TANIMOTO METRIC IN TREE-SOM FOR BETTER REPRESENTATION OF MASS SPECTROMETRY DATA WITH AN UNDERLYING TAXONOMIC STRUCTURE

Stephan Simmuteit(1), *Frank-Michael Schleif* (1)

(1) University Leipzig
Dept. of Medicine
04107 Leipzig, Germany
simmuteit@googlemail.com
schleif@informatik.uni-leipzig.de

Thomas Villmann(2), *Thomas Elssner*(3)

(2) University of Applied Sciences Mittweida
Dep. Mathematics/Physics/Computer Sciences
09648 Mittweida, Germany
villmann@hsmw.de

(3) Bruker Daltonik GmbH
04318 Leipzig, Germany
the@bdal.de

ABSTRACT

In this paper we develop a Tanimoto metric variant of the Evolving Tree for the analysis of mass spectrometric data of animal fur. The Evolving Tree is an extension of Self-Organizing Maps developed to analyze hierarchical clustering problems. Together with the Tanimoto similarity measure, which is intended to work with taxonomic structured data, the Evolving Tree is well suited for the identification of animal hair based on mass spectrometry fingerprints. Results show a suitable hierarchical clustering of the test data and also a good retrieval capability with a logarithmic number of comparisons.

1. INTRODUCTION

The identification of animal fur is important for verification of authenticity of fur or for finding illegal imports or detection of imitated fur, as recently shown in [1, 2]. The classification of fur by visual inspection is not obvious in some cases, e. g. for assembled products. The utilization of mass spectrometry (MS) provides a fast and reproducible way to receive a MS fingerprint of digested animal hair. Spectra with closely related species should have similar spectra, whereas further related species should have less similar spectra, which is well known, supported by keratin analysis as shown in [1]. Figure 1 demonstrates the workflow from animal fur to MS measure.

The standard Euclidean metric does not provide this in case of very high-dimensional data, because the data basis is too sparsely populated [3]. Tanimoto developed a measure that retains the taxonomic information [4]. For use in our prototype based methods we need the continuous variant of the Tanimoto metric [5].

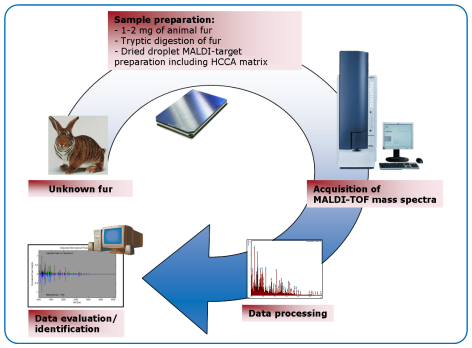

Fig. 1. Standardized workflow of MS based fur analysis

The ideal model for representing taxonomic data is hierarchical. We combine the Evolving Tree (ET) [6], which is a tree-shaped Self-Organizing Map (SOM) [7] with the Tanimoto metric to comply with the requirements referring to taxonomy representation. The Tanimoto coefficient winner determination has been applied to SOMs before in [8], but the update rule adaption by means of Tanimoto gradient descent is new. In [5] the Tanimoto metric has been used with Learning Vector Quantization and its variants.

This contribution provides new aspects for the analysis and representation of animal fur by mass spectrometry technology using a Tanimoto variant of the Evolving Tree.

1.1. Evolving Tree

The adequate model in taxonomic questions is tree structured. The Evolving Tree [6] is a tree structured growing variant of the Self-Organizing Map [7]. The SOM is a projection of high-dimensional vectorial data $\mathbf{v} \in \mathbb{R}^n$ to a predefined m-dimensional grid S with $m << n$. Each node i

has an assigned weight vector $w_i \in \mathbb{R}^n$, which is called prototype. The prototype w_r with the smallest distance to a presented vector $\mathbf{v}$ is the best matching unit (BMU), see equation (1). The distance measure usually used is Euclidean metric.

$$w_r = \arg\min_{i \in S} |\mathbf{v} - \mathbf{w_i}| \tag{1}$$

The set of data points mapped onto the same BMU is the receptive field of this prototype. The ET has a tree structured neighborhood and appends new nodes if a certain condition is satisfied. 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. 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

$$d_{\mathcal{T}}(r, r') = L_{p_{\hat{r} \to r}} + L_{p_{\hat{r} \to r'}} \tag{2}$$

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

$$s_{\mathcal{T}_r}(\mathbf{v}) = \arg\min_{r \in S_r} (d_V(\mathbf{v}, \mathbf{w}_r)) \tag{3}$$

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

$$\Omega_r = \{\mathbf{v} \in V | s(\mathbf{v}) = r\} \tag{4}$$

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

$$\Omega_{r'} = \cup_{r'' \in R_{\mathcal{T}_{r'}}} \Omega_{r''} \tag{5}$$

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:

$$\triangle \mathbf{w}_r = \epsilon h_{SOM}(r, s(\mathbf{v}))(\mathbf{v} - \mathbf{w}_r) \tag{6}$$

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

$$h_{SOM}(r, r') = \exp\Big(\frac{-(d_{\mathcal{T}}(r, r'))^2}{2\sigma^2}\Big). \tag{7}$$

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 as it is known for example from [9].

Since ETs are extended variants of usual SOM one can try to transfer evaluation methods known from SOMs to ETs. Unknown samples can be identified using the ET in the following way. The ET is fully labeled by assignment of a label to each node by an analysis of the receptive fields of the corresponding sub-trees. The root node remains unlabeled. For each receptive field a common label is determined by a majority voting of the contained samples and their labels. An unknown, new item is preprocessed as described later on. For this item the BMU in the tree is determined in accordance to Equation (3) and $s(\mathbf{v})$ is calculated. The label of the receptive field of $s(\mathbf{v})$ defines the label of the item.

1.2. Evolving Tree with Tanimoto metric

The standard Euclidean metric for BMU search (3) in the Evolving Tree is now replaced with the Tanimoto metric in the continuous case (8) [5].

$$d_V^t(\mathbf{v}, \mathbf{w}) = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle - \langle \mathbf{v}, \mathbf{w} \rangle} \tag{8}$$

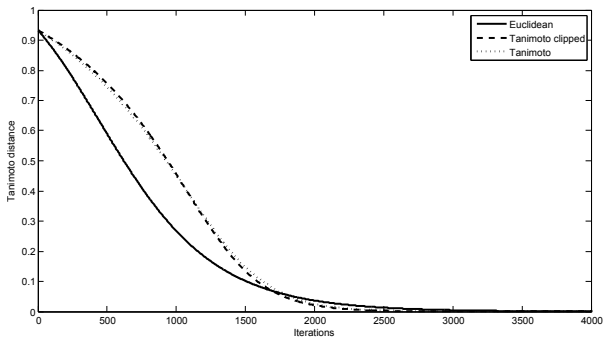

Fig. 2. Plots of Tanimoto distance through time with different update functions

The original tanimoto coefficient was defined for $\mathbf{v}, \mathbf{w} \in \{0,1\}^D$. The continuous case is hence restricted to $\mathbf{v}, \mathbf{w} \in [0,1]^D$. The weight update function (6) is replaced by the gradient (11) with

$$\zeta = \langle \mathbf{v}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle - \langle \mathbf{v}, \mathbf{w} \rangle \tag{9}$$

and

$$\varrho = \frac{\zeta}{\langle \mathbf{v}, \mathbf{w} \rangle} \tag{10}$$

$$\frac{1}{\zeta \varrho} \left((1 + \varrho) \mathbf{v} - 2\mathbf{w} \right). \tag{11}$$

This leads to the new weight update (12).

$$\Delta \mathbf{w} = \epsilon h \frac{\partial d_V^t}{\partial \mathbf{w}} (\mathbf{v}, \mathbf{w}) \tag{12}$$

The gradient descent of the weights points in some case out of the hypercube $[0,1]^D$, then it is appropriate to clip the vector back to the hypercube. Figure 2 shows examples of three update functions and the effect on the tanimoto distance. We recognize an explicit difference between the update functions in speed and dynamic of the process. The data used for this experiment were vectors $a, b \in \{0,1\}^{1000}$ and a randomized distribution of $0/1$.

2. APPLICATION ON MASS SPECTRA OF SOLUTED ANIMAL HAIR

The introduced Tanimoto TreeSOM is now applied on mass spectra of animal hair in solution. The spectra are preprocessed to aligned line spectra as shown in [10]. The preprocessing procedure includes smoothing, baseline reduction, peak picking and the alignment on a global mass axis. The measurements are cross-validated regarding to the respective higher taxonomic level to provide a meaningful evaluation.

Parameter	Value
Threshold	4848
Branching	3
Iterations	331851
Neighborhood width	2.0
Learning rate	0.1

Table 1. Parameter configuration for Tanimoto TreeSOM experiments

2.1. Data

The data used in the experiments are MS-spectra from 46 different animals, 27 canoidae and 19 feloidea. We have about 35 spectra from every animal and a total of 1651 measurements. The dimensionality of the dataset is 1974. The intensities of the peaks are recalculated according to equation (13).

$$v_d = \begin{cases} 1 & v_d > 0 \\ 0 & \text{else} \end{cases} \tag{13}$$

Most of the data points within a specific species are multiple measurements from the same individual. Hence they become either completely included or completely excluded in the tests.

2.2. Experiments

The settings for the tests are shown in Table 1. They are calculated according to equations motivated by [10]. Two species were excluded from the experiment, because they are unique, a crossvalidation experiment would be useless here. Figure 3 shows a typical tree with Tanimoto distance for the given animal hair MS data. The discrimination of higher taxonomic differences becomes clearly visible. A false assignment occurs at node number 113 with ursidae. In the first level one can observe a good discrimination of dog-like, cat-like and other animals (bears, martens, seals). The tree in Figure 4 is a visualization of a standard Evolving Tree with the same data and parameter settings like the one in Figure 3, but with Euclidean metric. Obviously the taxonomic representation is worse than the on from Tanimoto TreeSOM.

Table 2 shows the result of the crossvalidation experiment. We excluded one complete species per tree build and made a retrieval search with all measures of the excluded species and counted the correct and incorrect assignments. In the Table 2 the first column shows the family of the excluded animal and the second column its taxonomic placement (here: the genus). The two last columns provide the ratio how often the correct taxonomic placement was found. In each case the Tanimoto measure performs better than the Euclidean one. For the genus of *Procynoidae* the euclidean model is unable to generalize. This is reflected by the very

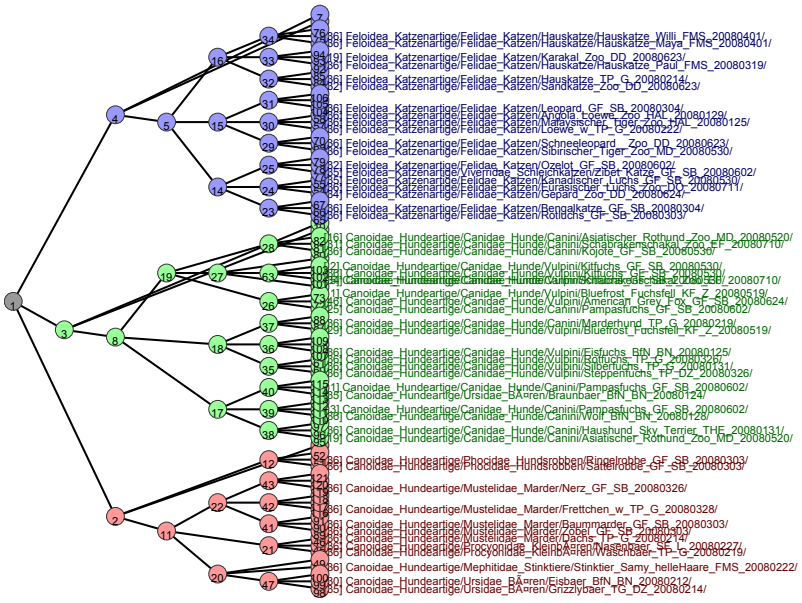

Fig. 3. Tanimoto TreeSOM visualization of the different animals with labeled leafs

Family	Genus	Samples	(Tan)	(Euc)
Canoidae	Canini	244	67.7%	52.2%
Canoidae	Vulpini	292	71.1%	65.3%
Canoidae	Mustelidae	180	95.0%	85.6%
Canoidae	Phocidae	72	100.0%	50.0%
Canoidae	Procyonidae	72	45.9%	1.4%
Canoidae	Ursidae	100	66.7%	43.3%
Feloidea	Felidae	552	100.0%	65.5%

Table 2. Result of the crossvalidation of the second-lowest taxonomic order. The first and third column show the label and number of the measurements, the second column the respective taxonomic classification. The fourth and fifth column shows the percentage of correct classification for tanimoto and euclidean updates, respectively.

low accuracy. The overall result using the Tanimoto measure is 84.8% in contrast to only 60.5% using the euclidean Tree-SOM. 100% correct classification rate is found for *Felidae* (different cat-likes) and *Phocidae* (different seals) using the Tanimoto distance. The classification result in third-lowest taxonomic level (*Canoidae* and *Feloidae*) is always 100%. A classification in the highest taxonomic order, i.e. the species or even subspecies level might be possible, but the data basis is not wide enough to make valid experiments.

3. CONCLUSIONS

A method for an unsupervised analysis of animal hair data from MS has been presented. We compared two different distance measures. It could be shown, that for the family level both measures perform very well, but for the genus level the Tanimoto measure outperforms the euclidean distance. The approach is capable to reflect the underlying hierarchical structure in the of data and allows a retrieval of new data, providing a taxonomic placement, even if no measurement of the particular animal species is in the database an assignment to the most similar taxonomic branch becomes possible[1].

4. REFERENCES

[1] Thomas Elssner, Stefan Klepel, Guido Mix, Klaus Hollemeyer, Wolfgang Altmeyer, and Markus Kostrzewa, "Identification of animal furs by MALDI-TOF mass spectrometry," American Society for Mass Spectrometry Conference (ASMS) 2008, Poster WPUUU 580, 2008.

[2] K. Hollemeyer, W. Altmeyer, E. Heinzle, and Christian Pitra, "Species identification of oetzis clothing with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry based on peptide pattern similarities of hair digests," *Rapid Commun. Mass Spectrom*, , no. 22, pp. 2751–2767, 2008.

[3] M. Verleysen and D. François, "The curse of dimensionality in data mining and time series prediction," in *Computational Intelligence and Bioinspired Systems, Proceedings of the 8th International*

[1]Acknowledgment: We like to thank Markus Kostrzewa and Stefan Klepel (Bruker Daltonik), Barbara Hammer (TU-Clausthal) and Jessica Simmuteit for supporting this work.

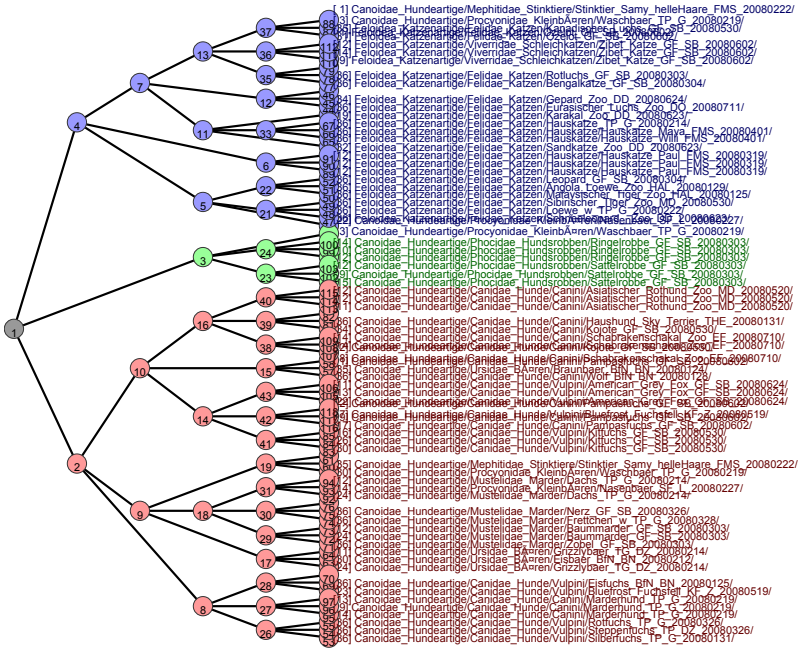

Fig. 4. Euclidean Evolving Tree visualization of the different animals with labeled leafs

Work-Conference on Artificial Neural Networks 2005 (IWANN), Barcelona, J. Cabestany, A. Prieto, and F. S. Hernández, Eds., Berlin, 2005, pp. 758–770, Springer.

[4] T.T. Tanimoto, *An Elementary Mathematical Theory of Classification and Prediction*, IBM Program IBCFL, 1959.

[5] F.-M. Schleif, *Prototype based Machine Learning for Clinical Proteomics*, Ph.D. thesis, Technical University Clausthal, Technical University Clausthal, Clausthal-Zellerfeld, Germany, 2006.

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

[7] Teuvo Kohonen, *Self-Organizing Maps*, vol. 30 of *Springer Series in Information Sciences*, Springer, Berlin, Heidelberg, 1995, (Second Extended Edition 1997).

[8] S. B. Garavaglia, "Statistical analysis of the tanimoto coefficient self-organizing map (TCSOM) applied to health behavioral survey data," in *Proceedings of the International Joint Conference on Neural Networks*. Integrated Therapeutics Group, Schering-Plough Corporation, 2001, vol. 4, pp. 2483–2488.

[9] H.-U. Bauer and Th. Villmann, "Growing a Hypercubical Output Space in a Self–Organizing Feature Map," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 218–226, 1997.

[10] Stephan Simmuteit, Frank-Michael Schleif, Thomas Villmann, and Markus Kostrzewa, "Hierarchical pca using tree-som for the identification of bacteria," in *Proceedings of WSOM 2009*. 2009, p. in press, Springer.