Mnemonic SOMs: Recognizable Shapes for Self-Organizing Maps

Rudolf Mayer[1], **Dieter Merkl**[2], **Andreas Rauber**[1]
[1]Institute of Software Technology and Interactive Systems
Vienna University of Technology, Vienna, Austria
http://www.ifs.tuwien.ac.at/~{mayer,andi}
[2] School of Computing and Information Technology
University of Western Sydney
Sydney, Australia
d.merkl@uws.edu.au

Abstract - *The Self-Organizing Map (SOM) enjoys significant popularity in the field of data mining and visualization. While its topology-preserving mapping allows easier interpretation of complex data, communicating the location of clusters and individual data items as well as memorizing locations are not solved satisfactorily in conventional rectangular maps. In this paper, a variant of self-organizing maps following standard SOM training practices and having a regular grid structure, but in non-rectangular map shapes, is introduced. Utilizing different recognizable map shapes, such as country or continent maps, or geometrical shapes such as icons, easy description of the location of certain data items becomes possible, and provides an additional mnemonic clue for remembering the locations and relationships between clusters.*

Key words - Self-Organizing Maps, Architecture, Visualization, Mnemonic

1 Introduction

The Self-Organizing Map (SOM) [5], and related self-organizing architectures, enjoy a significant popularity for data mining applications. This is due to their ability to generate a topology-preserving mapping from a high-dimensional input space to a lower dimensional output space. The data thus structured and organized enables an easier interpretation of complex inherent structures and correlations in the data. In many applications the output space is constituted by a two-dimensional rectangular or hexagonal grid. This provides a representation which is easy to read and interpret. A number of techniques for visualizing the maps and labeling the nodes have been suggested to further assist the user in interpreting the maps, e.g. [7] [8] [10].

However, there are still shortcomings when it comes to *explaining* a Self-Organizing Map: when describing a special region of the map, one often has to refer to positions relative to the corners of the map. For bigger maps, only a part of the map can be referred to like this, and therefore often the respective node(s) are directly denoted by their X/Y coordinates in the grid. This method is inappropriate and counter-intuitive in many cases, as the reader first has to search the specific area on the map reverting to the technical notation of nodes and grid

structures. Furthermore, even though advanced visualizations such as the U- Matrix [10] or SDH-based visualizations [7] provide additional structure within the SOM, remembering the relative locations of clusters and their interrelations is usually cumbersome for conventional shapes.

What the user would need is some support in the shape of location indicators within the map that one can refer to when describing locations apart from the center and the four corner regions. Using conventional maps, users refer to locations by different provinces within the country, vicinity to typical landmarks, major cities or rivers. Well-identifiable shapes, such as pictograms, support describing areas and remembering locations better by referring to areas such as the left arm, head region, etc., of an iconic representation of a human body.

We therefore propose using standard self-organizing maps where nodes are arranged in the standard rectangular or hexagonal grid locations; not covering a complete rectangular grid, they would have map shapes following specific, recognizable shapes. The training procedures follow the standard SOM training, with the only adaptation required being the neighborhood calculation within the irregular structure. In this paper we present such a *Mnemonic SOM* model where map nodes are assigned according to standard size parameters such as rows and columns of nodes, or approximate number of nodes, with the required map nodes being projected onto a basically arbitrary map shape. Experiments using a map in the layout of a pictogram of a human body, and a second series of experiments with maps in the shape of the country map of *Austria*, are presented and discussed. The data used consist of (1) an artificial data set of 6 Gaussians, (2) the well-known Iris reference data set, and (3) a small-scale example from the text-mining domain.

The remainder of this paper is organized as follows: first, we present the principles of the Mnemonic SOM, describing the adaptations made to the standard training algorithm in Section 2. The three experimental settings are presented in Section 3, followed by conclusions, caveats, as well as a brief outlook on future work in Section 4.

2 Mnemonic SOMs

Various different neural network models based on the SOM, but exhibiting an irregular shape, have been developed. Among them is the *Incremental Grid Growing* (IGG) [2], which differs from the standard SOM in that the grid is not predefined in size, but iteratively expanded during the learning process, until some stopping criterium is fulfilled. Related to the IGG is the *Adaptive Hierarchical Incremental Grid Growing* [6], which additionally provides hierarchically arranged SOMs. Other models using an irregular grid shape are, for example, the *Growing Cell Structures* [4] or the *Growing SOM* [1], both, similarly to the IGG, having an incrementally growing grid. These models were developed as an answer to the problem of having to specify the size of the grid prior to training in the SOM. In contrast to that, the approach described in this paper primarily aims at providing the user with more easily recognizable and memorizable shapes of SOMs.

For generating the irregular-shaped maps, we process a black and white image representing that shape. The user can specify the number of columns and rows the grid has to contain. Alternatively, the (approximate) total number of requested grid positions can be specified from which the number of rows and columns are computed. Nodes are then placed only on grid positions within the black area, allowing the flexible creation of SOMs in a range of

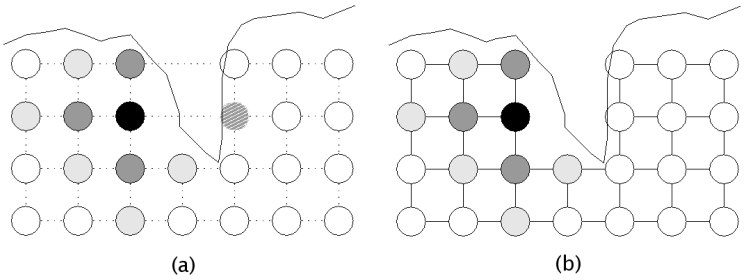

Figure 1: Modeling concave shapes in a standard SOM (a) and in a Mnemonic SOM (b).

shapes.
In general, shapes where individual parts are easy to refer to are preferable, as e.g. geographical maps, which also exhibit an internal structure - borders, rivers, cities - as mnemonic reference points. Another example would be the human body, where one can refer for example to the elbow or knee.
Care has to be taken not to mix the domain of the data collection with the one of the chosen map shape. For example, using a human body as a shape for a map while analyzing a medical document collection will lead to confusion and discussions, rather than an easier explanation of the map (e.g. when documents describing knee ligament injuries are mapped onto nodes that represent the region of the lungs, etc.). Similarly, using a map shaped like a country or continent for data that contains primarily geographic information will most likely be counter-productive (as a reader would just wonder why, for example, some documents originally describing province A get mapped onto nodes in province B). Additionally, one has to keep in mind that specific shapes might not be well-known to all users. Therefore, it is vital to have a clear definition of one's target audience.
Another aspect is that not all shapes may lead to good results for all kinds of data - some shapes may rather impose a clustering on the data by the "clusters" in the shape. This may be the case when using strongly irregular shapes like e.g. the human body in our first experiments.

The feature that the models mentioned in the beginning of this section share with our model is the different way of calculating the *neighborhood* as compared to the conventional SOM model. Basically, neighborhood is no longer based on the grid position of two nodes alone but rather on the existence and the length of the path between these two nodes.
Conventionally, the SOM consists of a grid which is fully occupied, i.e., every position on the grid represents a node. When using irregular shapes, only grid positions within this shape will be occupied by map nodes. This also requires changes to the way the distances between nodes are calculated. Usually, adaptation of nodes will happen on all the neighboring grid positions that are within a certain neighborhood range. With this algorithm, it is not possible to correctly represent concave shapes, which are, however, likely to occur in the irregular shapes we propose. An illustration of of such a situation is given in Figure 1(a), where the black-colored node is the *winner*, and the gray nodes are going to be adapted. Note, however, that the gray-hatched node to the right of the winner is also going to be adapted, while it should actually not be adapted, as there is no direct connection between the nodes.
The metric we utilize for calculating the distances between the nodes in the irregular shapes

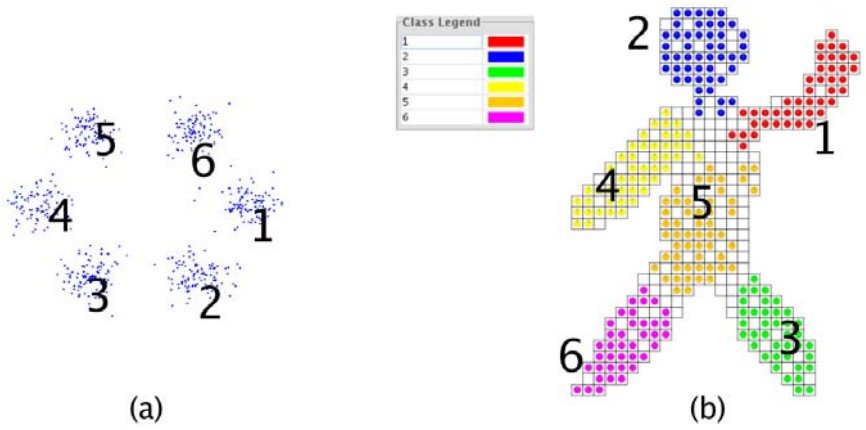

Figure 2: The 2-dimensional data set (a), mapped onto a mnemonic, human-body-shaped SOM (b).

is based on the L1 or *city block* metric, but possible paths between nodes are only found via occupied grid positions. The same map layout as before is given in Figure 1(b). Using a metric that considers empty grid positions, we can also correctly represent concave shapes - the node to the right of the winner is now not going to be adapted, as there is no connection between the nodes within the neighborhood range (in these examples, the neighborhood range has the value of 2). As the way required for calculating the distances between nodes is a costly operation, distances between all pairs of nodes are pre-calculated during the map initialization stage. Utilizing this distance matrix in the neighborhood function, the conventional SOM training algorithm can then be applied.

3 Experiments

3.1 Artificial data on a pictogram

In our first experiment, we use a pictogram in the shape of a human body, i.e. a stickman. The shape itself is interesting as it consists of six easily distinguishable parts: the legs, arms, head and the upper part of the body. For this experiment, we generate a two dimensional data set, depicted in Figure 2(a), which seems to be predestined for the stickman shape. It consists of 600 vectors from six Gaussian clusters.
In Figure 2(b), a trained Mnemonic SOM for the shape of a human body is depicted. The numbers denote the cluster IDs in both subfigures. The circles represent nodes where input data has been mapped onto, while the color denotes the class(es)to which these input vectors belong. The map has a grid size of 30×43, with 367 of these 1290 grid positions actually being occupied by nodes. With these results, it becomes obvious that the clustering capabilities of the SOM are still preserved using the new shapes we propose. The advantage of this irregularly shaped SOM, however, becomes apparent when explaining the results - one can then simply refer to the data points belonging to class '1' as the ones being located in the upturned arm, and the user will immediately know the respective section of the map. When going into more details and describing the location of specific data items, one can take advantage of the knowledge of the underlying shape, and describe them e.g. as being located in the hand, or around the elbow, of the upturned arm.

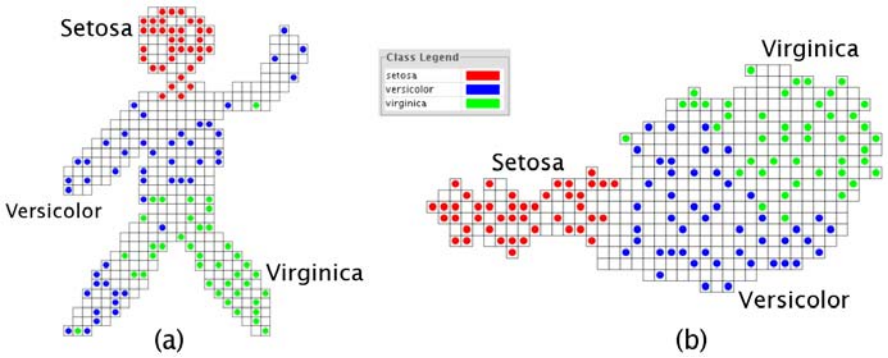

Figure 3: Mnemonic SOMs trained on the Iris data set.

3.2 Iris data set on the pictogram and the map of Austria

In a second series of experiments, we use the *Iris* data set [3], a standard reference data set containing 150 input items equally distributed over three clusters, where the first cluster is linearly separable from the other two clusters, which are somewhat overlapped.
Figure 3 shows the results from two experiments, mapping the data again onto the stickman map, as well as onto a map in the shape of Austria. The stickman map consists of a grid of the size of 30×43 holding 367 nodes, while the map of Austria is represented by a grid of a dimension of 40×21, with 416 grid positions occupied.
In (a) we can observe the first cluster being located on the head of the body, separated from the two other clusters. In (b), we can see the first cluster being mapped onto the provinces of Vorarlberg and Tyrol in western Austria, while the two other clusters are mapped onto central and eastern Austria.

3.3 Time magazine articles on a map of Austria

In our third series of experiments, we are dealing with a setting from the text mining domain. We use the shape of the country map of Austria as the layout. The data used in these experiments is a *Time Magazine* article collection consisting of 420 articles from the 1960's, covering news from politics to social gossip [1]. Each article contains an average 3,700 characters. The aim of our experiments is to generate a thematic clustering of the texts in the collection. For generating our input vectors for the SOM training algorithm from the text articles, we apply a standard bag-of-words document representation, and utilize a $tf \times idf$ weighting scheme [9]. After removing both too low and too high frequency terms, we obtain feature vectors of 3819 dimensions.
In our experiments, we use different map sizes for representing the shape of the country map of Austria. Figure 4(a) shows the full map with a grid size of 43×22, out of which 470 grid positions are actually occupied by nodes. The labels were added after manually inspecting the document mapping. The western provinces Vorarlberg and Tyrol, as well as the south, Carinthia and southern Styria, feature mostly European politics topics. The province of Burgenland in eastern Austria features mostly topics covering Africa, while the parts of the map in and north-east of Vienna hold documents about the India-Pakistan conflict over

[1] The collection is available at `http://www.ifs.tuwien.ac.at/~andi/somlib/experiments_time60.html`

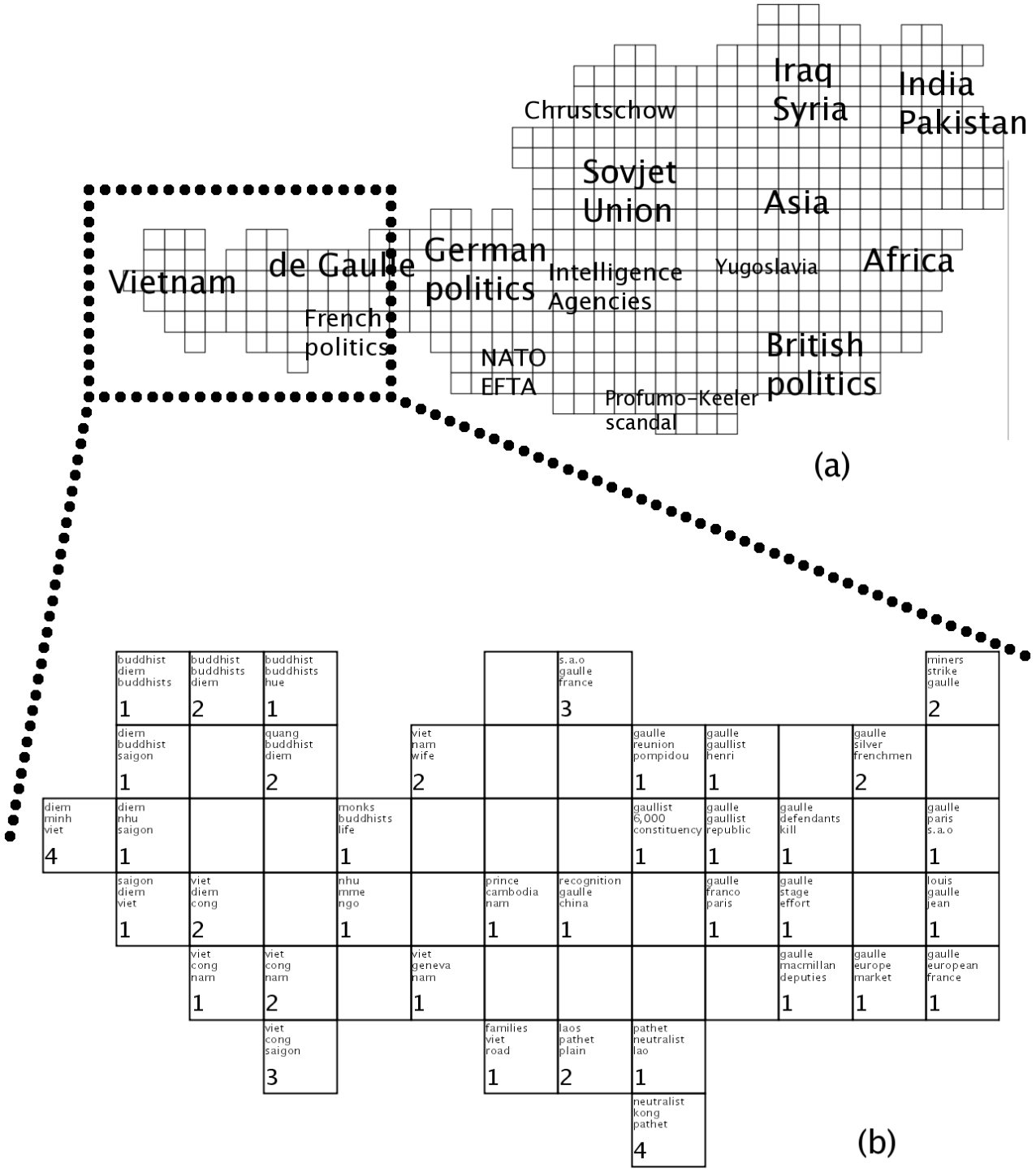

Figure 4: A trained mnemonic SOM in the shape of Austria.

Kashmir. The province of Lower Austria features documents about Iraq, Syria and Asia, while Salzburg has documents about the Soviet Union mapped onto it. Interestingly, the area of southern Salzburg / northern Carinthia, situated between the western European and Soviet "block" has as the main topic reports about espionage and intelligence agencies.
In Figure 4(b), we show a section of the map, namely the provinces of Vorarlberg and Tyrol, in western Austria, in more detail. The nodes presented in this figure bear labels to aid the user in interpreting the mapping; these labels were automatically extracted from the document by using the LabelSOM method [8]. When explaining to the reader what part of the map we show, we can now easily name the provinces, while otherwise we would need to refer to the coordinates of the nodes in order to allow the reader to locate the section concerned in the complete map.
Obviously, care has to be taken that the respective shapes and metaphors are known to the target audience of the data analysis task. As an example of such a more narrative and more intuitive explanation of the cluster structure (at least for readers familiar with Austrian geography), we will now discuss the results in greater detail. The part of the map that depicts the province of Vorarlberg (the left part of the map) holds input data on the war in Vietnam, with the area of Bregenzer Wald in northern Vorarlberg holding documents describing the religious aspects of the conflict. The area south of Feldkirch and Bludenz, i.e. the Montafon, represents the military aspects. Crossing the mountain range of Arlberg (in this case the SOM cluster boundary nicely coincides with a significant topological barrier in reality) into Tyrol, we arrive at a different cluster holding documents on France under President de Gaulle.

4 Conclusion

While the SOM has become a prominent tool for data analysis, describing regions on the SOM or the location of data items remains a cumbersome experience. We therefore introduce the Mnemonic SOM, an adaptation of the standard SOM model allowing the use of easily recognizable map shapes. We propose using shapes that are familiar to the user and allow an easier explanation of a SOM. It allows to refer to regions of the map not by addressing them by the corner of the map they are located in, or by defining the region by its X/Y coordinates in the grid. In the example of our experiments, we can refer to (parts of) provinces and even cities, or parts of the human body, respectively, and the user - assuming that he or she is familiar with the shape - will immediately know what area of the map we talk about. This technique can become even more useful when we show and describe only sections of a map - the user will easily know where it belongs to in the complete map.

Though there are some obvious advantages, there are also a couple of issues one has to pay attention to when using the Mnemonic SOM. Most importantly, one should avoid mixing the domain of the data collection with the one of the chosen map shape - using a human body as a shape for a map analyzing a medical data collection will more likely lead to confusion, rather than to an easier explanation of the map. Moreover, it is important to keep in mind that specific shapes might not be well known to all users - it is necessary to have a clear definition of one's target audience.
Another aspect is that not all shapes may lead to good results for all kinds of data - some shapes may rather predetermine a clustering on the data by the "clusters" in the shape.

On the other hand, maps of too regular, symmetric shapes may not support easier region description than the conventional rectangular SOM shape. Thus, the optimal choices seem to be shapes that exhibit a high internal structure well-known to the target audience, with a preferably not too dominant external structure of the map shape itself, and where the shape is unrelated to the target data domain.
We are currently in the process of performing a series of user studies using different map shapes in a range of target domains, evaluating their describability and mnemonic characteristics.

References

[1] Damminda Alahakoon, Saman K. Halgamuge, and Bala Srinivasan. Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks*, 11(3):601–614, May 2000.

[2] Justine Blackmore and Risto Miikkulainen. Incremental grid growing: Encoding high-dimensional structure into a two-dimensional feature map. In *Proc. ICNN'93, International Conference on Neural Networks*, volume I, pages 450–455, Piscataway, NJ, 1993. IEEE Service Center.

[3] R.A. Fisher. The use of multiple measurements in taxonomic problems. In *Annual Eugenics, 7, Part II*, pages 179–188, 1936.

[4] Bernd Fritzke. Growing cell structures—a self-organizing network in k dimensions. In I. Aleksander and J. Taylor, editors, *Artificial Neural Networks, 2*, volume II, pages 1051–1056, Amsterdam, Netherlands, 1992. North-Holland.

[5] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995.

[6] Dieter Merkl, Shao Hui He, Michael Dittenbach, and Andreas Rauber. Adaptive hierarchical incremental grid growing: An architecture for high-dimensional data visualization. In *Proceedings of the 4th Workshop on Self-Organizing Maps*, Advances in Self-Organizing Maps, pages 293–298, Kitakyushu, Japan, September 11-14 2003.

[7] E. Pampalk, A. Rauber, and D. Merkl. Using smoothed data histograms for cluster visualization in self-organizing maps. In *Proceedings of the Intl Conf on Artificial Neural Networks (ICANN 2002)*, pages 871–876, Madrid, Spain, August 27-30 2002.

[8] Andreas Rauber and Dieter Merkl. Automatic labeling of self-organizing maps for information retrieval. *Journal of Systems Research and Information Systems (JSRIS)*, 10(10):23–45, December 2001.

[9] Gerald Salton. *Automatic text processing – The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[10] Alfred Ultsch and H.P. Siemon. Kohonen's self-organizing feature maps for exploratory data analysis. In *Proc. Intern. Neural Networks*, pages 305–308, Paris, 1990. Kluwer Academic Press.