Technical Report OeFAI-TR-2002-29, extended version published in
Proceedings of the International Conference on Artificial Neural Networks,
Springer Lecture Notes in Computer Science, Madrid, Spain, 2002.

Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps*

Elias Pampalk[1], Andreas Rauber[2,**], and Dieter Merkl[2]

[1] Austrian Research Institute for Artificial Intelligence (OeFAI)
Schottengasse 3, A-1010 Vienna, Austria
elias@oefai.at
[2] Department of Software Technology and Interactive Systems
Vienna University of Technology
Favoritenstr. 9-11/188, A-1040 Vienna, Austria
{andi, dieter}@ifs.tuwien.ac.at

Abstract. Several methods to visualize clusters in high-dimensional data sets using the Self-Organizing Map (SOM) have been proposed. However, most of these methods only focus on the information extracted from the model vectors of the SOM. This paper introduces a novel method to visualize the clusters of a SOM based on smoothed data histograms. The method is illustrated using a simple 2-dimensional data set and similarities to other SOM based visualizations and to the posterior probability distribution of the Generative Topographic Mapping are discussed. Furthermore, the method is evaluated on a real world data set consisting of pieces of music.

1 Introduction

The Self-Organizing Map (SOM) [1] is frequently employed in exploratory data analysis to project multivariate data onto a 2-dimensional map in such a way that data items close to each other in the high-dimensional data space are close to each other on the map. In the interactive process of data mining such maps can be utilized to visualize clusters in the data set to support the user in understanding the inherent structure of the data.

Several methods to visualize clusters based on the SOM can be found in the literature. The most commonly used method, i.e. the U-Matrix [2], visualizes the *distances between the model vectors* of units which are immediate neighbors. Alternatively, the *model vectors can be clustered* [3] using techniques such as

* Part of this work was supported by the Austrian FWF under grant Y99-INF.
** Part of the work was performed while the author was an ERCIM Research Fellow at IEI, Consiglio Nazionale delle Ricerche (CNR), Pisa, Italy.

k-means or hierarchical agglomerative clustering. The clusters or dendrograms can be visualized on top of the map. Another possibility to analyze the cluster structure of the SOM is to *project the model vectors* into low-dimensional spaces and use a color coding to link the projections with the original map [4, 5]. A similar approach is to *mirror the movement of the model vectors* during SOM training in a two-dimensional output space using Adaptive Coordinates [6]. A rather different approach to visualize clusters are *data histograms* which count how many data items are best represented by a specific unit. For an overview of different possibilities to visualize the data histograms see e.g. [7]. However, the problem with data histograms is that considering only the best matching unit for each data item ignores the fact that data items are usually represented well by more than just one unit.

The visualization method presented in this paper is based on smoothed data histograms (SDH). The underlying idea is, that clusters are areas in the data space with a high density of data items. The results are related to the posterior probability distribution obtained by e.g. the Generative Topographic Mapping (GTM) [8], however the technique itself is very simple and computationally not heavier than the calculation of the standard data histogram.

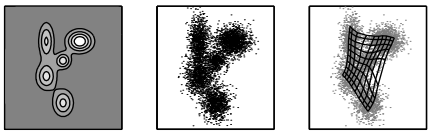
The remainder of this paper is organized as follows. Section 2 presents the calculation of SDHs and discusses the similarities to other methods. In Section 3 the method is evaluated using a data set consisting of 359 pieces of music, and in Section 4 some conclusions are drawn.

2 Smoothed Data Histograms

2.1 Principles

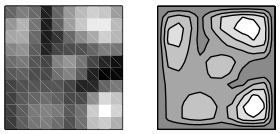
The SOM consists of units which are usually ordered on a rectangular 2-dimensional grid which is referred to as map. A model vector in the high-dimensional data space is assigned to each of the units. During the training process the model vectors are fitted to the data in such a way that the distances between the data items and the corresponding closest model vectors are minimized under the constraint that model vectors which belong to units close to each other on the map, are also close to each other in the data space.

The objective of the SDH is to visualize the clusters in the data set through estimation of the probability density of the high-dimensional data on the map. This is achieved by using the SOM as basis for a smoothed data histogram. The map units are interpreted as bins. The bin centers in the data space are defined by the model vectors and the varying bin widths are defined through the distances between the model vectors. The membership degree of a data item to a specific bin is governed by the smoothing parameter s and calculated based on the rank of the distances between the data item and all bin centers. In particular, the membership degree is s/c_s to the closest bin, $(s-1)/c_s$ to the second, $(s-2)/c_s$ to the third, and so forth. The membership to all but the closest s bins is 0. The constant $c_s = \sum_{i=0}^{s-1} s - i$ ensures that the total membership of each data item adds up to 1.



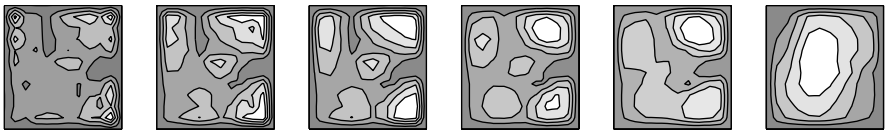
(a) (b) (c)

Fig. 1. The 2-dimensional data space. (a) The probability distribution from which (b) the sample was drawn and (c) the model vectors of the SOM.



(a) (b)

Fig. 2. The SDH ($s = 8$) visualized using (a) gray shadings and (b) contours.



(a) $s = 1$ (b) $s = 3$ (c) $s = 5$ (d) $s = 15$ (e) $s = 20$ (f) $s = 50$

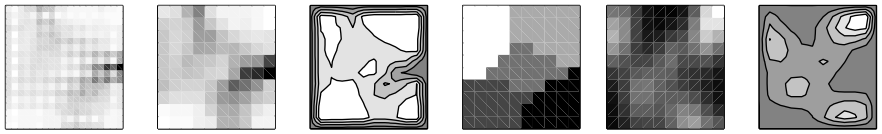
Fig. 3. Different values for the smoothing parameters s and their effects on the SDH.

The SDH is illustrated and the similarities to other methods are discussed using the 2-dimensional data set presented in Figures 1(a) and 1(b). The data set consists of 5000 samples, randomly drawn from a probability distribution that is a mixture of 5 Gaussians. The SOM consists of 10×10 units which are arranged on a rectangular grid. The model vectors in the data space, where the immediate neighbors are connected by a line, are shown in Figure 1(c).

Figure 2(a) illustrates the SDH for $s = 8$. The SOM has been rotated so that the orientation of the latent space corresponds to the orientation of the data space. The gray shadings are chosen so that darker shadings correspond to lower values, while areas shown in lighter shadings correspond to higher values of the SDH. To further simplify the SDH visualization a 5-level contour plot is presented in Figure 2(b), where the values between units are interpolated. A detailed comparison of the SDH and the model vectors in the data space reveals, that the cluster centers found by the SDH correspond well to the cluster centers of the data set.

The influence of the parameter s can be seen in Figure 3. For $s = 1$ the results are identical with those of the data histogram, where clusters cannot easily be identified. For example, the cluster in the upper left corner is represented by three different peaks. For increasing values of s the general characteristics of the data space become more obvious until levelling out into a coarser cluster representation. For extremely high values of s there is only one big cluster which has its peak approximately in the center of the map.

The different cluster shapes reflect the hierarchical structure of the clusters in the data. In particular, $s \geq 50$ resembles the top level in the hierarchy, where the whole data set appears as unity. On the next level, with $s \approx 20$, the difference between the upper right cluster and the rest of the data set becomes noticeable.



(a) (b) (c) (d) (e) (f)

Fig. 4. Alternatives to the SDH, in particular: (a) U-matrix, (b) distance matrix, (c) as contour plot, (d) k-means, (e) posterior distribution of GTM, (f) as contour plot.

In the range $5 \leq s \leq 15$ the 5 clusters can easily be identified, and $s < 5$ depicts the noise within the data. The appropriate hierarchical level and thus the optimal value for the smoothing parameter s depends on the respective application and in particular on the noise level contained in the data.

2.2 Similarities to other Visualization Methods

The U-matrix of the SOM presented in Figure 1(c) can be seen in Figure 4(a). The large distances between the model vectors which are in the center right of the map become clearly visible in the U-matrix. However, these large distances are so dominant that the other distances seem less significant. Figure 4(b) depicts the distance matrix which is calculated as the median of the all the distances of the model vectors between a unit and its immediate neighbors. Although the big gap in the center right of the map is still very dominating, the contour plot depicts that there are 5 clusters (cf. Figure 4(c)).

Generally, there is a close relationship between the distance of the model vectors and the probability density of the data due to an important characteristic of the SOM algorithm known as *magnification factors*. Areas with a high density of data items are represented by more model vectors and thus with more detail than sparse areas. Another advantage of visualizing the distances between the model vectors is that there is no need for further parameters. Note that the cluster centers found by the SDH with $s = 8$ and with the distance matrix are approximately identical. Yet, as we will see in Section 3, the distance matrix provides sub-optimal results in many real-world data mining applications.

Figure 4(d) illustrates the k-means (with $k = 5$) clustering of the model vectors where all map units belonging to one cluster have the same gray shading. Although the 5 clusters can clearly be identified, it is necessary to know how many clusters there are beforehand. The Davies-Bouldin index [9], for example, indicates that $k = 2$ would be a better choice, resulting only in the separation of the upper right cluster from the other four clusters.

The main advantage of the posterior probability of the GTM (cf. Figures 4(e - f)) is the precisely defined statistical interpretation. The GTM was trained using 10×10 latent points, 4 basis functions, and $\sigma = 3$. Although this paper does not deal with visualizing the GTM, it is interesting to note the similarities between the posterior probability and the SDH.

(a) $s = 1$ (b) $s = 2$ (c) $s = 4$ (d) $s = 15$

Fig. 5. Influence of the parameter s on the SDH visualization of the music collection.

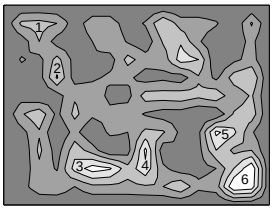
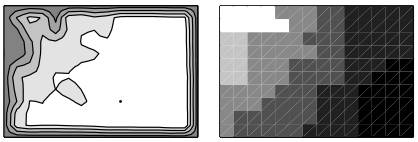


Fig. 6. Music collection with SDH, $s = 3$.



(a) (b)

Fig. 7. Alternative visualizations: (a) distance matrix, (b) k-means, $k = 6$.

3 Visualizing a Music Collection

The SDH was evaluated using a data set consisting of 359 pieces of music which are represented by 1200 features based on Fourier-transformed frequency spectra. The main goal of this experiment is to obtain a clustering of music according to perceived acoustic similarities similar to the results presented in [10]. For full details on the feature extraction and experiments see [11]. Figure 5 depicts the effects of different s values for the SDH of the 14×10 SOM trained with the music collection. The best results were obtained with $s = 3$ (cf. Figure 6). Using, for example, $s = 15$ the cluster in the upper left of the map disappears. Using values below $s = 3$ the cluster in the lower right becomes too dominant. Yet, in all cases the cluster structure is clearly visible.

We will now take a closer look at the interpretation of some of the clusters. Cluster 1 in Figure 6 represents music with very strong beats. In particular, several songs of the group *Bomfunk MCs* are located there, but also songs such as *Blue* by *Eiffel 65* or *Let's get loud* by *Jennifer Lopez*. Cluster 2 represents music mainly by the rock band *Red Hot Chili Peppers* including songs like *Californication* and *Otherside*. Cluster 3 represents more aggressive music by bands such as *Limp Bizkit*, *Papa Roaches*, *Korn*. Cluster 4 represents slightly less aggressive music by groups such as *Guano Apes* and *K's Choice*. Cluster 5 represents concert music and classical music used for films, e.g., the well known *Starwars* theme. Cluster 6 represents peaceful, classical pieces such as *Für Elise* by *Beethoven* or *Eine kleine Nachtmusik* by *Mozart*.

Figure 7 depicts the distance matrix and the k-means visualization of the same SOM. While the former clearly identifies Cluster 1 in the upper left corner, all other clusters are not revealed, mainly because of the general similarity of data in that area and the dominance of the difference to the cluster in the upper

left. The features extracted represent the dynamic patterns of the music pieces and emphasize in particular beats reoccurring in fixed time intervals. Thus, for example, the values of songs by *Bomfunk MCs* are much higher than those of *Für Elise*. The general tendency of the increasing strength of the beats can also be guessed from the k-means visualization for $k = 6$ in Figure 7(b), yet, the cluster structure itself is not visible.

4 Conclusion

The smoothed data histogram is a simple, robust, and computationally light cluster visualization method for self-organizing maps. We illustrated SDHs using a simple 2-dimensional data set and discussed the similarities to other SOM based visualizations as well as to the posterior probability distribution of the GTM. Furthermore, the SDH was evaluated on a real world data set consisting of pieces of music. Observed results show improved cluster identification compared to alternative visualizations based solely on the information extracted from the model vectors of the SOM. The SDH is able to identify clusters by resembling the probability distribution of the data on the map. The low complexity of the SDH calculation allows interactive determination of the smoothing parameter s, and thus is well suited for interactive data analysis. We provide a Matlab® toolbox to visualize SDHs together with several demonstrations at *http://www.oefai.at/~elias/sdh*.

References

1. Kohonen, T.: Self-Organizing Maps. 3rd edn. Springer-Verlag, Berlin (2001)
2. Ultsch, A., Siemon, H.: Kohonen's self-organizing feature maps for exploratory data analysis. In: Proc Int'l Neural Network Conference (INNC'90), Dordrecht, Netherlands, Kluwer (1990) 305-308
3. Vesanto, J., Alhoniemi, E.: Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks **11** (2000) 586-600
4. Himberg, J.: Enhancing the SOM based data visualization by linking different data projections. In: Proc Int'l Symp on Intelligent Data Engineering and Learning (IDEAL'98), Hong Kong (1998) 427-434
5. Kaski, S., Kohonen, T., Venna, J.: Tips for SOM processing and colorcoding of maps. In: Visual Explorations in Finance. Springer Verlag, Berlin, Germany (1998)
6. Merkl, D., Rauber, A.: Alternative ways for cluster visualization in self-organizing maps. In: Proc Workshop on Self-Organizing Maps (WSOM97), Espoo, Finland
7. Vesanto, J.: SOM-Based data visualization methods. Intelligent Data Analysis **3** (1999) 111-126
8. Bishop, C., Svensen, M., Williams, C.: GTM: The generative topographic mapping. Neural Computation **10** (1998) 215-235
9. Davies, D., Bouldin, D.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **1** (1979) 224-227
10. Rauber, A., Frühwirth, M.: Automatically analyzing and organizing music archives. In: Proc 5. Europ Conf on Research and Advanced Technology for Digital Libraries (ECDL 2001). Darmstadt, Germany, Springer (2001)
11. Pampalk, E.: Islands of Music: Analysis, Organization, and Visualization of Music Archives. Master's thesis, Vienna University of Technology, Austria (2001) *http://www.oefai.at/~elias/music*