HIERARCHICAL EVOLVING TREES TOGETHER WITH GLOBAL AND LOCAL LEARNING FOR LARGE DATASETS IN MALDI IMAGING

Stephan Simmuteit[1] *and Frank-Michael Schleif*[2] *and Thomas Villmann*[3]

[1]Univ. of Leipzig, Medical Dept., Semmelweisstrasse 10, 04103 Leipzig, Germany
[2]Univ. of Bielefeld, CITEC, Universitätsstrasse 21-23, 33615 Bielefeld , Germany
[3]Univ. of Appl. Sc. Mittweida, MPI, Technikumplatz 17, 09648 Mittweida, Germany
simmuteit@googlemail.com, schleif@informatik.uni-leipzig.de, thomas.villmann@hs-mittweida.de

ABSTRACT

The analysis of very large sets of data with multiple thousand measurements is an increasing problem. High-throughput approaches in the life science lead to large amounts of data which need to be analyzed by data mining approaches. Focusing on clustering and visualization approaches a common problem are very large similarity matrices. Standard techniques suffer from memory and runtime limitations for such complex settings or are not applicable at all. Here we present a hierarchical composite clustering employing data specific properties to deal with this problem for data with an inherent hierarchical order. As an additional advantage our algorithm allows easy control of the clustering depth. The method is a prototype based approach leading to sparse, compact and interpretable models. We derive the algorithm and present it on data taken from tissues slices of high resolution MALDI Imaging. Results show an effective clustering as well as significant improvements of the computational complexity for this type of data.

1. INTRODUCTION

Life science experiments generate large sets of measurements as a more and more challenging problem for the used machine learning methods. Often similarity matrices, generated from the data, are used as an input to the machine learning approaches and clusterings or a low dimensional visualization of the data is in focus. Prominent techniques of this type are e.g. Single Linkage Clustering (SLC) or Multi Dimensional Scaling (MDS) [1]. In the last decade the number of processed samples and the resolution in life science experiments has increased a lot and hence also the amount of data. The similarity matrices of such data are getting huge now and calculation with it becomes very challenging. One potential solution for this problem is to use iterative approaches, but this would slow down the analysis process significantly in most cases and will in parts also be suboptimal. An alternative strategy is to employ specific structural information of the data as we will show subsequently for a clustering task. In this study we focus on a specific type of *prototype* based clustering based on Self-Organizing Maps (SOM) as introduced by Kohonen [2] and a specific type of data with an inherent hierarchical structure as it is common e.g. for tissues data (consisting of different types of sub-tissues).

An especially well suited approach to deal with hierarchical data is the Tree based Self Organizing Map or Evolving Tree-SOM (ET) [3]. Although very promising, it is limited in case of very huge sets of data. In this paper we will extend the ET to a hierarchical composite clustering approach (HCC) using local and global learning strategies. The approach will be evaluated on two distinct sets of data taken from tissue studies as common in pathology. The paper is organized as follows: in Sec. 2 we give a brief introduction to prototype based learning and explain the used basic algorithms SOM, ET and an additional prototype based clustering developed for the handling of very large data sets. In Sec. 4 the HCC approach is presented and applied on real life experimental data as described in Sec 3. We conclude with a summary of the results and an outlook for further improvements and research directions.

2. METHODS

2.1. Self organizing maps

The SOM constitutes one of the most popular unsupervised approaches for clustering, visualization and data mining of high-dimensional data [2]. SOMs belong to the prototype based methods of data representation. Due to its inherent regularization abilities SOMs are also applicable in case of sparse data sets. SOMs can be taken as unsupervised learning of topographic vector quantization with a topological structure (grid) within the set of prototypes (codebook vectors). Thereby, roughly speaking, topology preservation means that similar data points $\mathbf{v} \in V$ with $V \subseteq \mathbb{R}^D$ and D the data dimensionality are mapped onto identical or neighbored grid locations which have pointers into the data space (weight vectors). The weight vectors also are called prototypes, because they represent parts of the data space.

Assume that data $\mathbf{v} \in V \subseteq \mathbb{R}^D$ are given distributed according to an underlying distribution $P(V)$. A SOM is determined by a set A of neurons $\mathbf{r}$ equipped with weight vectors (prototypes) $\mathbf{w_r} \in \mathbb{R}^D$. The neurons are arranged on a lattice structure, which determines the neighborhood relation $N(\mathbf{r}, \mathbf{r}')$ between the neurons $\mathbf{r}$ and $\mathbf{r}'$. Denote the set of prototypes by $\mathbf{W} = \{\mathbf{w_r}\}_{\mathbf{r} \in A}$. The mapping description of a trained Heskes-SOM[1] defines a function

$$\Psi_{V \to A} : \mathbf{v} \mapsto s(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{r} \in A} le(\mathbf{r}) \tag{1}$$

where

$$le(\mathbf{r}) = \sum_{\mathbf{r}' \in A} h_\sigma(\mathbf{r}, \mathbf{r}') \xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) \tag{2}$$

[1]An extension provided by Heskes incorporating a cost function

is the local neighborhood weighted error of distances $\xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'})$. $\xi(\mathbf{v}, \mathbf{w})$ is an appropriate distance measure, usually the quadratic Euclidean norm $\xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}}) = (\mathbf{v} - \mathbf{w}_{\mathbf{r}})^2$. However, here we only suppose $\xi(\mathbf{v}, \mathbf{w})$ to be arbitrary assuming differentiability and symmetry and assessing some dissimilarity. The function

$$h_\sigma(\mathbf{r}, \mathbf{r}') = \exp\left(\frac{N(\mathbf{r}, \mathbf{r}')}{2\sigma^2}\right) \tag{3}$$

determines the neighborhood cooperation with range $\sigma > 0$. In this formulation, an input stimulus $\mathbf{v}$ is mapped onto that position $\mathbf{r} \in A$ of the SOM, the local error $le(\mathbf{r})$ of which is minimum, whereby the average over all neurons according to the neighborhood is taken. We refer to this neuron $s(\mathbf{v})$ as the winner.

During the adaptation process a sequence of data points $\mathbf{v} \in V$ is presented to the map representative for the data distribution $P(\mathcal{V})$. Each time the currently most proximate neuron $s(\mathbf{v})$ according to (1) is determined. All prototypes are gradually adapted according to the neighborhood degree of the respective neuron to the winning one by

$$\triangle \mathbf{w}_{\mathbf{r}} = -\epsilon h_\sigma(\mathbf{r}, s(\mathbf{v})) \frac{\partial \xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}})}{\partial \mathbf{w}_{\mathbf{r}}} \tag{4}$$

with a small learning rate $\epsilon > 0$. This adaptation follows a stochastic gradient descent of the cost function for the SOM as introduced by HESKES [4]:

$$E_{\mathrm{SOM}} = \frac{1}{2C(\sigma)} \int P(\mathbf{v}) \sum_{\mathbf{r}} \delta_{\mathbf{r}}^{s(\mathbf{v})} \sum_{\mathbf{r}'} h_\sigma(\mathbf{r}, \mathbf{r}') \xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) d\mathbf{v} \tag{5}$$

where $C(\sigma)$ is a constant which we will drop in the following, and $\delta_{\mathbf{r}}^{\mathbf{r}'}$ is the usual Kronecker symbol checking the identity of $\mathbf{r}$ and $\mathbf{r}'$.

One main aspect of SOMs is the visualization ability of the resulting map due to its topological structure. Under certain conditions the resulting non-linear projection $\Psi_{V \to A}$ generates a continuous mapping from the data space V onto the grid structure on A. This mapping can mathematically be interpreted as an approximation of the principal curve or its higher-dimensional equivalents [5]. Thus, as pointed out above, similar data points are projected on prototypes which are neighbored in the grid space A. Further, prototypes neighbored in the lattice space should code similar data properties, i.e. their weight vectors shoud be close together in the data space according to the dissimilarity measure ξ. This property of SOMs is called topology preserving (or topographic) mapping realizing the mathematical concept of continuity (see also [6]). For data sets with an inherent hierarchical structure the rectangular grid topology is not any longer appropriate and a tree topology may be more appropriate as shown subsequently.

2.2. Evolving Tree

Suppose we consider an ET $\mathcal{T}$ with nodes $r \in R_{\mathcal{T}}$ (set of nodes) and root r_0 which has the depth level $l_{r_0} = 0$. A node r with depth level $l_r = k$ is connected to its successors r' with level $l_{r'} = k + 1$ by directed edges $\varepsilon_{r \to r'}$ with length is unit. The set of all direct successors of the node r is denoted by S_r. If $S_r = \varnothing$ is valid, the node r is called a leaf and denoted by $\odot$. The degree of a node r is $\delta_r = \#S_r$, here assumed to be constant δ for all nodes except the leafs. A sub-tree $\mathcal{T}_r$ with node r as root is the set of all nodes $r' \in R_{\mathcal{T}_r}$ such that there exists a directed cycle-free path $p_{r \to r'} = \varepsilon_{r \to m} \circ \ldots \circ \varepsilon_{m' \to r'}$ with $m, \ldots, m' \in R_{\mathcal{T}_r}$ and $\circ$ as the concatenation operation. $L_{p_{r \to r'}}$ is the length of path $p_{r \to r'}$, i.e. the number of concatenations plus 1. The distance $d_{\mathcal{T}}(r, r')$ between nodes r, r' is defined as

$$d_{\mathcal{T}}(r, r') = L_{p_{\hat{r} \to r}} + L_{p_{\hat{r} \to r'}} \tag{6}$$

with paths $p_{\hat{r} \to r}$ and $p_{\hat{r} \to r'}$ in the sub-tree $\mathcal{T}_{\hat{r}}$ and $R_{\mathcal{T}_{\hat{r}}}$ contains both r and r' and the depth level $l_{\hat{r}}$ is maximum for all sub-trees $\mathcal{T}_{\hat{r}'}$ which contain r and r'. A connecting path between a node r and a node r' is defined as follows: let $p_{\hat{r} \to r'}$ and $p_{\hat{r} \to r}$ be direct paths such that $L_{p_{\hat{r} \to r'}} \cdot L_{p_{\hat{r} \to r}}$ is $d_{\mathcal{T}}(r, r')$. Then $p_{r \to r'}$ is the reverse path $p_{r' \to \hat{r}} \cdot p_{\hat{r} \to r}$ and the node set of P is denoted by $\mathcal{N}_{p_{r \to r'}}$. As for usual SOMs, each node r is equipped with a prototype $\mathbf{w}_r \in \mathbb{R}^D$, provided that the data to be processed are given by $\mathbf{v} \in V \subseteq \mathbb{R}^D$. Further, we assume a differentiable similarity measure $d_V : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. The winner detection is different from usual SOM but remains the concept of winner-take-all. For a given subtree $\mathcal{T}_r$ with root r the *local winner* is

$$s_{\mathcal{T}_r}(\mathbf{v}) = \arg\min_{r \in S_r} \left(d_V(\mathbf{v}, \mathbf{w}_r)\right) \tag{7}$$

If $s_{\mathcal{T}_r}(\mathbf{v})$ is a leaf then it is also the overall winner node $s(\mathbf{v})$. Otherwise, the procedure is repeated recursively for the sub-tree $\mathcal{T}_{s_{\mathcal{T}_r}}$. The *receptive field* Ω_r of a leaf r (or its prototype) is defined as

$$\Omega_r = \{\mathbf{v} \in V | s(\mathbf{v}) = r\} \tag{8}$$

and the receptive field of root r' of a sub-tree $\mathcal{T}_{r'}$ is defined as

$$\Omega_{r'} = \cup_{r'' \in R_{\mathcal{T}_{r'}}} \Omega_{r''} \tag{9}$$

The adaptation of the prototypes $\mathbf{w}_r$ takes place only for those prototypes, where the nodes r of are leafs. The other nodes remain fixed. This learning for a randomly selected data point $\mathbf{v} \in V$ is neighborhood-cooperatively as in usual SOM:

$$\triangle \mathbf{w}_r = \epsilon h_{SOM}(r, s(\mathbf{v}))(\mathbf{v} - \mathbf{w}_r) \tag{10}$$

with $s(\mathbf{v})$ being the overall winner and $\epsilon > 0$ a small learning rate. The neighborhood function $h_{SOM}(r, r')$ is defined as a function depending on the tree distance $d_{\mathcal{T}}$ usually of Gaussian shape

$$h_{SOM}(r, r') = \exp\Big(\frac{-\left(d_{\mathcal{T}}(r, r')\right)^2}{2\sigma^2}\Big). \tag{11}$$

with neighborhood range σ.

Unlike for the SOM we cannot guarantee that $s(\mathbf{v})$ is the true best matching unit (*bmu*), because the tree model is subject of a stochastic optimization process.

The whole ET learning is a repeated sequence of adaptation phases according to the above mentioned prototype adaptation and tree growing beginning with a minimum tree of root r_0 and its δ successors as leafs. The decision, which leafs become roots of sub-trees at a certain

time can be specified by the user. Subsequently for each node r a counter b_r is defined. This counter is increased if the corresponding node becomes a winner and the node is branched if a given threshold $\theta \in \mathbb{N}, \theta > 0$ is reached.

Possible criteria might be the variance of the receptive fields of the prototypes or the number of winner hits during the competition. The prototypes of the new leafs should be initialized in a local neighborhood of the root prototype according to d_V. Hence, the ET also can be taken as a special growing variant of SOM.

2.3. Patch clustering for large datasets

Patch Clustering is a method to cluster datasets, which are too big to fit into the main memory. The idea is to cluster patches of the data separatly from another and to employ statistics about the last patch in the learning procedure. This statistics include the old prototypes and how often they became the BMU. The statistics and the datapoints from the next patch are used to learn new prototypes, until every patch has been learned. Details are provided in [7].

2.4. Evolved Tree of Subtrees

In Order to process very large datasets we develop an evolved tree of subtrees, called Hierarchical Composite Clustering (HCC). Therefore we employ the aforementioned patch clustering. The data set V is divided into K patches which are then processed using patch-clustering with a defined in general smaller number of prototypes. This results in a patch clustering model (PCM) which constitutes the global learning model. Now the receptive fields of the prototypes of the PCM are analyzed equivalent to (8) generating a subset of V denoted as V_i. For a V_i a ET is generated and the obtained subtree $\mathbb{T}_i$ is assigned to the leaf $\odot_i$ of the global tree $\mathbb{T}_G$ with the root node r_o as the mean of V. Using this procedure even for large data sets V the problem can be divided reasonably well into solvable sub-problems. The potential loss of accuracy with respect to a full model is only minor for our data, see also [7]. The global tree represents the rough hierarchical structure of the data and its leaf prototypes represent subsets of the original data. Hence the leafs $\odot_i$ of the global tree are used as the roots for the local trees $\mathbb{T}_i$, which are independently of each other learned with the respective subset of the large data set. Both steps describe our HCC approach and generate the complete model. Prototype-to-prototype distances and the placement of new data points in the clustering can be calculated as described in 2.2. This procedure reduces the complexity for the local trees if the global tree is of moderate size and scales with the number of leafs in $\mathbb{T}_G$ with regard to memory restrictions and computation time. Further, the independence of the local trees allows direct parallelization. Moreover the dimensionality D of the subsets V_i can be reduced in general, because of the appearance of dimensions with no or constant entries, resulting in a sparsely populated data matrix, further reducing the model complexity. In this way the subtrees $\mathbb{T}_i$ exist in a lower dimensional space. Visualizations of ET and $\mathbb{T}_G$ (with subtrees) are obtained by calculating the prototype tree distances projected by MDS to 1D or 3D.

3. DATA AND EXPERIMENTAL SETTINGS

We apply our approach on mass spectrometric data from two matrix assisted laser desorption ionization (MALDI) imaging experiments. The original data are tissue slices from rat brain and breast cancer tissues measured by a mass spectrometer and preprocessed as described in [8]. The data after preprocessing are given as peak lists with pairs $(m/c, I)$ indicating a mass position and a corresponding intensity. All peak lists are mapped on a common mass axis using a tolerance of 500 ppm. The first data set is from a rat and denoted as D1. It is still with a low resolution of 1062 measurement and $D = 121$. The second data set is taken from a breast cancer tissue, denoted as D2 and contains 100594 spectra with $D = 76$. All data sets have been analyzed using the above mentioned methods. It should be noted that D1 is still processable with standard tools and D2 is already very challenging for regular methods. Both data sets should be considered as toy examples without a stronger interpretation objective. Standard parameter settings are as follows: the branching factor was fixed to 3 for all regular ET trees and 20 for the $\mathbb{T}_G$. The maximal number of iterations until convergence is $3e6$. Standard ET settings (see [9]) are defined such that all ET trees have in average 40 leafs. Neigborhood range $\sigma = 2$ and the start learning rate $\epsilon = 0.1$.

4. EXPERIMENTAL RESULTS AND INTERPRETATION

Table 1 shows the computation times for the tree model decomposition of D2 in comparison with a standard ET (the model complexity was chosen such that it was still computable), we also show a result for tree with higher depth utilizing the new freedom in the calculation complexity. One observes that the processing time is significantly reduced for our approach. All obtained models have been additionally evaluated visual in comparison to known ground truth labeling provided by an expert. We found that all models performed well in clustering the data in a structural or biological meaningful way but with HCC in a significant quicker time. Considering the ongoing technological progress in this field, especially in MALDI imaging, resolutions of $10\mu m$ are already appearing, leading to at least 1 million spectra per cm^2 of tissue such that our method becomes a valuable analysis approach.

Approach	Number of nodes	calculation time
ET	184	10624 sec
HCC (simple)	159	713 sec
HCC (deep)	1244	2208 sec

Table 1. Comparison of the computation times

For D1 all approaches learned the clustering very quick, the corresponding coloring in comparison to the regular microscope image is shown in Figure 1 including a sketch of a hierarchical analysis of the data. Employing the given topology of ET, motivated by an expected inherent hierarchical ordering of the data, we are able to browse through the different clusters of the ET which nicely corresponds to different sub structures in the original data.

Figure 2 shows a calculated clustering and colorization of D2 using the HCC approach. The right plot in the same figure shows the corresponding microscope image, partially labeled by an expert. One can clearly identify the connnective tissue region (dark intensities). In a RGB

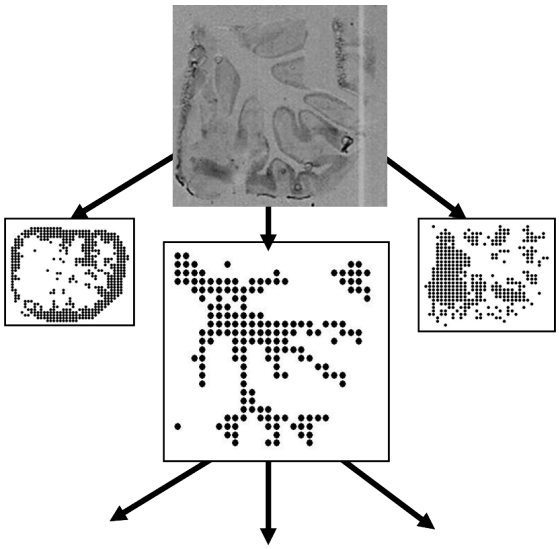

Figure 1. Clustering result for D1 obtained from the ET model and employing the hierarchy.

colored version[2] also the inflamation region (lower center) and large parts of the cancer regions can be distinguished.

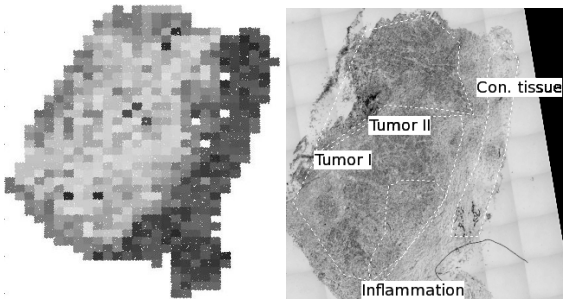

Figure 2. Clustering for D2 obtained by HCC, the ET model is similar; corresp. microscope image (right)

In the HCC approach it is possible that the first hierarchical level (step one of HCC) generates a sufficiently precise model of the problem on the basis of the rough level data set whereas the precise adaptation is delegated to the local trees (step two of HCC). Moreover, the independence of the local learning allows parallelization.

Figure 3 shows an approximated number of weight updates for a given dataset, depending on the number of leafs in $\mathbb{T}_G$, not incorporating possible parallelization. As expected, there exists an optimal number of nodes in the global tree, since a one-prototype global tree (see Figure 3 left side) equates to a fully learned Evolving Tree and the other extreme, giving the global tree the desired size of the full model (see Figure 3 right side) equates the fully learned Evolving Tree as well. The influence of faster BMU search and dimensionality reduction is not depicted here, but the overall characteristics are similar. An improvement of two orders of magnitude is observed here.

[2]Available on request

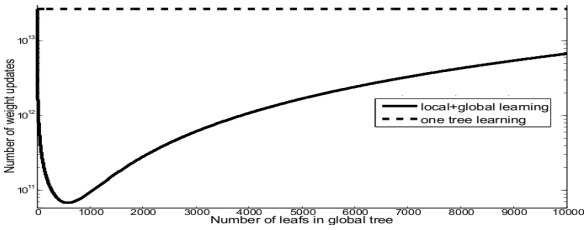

Figure 3. Estimated logarithmic number of weight updates for a dataset of size 1e6 depending on the size of $\mathbb{T}_G$. Dashed line for ET.

5. CONCLUSIONS

We presented a hierarchical composite clustering approach for evolving trees to cluster very large sets of data. Inital experiments show that our approach is effective in computation time as well as clustering performance. We were able to provide a good labelings or data colorings for MALDI imaging data on two data sets in agreement to a provided expert labeling. The different hierarchical levels allow a view onto the data with a specific granularity and a plausible tissue section colorization, being in accordance with the usual analysis approach of pathologists and therefore a progressive step forward to an automatic processing of tissue with high resolution MALDI Imaging. In the next steps we will evaluate our approach on further, larger sets of hierarchical data and derive quality measures for the obtained clusterings. Also the incorporation of label information in the clustering and an effective parallelization of the algorithm, to generate real-time tissue colorizations are of interest.

6. ACKNOWLEDGMENTS

This work was supported by the German Res. Fund. (DFG),HA2719/4-1 (Relevance Learning for Temporal Neural Maps). We would like to thank Dr. Axel Walch (Helmholtz-Zentrum München) for former discussions on the data and Dr. Sören O. Deininger and A. Fütterer (both Bruker Daltonik) for support with MALDI imaging.

7. REFERENCES

[1] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, Springer, corrected edition, July 2003.

[2] T. Kohonen, *Self-Organizing Maps*, vol. 30 of *Springer Series in Information Sciences*, Springer, Berlin, Heidelberg, 1995, (2nd Ed. 1997).

[3] J. Pakkanen, J. Iivarinen, and E. Oja, "The evolving tree—a novel self-organizing network for data analysis," *Neural Process. Lett.*, vol. 20, no. 3, pp. 199–211, 2004.

[4] T. Heskes, "Energy functions for self-organizing maps," in *Kohonen Maps*, E. Oja and S. Kaski, Eds., pp. 303–316. Elsevier, Amsterdam, 1999.

[5] T. Hastie and W. Stuetzle, "Principal curves," *J. Am. Stat. Assn.*, vol. 84, pp. 502–516, 1989.

[6] T. Villmann, R. Der, M. Herrmann, and T. Martinetz, "Topology Preservation in Self–Organizing Feature Maps: Exact Definition and Measurement," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 256–266, 1997.

[7] N. Alex, A. Hasenfuss, and B. Hammer, "Patch clustering for massive data sets," *Neurocomputing*, vol. 72, no. 7-9, Sp. Iss. SI, pp. 1455–1469, MAR 2009.

[8] T. Villmann, F.-M. Schleif, B. Hammer, and M. Kostrzewa, "Exploration of mass-spectrometric data in clinical proteomics using learning vector quantization methods," *Briefings in Bioinformatics*, vol. 9, no. 2, pp. 129–143, 2008.

[9] S. Simmuteit, F.-M. Schleif, T. Villmann, and B. Hammer, "Evolving trees for the retrieval of mass spectrometry-based bacteria fingerprints," *Knowledge and Information Systems*, p. in press, Oct 2009.