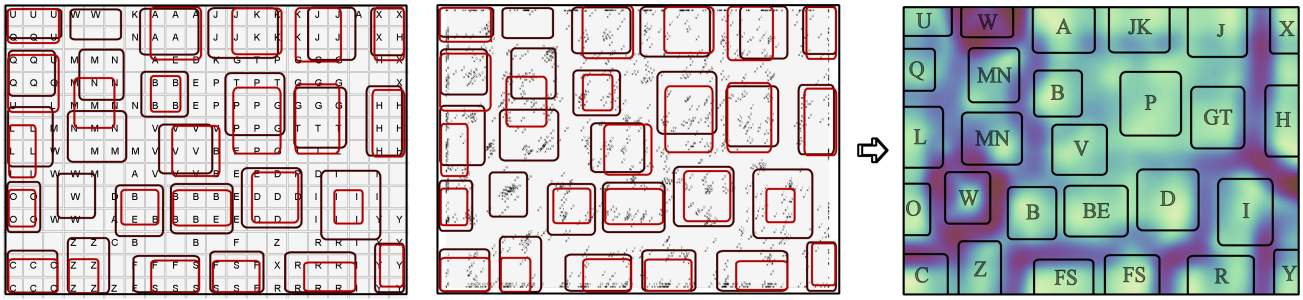

Figure 12. Evaluation and Classification. Left: clustering results before (red) and after (black) the cluster correspondence visualization and cluster refinement phases, mapped on the labeled SOM grid. Center: clustering results before (red) and after (black) the cluster correspondence visualization and cluster refinement phases, mapped on our new high resolution data mapping plane. Right: combining the information from image a and image b, the final labeled cluster result is shown on the high resolution U-matrix

8. CONCLUSIONS

Incorporating the notion of quality into visual cluster analysis applications is important due to the non-deterministic 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 on varying levels of abstraction, and allows to validate the results by visual comparison with supportive cluster analysis results. Our approach is useful for improoving SOM clustering assessment, and for interactively finding good cluster results.

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 evaluation of the process of converging to best clustering results by means of interactive and quality-aware visual clustering should be performed.

REFERENCES

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

[2] Pearson, K., "LIII. On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6* **2**(11), 559–572 (1901).

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

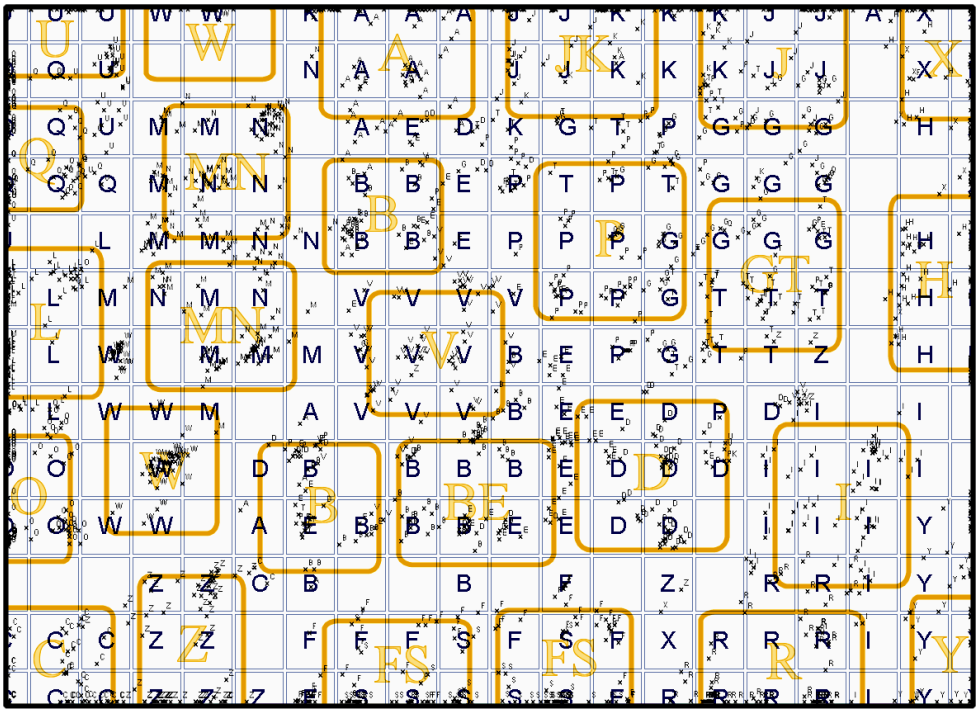

Figure 13. The final result of our visual clustering approach with iterative cluster refinement techniques. The illustration combines three different scales of detail: the cluster scale, the SOM unit scale and the granularity of datapoints.

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

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

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

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

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

[14] Bauer, H., Herrmann, M., and Villmann, T., "Neural maps and topographic vector quantization," *Neural Networks* **12**(4-5), 659–676 (1999).

[15] Bauer, H. and Pawelzik, K., "Quantifying the neighborhood preservation of self-organizing feature maps," *IEEE Transactions on neural networks* **3**(4), 570–579 (1992).

[16] 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).

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

[18] Vesanto, J., "SOM-based data visualization methods," *Intelligent Data Analysis* **3**(2), 111–126 (1999).
[19] 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).
[20] Vesanto, J., "Using SOM in data mining," *Licentiates thesis, Helsinki University of Technology, Espoo, Finland* (2000).
[21] Pölzlbauer, G., Dittenbach, M., and Rauber, A., "Advanced visualization of self-organizing maps with vector fields," *Neural Networks* **19**(6-7), 911–922 (2006).
[22] Kaski, S., Venna, J., and Kohonen, T., "Coloring that reveals high-dimensional structures in data," in [*6th International Conference on Neural Information Processing, 1999. Proceedings. ICONIP'99*], **2** (1999).
[23] Deboeck, G. and Kohonen, T., [*Visual Explorations in Finance with self-organizing maps*], Springer New York (1998).
[24] 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).
[25] Ultsch, A., "Maps for the visualization of high-dimensional data spaces," in [*Proc. Workshop on Self organizing Maps*], 225–230 (2003).
[26] 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).
[27] 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).
[28] Lagus, K., Kaski, S., and Kohonen, T., "Mining massive document collections by the WEBSOM method," *Information Sciences* **163**(1-3), 135–156 (2004).
[29] 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).
[30] Agarwal, P. and Skupin, A., [*Self-organising maps: applications in geographic information science*], Wiley (2008).
[31] 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).
[32] 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).
[33] 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).
[34] Schreck, T., Tekušová, T., Kohlhammer, J., and Fellner, D., "Trajectory-based visual analysis of large financial time series data," (2007).
[35] 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.
[36] Lampinen, J. and Oja, E., "Clustering properties of hierarchical self-organizing maps," *Journal of Mathematical Imaging and Vision* **2**(2), 261–272 (1992).
[37] Bernard, J., von Landesberger, T., Bremm, S., and Schreck, T., "Micro-Macro Views for Visual Trajectory Cluster Analysis," IEEE Information Visualization (2009).
[38] Blake, C. and Merz, C., "UCI repository of machine learning databases," (1998).
[39] Cole, R., Muthusamy, Y., and Fanty, M., "The ISOLET spoken letter database," *Tect. Rep* , 90–004 (1990).