Data Dependent Evaluation of Dissimilarities in Nearest Prototype Vector Quantizers Regarding Their Discriminating Abilities

M. Kaden*, D. Nebel*, F. Melchert†‡, A. Backhaus†, U. Seiffert†, and T. Villmann*

*University of Applied Sciences Mittweida
Computational Intelligence Group
Mittweida, Germany

†Fraunhofer IFF Magdeburg
Biosystems Engineering
Magdeburg, Germany

‡University Groningen
Intelligent Systems Group
Groningen, The Netherlands

***Abstract*—In this paper we propose a rank measure for comparison of (dis-)similarities regarding their behavior to reflect data dependencies. It is based on evaluation of dissimilarity ranks, which reflects the topological structure of the data in dependence of the dissimilarity measure. The introduced rank measure can be used to select dissimilarity measures in advance before cluster or classification learning algorithms are applied. Thus time consuming learning of models with different dissimilarities can be avoided.**

I. Introduction

Similarity or dissimilarity based data processing is one of the key features in almost all areas of data analysis and visualization [1]. A comparison of data in terms of numerical quantities is frequently the basic principle for any regarding algorithmic approach including machine learning. For example, prototype based vector quantizers in unsupervised and supervised learning as well as nearest neighbor approaches rely in appropriately chosen (dis-)similarity measures such that classes can be discriminated or clusters become separated as well as possible [2], [3]. Generally, all variants of nearest neighbor classifiers (NPC, [4]) and clustering approaches crucially depend on the used dissimilarity or similarity.

There exists a broad range of those measures reflecting different aspects like symmetry, transitivity and others. The most prominent are the Euclidean distance or the correlation measure. Kernel distances became popular during the last years whereas divergences and mutual information are used in information theoretic data analysis [5], [6], [7]. A systematic categorization scheme of (dis-)similarity types regarding their mathematical properties is provided in [8].

Other decision approaches like linear discriminant analysis or support vector machines (SVM) do not belong to the class of (dis-)similarity based approaches. Here the classification decision is made in dependence on the localization of an object regarding a separating hyperplane. The key ingredients for those approaches are semi-inner products, which are not (dis-)similarities in general [8]. However, semi-inner products are directly related to norms and, hence, also to distances [9].

Unfortunately, it is difficult to decide in advance, which measure is the best one for a given task. Here, the experience of the researcher is important to find a suitable measure. Yet, frequently one has to consider and to compare several measures for a given problem, which usually requires a time consuming learning process for adaptive machine learning models. One possibility for problem reduction is the application of adaptive (parameterized) measures which are optimized during the learning process according to the specific task [10], [11], [12]. Another idea is to combine several measures in one model like in fusion trees or as a linear combination and to adapt the respective coefficients [13], [14], [15]. The latter strategy requests, however, again a thorough selection of measures to be involved. More precisely, if all used measures would share the same (mathematical) properties, the combination thereof can not generate better solutions.

Hence, there is a need for a practicable scheme to get an advice in advance. This scheme should suggest a subset out of a given set of measures at a glance for reasonable computational costs. One possible strategy is the topic of this paper. In particular, we propose to compare (dis-)similarities preselected for a given data analysis task. If the measures behave differently we can assume that they reflect distinct properties of the data. Only in that case a nearest neighbor classifier will deliver different results. Hence, a diverse collection of dissimilarities would be the basis to built up a linear combination which offers the possibility for substantial improvement [15]. To evaluate the degree of diversity for dissimilarities, we propose in this paper a rank measure, which allows to compare (dis-)similarities based on the rank statistics. Thus we are able to select diverse dissimilarity measures in advance before time-consuming training of a classifier model takes place.

The paper is organized as follows: first, we briefly review basic properties of (dis-)similarities. Thereafter, we introduce the rank measure for comparison of pairs of dissimilarities. Based on this rank measure we compare the behavior of

[0]978-1-5090-6638-4/17/$31.00 ©2017 IEEE

several dissimilarities regarding two real world datasets. After a respective selection of the most differently behaving dissimilarities, these were combined in a median variant of learning vector quantization to obtain an optimal linear combination for best class discrimination.

II. Comparison of Dissimilarities Based on Rank Statistics

A. The General Absolute Rank Equivalence Measure

Similarities as well as dissimilarities can be categorized regarding their mathematical properties like boundedness, definiteness, symmetry or the validity of the triangle inequality [16]. There are different possibilities to transfer a similarity to dissimilarity and vise versa. A simple one is: if we assume a bounded similarity measure $s(x,y)$ with the upper bound b_u then $d_s(x,y) = b_u - s(x,y)$ will obviously establish a dissimilarity measure. Therefore, we will concentrate in the following to dissimilarity measures with the further assumption that $d(x,x) = 0$ is valid.

We start with a distance measure d for a non-empty dataset X, which fulfills all mathematical properties of be a metric and, thus the pair (X,d) is a metric space. Transformations of data keeping their distance relations are therefore denoted as isometric transformations. Otherwise, if we consider another metric δ over X, the pair (X,δ) is another metric space different from (X,d). Hence, the transition from one metric space (X,d) to another one (X,δ) can also be seen as a formal data transformation.

A weaker concept than isometric between metric spaces for data comparison is the concept of topological spaces, which essentially only compare neighborhood relations instead of distance values. Suppose a (symmetric) dissimilarity measure d and for an arbitrary $x \in X$ the open ball around x as $B(x,\varepsilon) = \{y \in X | d(x,y) < \varepsilon\}$. Then the system

$$\tau_d(x) = \{U(x) \subseteq X | x \in U(x) \text{ and } \exists \varepsilon > 0 : B(x,\varepsilon) \subset U\} \tag{1}$$

determines a *local* topology on X with the neighborhood $U(x)$. If X is not continuous, $\tau_d(x)$ is denoted as a discrete topology and $(X, \tau_d(x))$ is a *local* topological space. For those local spaces the concept of neighborhood can be defined as follows: an element $y \in X$ is denoted as rank-one neighbor of $x \in X$ if for all $z \in X$ with $z \neq x$ the relation $d(x,y) \leq d(x,z)$ is valid. We denote this fact by $r_x(y,(X,\tau_d(x))) = 1$ whereas $r_x(x,(X,\tau_d(x))) = 0$. Further, we iteratively define: An element $y \in X$ has the neighborhood rank $r_x(y,(X,\tau_d(x))) = k$ if y is rank-one neighbor for $X \setminus N_k(x)$, where the set $N_k(x) = \{z \in X | r_x(z,(X,\tau_d(x))) < k\}$ contains all neighbors of x with lower ranks than k.

Two discrete local topologies $\tau_d(x)$ and $\tau_{\widehat{d}}(x)$ generated by the dissimilarities d and $\widehat{d}$ are said to be locally equivalent for the dataset X with respect to $x \in X$, if for all $y \in X$ the neighborhood ranks are equal, i.e. $r_x(y,(X,\tau_d(x))) = r_x\left(y,\left(X,\tau_{\widehat{d}}(x)\right)\right)$ is valid for all $y \in X$. Consequently, two discrete local topologies $\tau_d(x)$ and $\tau_{\widehat{d}}(x)$ are said to be globally equivalent for the dataset X if they are equivalent with respect to all $x \in X$.

Since the above topologies are determined by the generating dissimilarities we analogously define rank-equivalence for dissimilarities. More precisely, two dissimilarities are defined to be locally rank-equivalent for the (discrete) dataset X with respect to $x \in X$ iff the related discrete local topologies $\tau_d(x)$ and $\tau_{\widehat{d}}(x)$ are locally equivalent. The global equivalence is defined accordingly. We denote this relation between the dissimilarities by $d \underset{X}{\doteq} \widehat{d}$.

Based on these definitions we are able to introduce a comparison measure for dissimilarities d and $\widehat{d}$ to judge their behavior regarding a given dataset X of N_X samples. For this purpose, we calculate the dissimilarity matrices $\mathbf{D}(X)$ with $d_{ij} = d(x_i,x_j)$ and $\hat{\mathbf{D}}(X)$ with $\hat{d}_{ij} = \hat{d}(x_i,x_j)$: such that the corresponding rank matrices $\mathbf{R}(X)$ and $\hat{\mathbf{R}}(X)$ are determined by

$$r_{ij}^{d} = r_{x_i}(x_j,(X,\tau_d(x_i))) \tag{2}$$

and

$$r_{ij}^{\hat{d}} = r_{x_i}\left(x_j,\left(X,\tau_{\hat{d}}(x_i)\right)\right) \tag{3}$$

respectively. We can calculate the matrix entries r_{ij}^{d} and $r_{ij}^{\hat{d}}$ as

$$r_{ij}^{d} = \sum_{k=1}^{N_X} H(d_{ij} - d_{ik}) \text{ and } r_{ij}^{\hat{d}} = \sum_{k=1}^{N_X} H\left(\hat{d}_{ij} - \hat{d}_{ik}\right) \tag{4}$$

using the Heaviside-Function

$$H(z) = \begin{cases} 1 & \text{for } z > 0 \\ 0 & \text{otherwise} \end{cases}$$

as it was suggested in [17].

We emphasize at this point again that comparing ranks is weaker than consideration of dissimilarities. To demonstrate this effect, we have visualized in Fig.1 the dissimilarity matrix for the Euclidean distance d_E and the RBF-kernel dissimilarity

$$d_{RBF,\sigma}(x,y) = 1 - \exp\left(\frac{-d_E^2(x,y)}{\sigma}\right) \tag{5}$$

for a real world dataset X described more detailed in the application section (sugar dataset). Although the resulted dissimilarity matrices look differently, the corresponding rank matrices seem to be equal, see Fig.1.

An intuitive measure to compare two dissimilarity rank matrices $\mathbf{R}_d(X)$ and $\mathbf{R}_{\hat{d}}(X)$ is

$$\widetilde{\vartheta}_X\left(p,\mathbf{D},\widehat{\mathbf{D}}\right) = \sum_{i=1}^{N_X}\sum_{j=1}^{N_X}\left|r_{ij}^{d} - r_{ij}^{\hat{d}}\right|^p$$

as the p-th power of the p-norm with the usual choices $p = 2$ obeying the squared error or $p = 1$ yielding the absolute error. In the following we will use the latter option taking the short hand notation $\widetilde{\vartheta}_X\left(\mathbf{D},\widehat{\mathbf{D}}\right) = \widetilde{\vartheta}_X\left(1,\mathbf{D},\widehat{\mathbf{D}}\right)$.

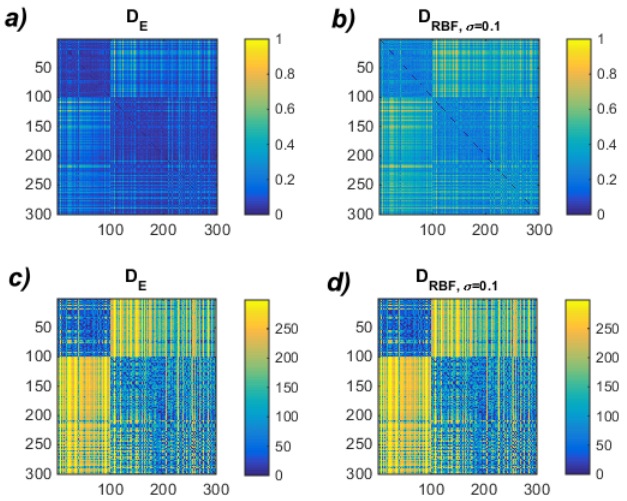


Figure 1. Visualization of the dissimilarity matrices for the Euclidean distance ($\mathbf{D}_E$) **a)** and the RBF-dissimilarity ($\mathbf{D}_{RBF}$) **b)** regarding the sugar dataset using 100 data points per class. Although these matrices look different the respective rank matrices ($\mathbf{R}_{d_E}$) and ($\mathbf{D}_{d_{RBF}}$) depicted in **c)** and **d)** are equal.

To be independent of the number N_X of data we introduce the normalization

$$\vartheta_X\left(\mathbf{D},\widehat{\mathbf{D}}\right)=\frac{\widetilde{\vartheta}_X\left(\mathbf{D},\widehat{\mathbf{D}}\right)}{c_X} \tag{6}$$

with

$$c_X=\begin{cases}4\left(\frac{N_X}{2}-1\right)\frac{N_X}{2}-\left(\frac{N_X}{2}+1\right) & \text{, if } N_X \text{ is even}\\ 4\left(\frac{N_X-1}{2}-1\right)\frac{N_X-1}{2} & \text{, if } N_X \text{ is odd}\end{cases}$$

as the normalization constant. We denote $\vartheta_X\left(\mathbf{D},\widehat{\mathbf{D}}\right)$ as the absolute rank equivalence measure (ARE).

Thus, the ARE measure $\vartheta_X\left(\mathbf{D}_E,\mathbf{D}_{RBF}\right)$ of the distance matrices given in Fig.1 is zero, which underlines the previous observation of equal rank matrices for the sugar dataset.

B. Variants of the Absolute Rank Equivalence Measure

So far, the ARE measure takes all data points of the set X into account. This is maybe too restrictive. In the following we present variants of the ARE measure, which might be useful for specific tasks.

Considering the frequently applied k-nearest-neighbor classifier, it would be desirable to compare dissimilarities on this level. Therefore, we propose for this case the k-ARE measure involving only the first k ranks. It is calculated as

$$\widetilde{\vartheta}_{X,k}\left(\mathbf{D},\widehat{\mathbf{D}}\right)=\sum_{i}^{N_X}\sum_{j}^{N_X}\gamma_{k(i,j)} \tag{7}$$

with

$$\gamma_{k(i,j)}=\begin{cases}\left|r_{ij}^{d}-r_{ij}^{\hat{d}}\right| & \text{if } r_{ij}^{d}<k\\ 0 & \text{otherwise.}\end{cases}$$

The normalization from (6) has to be adapted to

$$\vartheta_{X,k}\left(\mathbf{D},\widehat{\mathbf{D}}\right)=\frac{\widetilde{\vartheta}_{X,k}\left(\mathbf{D},\widehat{\mathbf{D}}\right)}{c_{X,k}} \tag{8}$$

with

$$c_{X,k}=N_x\cdot k^2$$

as the new normalization constant.

A class dependent variant of the k-ARE is obtained by

$$\psi_{X,k}\left(\mathbf{D},\widehat{\mathbf{D}}\right)=\frac{\sum_i\sum_j\tau_{k(i,j)}}{c_{X,k}}$$

with

$$\tau_{k(i,j)}=\begin{cases}\left|r_{ij}^{d}-r_{ij}^{\hat{d}}\right| & \text{if } r_{ij}^{d}<k \text{ and } c\left(x_i\right)\neq c\left(x_j\right)\\ 0 & \text{otherwise.}\end{cases}$$

where $c\left(x_i\right)$ denotes the class label of the data point x_i. The k-ARE measure considers only the rank difference r_{ij} if the data points x_i and x_j belonging to difference classes.

III. Applications

We demonstrate the application of the absolute rank equivalence measure for two real world datasets. The first one is a publicly available benchmark dataset that compiles the reflectance spectra of sugar and sugar related compounds. Here we investigate the behavior of different dissimilarities. The second one is a set of laser-diffuse-light images of Salmonella serovar colonies. For this dataset we investigate the parameter dependent behavior in case of a parametrized dissimilarity.

A. Investigation of Different Dissimilarities for the Analysis of the Sugar Dataset

The publicly available Sugar dataset is consist of reflectance spectra of nine different sugars and sugar related compounds and is described in [18]. A special feature for this dataset is that the spectral information is not only acquired with a single sensor, but with different sensors covering a wide spectral range. Here we make only use of a subset of the data, namely only data acquired by the Neo VNIR 1800 Sensor. The spectral information consists of 186 sampling points equidistantly distributed over the range from 400 to 1000nm wavelength and normalized according to the l_1-norm. Although the spectra are obtained from 9 different compounds, the classification is reduced to a three class problem only considering the three main classes sugar esters, sugar alcohols and sugars.

The dissimilarities in consideration for this dataset are depicted in Tab.I. The resulting dissimilarity matrices are depicted in Fig.2 whereas the respective rank matrices are visualized in Fig.3. We already know from Fig.1 that the Euclidean distance d_E and the RBF-dissimilarity d_{RBF} generate the same ranks. Further, we observe similar behavior for the Euclidean distance d_E and the Cauchy-Schwarz-divergence d_{CS}. This observation is verified by the ARE measures given

dissimilarity measure	formula
Euclidean distance	$d_E(x,y) = \sqrt{\sum_{i=1}^{n} \lvert[x]_i - [y]_i\rvert^2}$
l_1-distance	$d_{l_1}(x,y) = \sum_{i=1}^{n} \lvert[x]_i - [y]_i\rvert$
l_{max}-distance	$d_{l_{max}}(x,y) = \max_{i=1,\ldots,n} \lvert[x]_i - [y]_i\rvert$
Sobolev distance	$d_{sob}(x,y) = \frac{1}{2} d_E(x,y) + \frac{1}{2} d_E(x',y')$ with $z' = \frac{dz}{dt}$
RBF-distance	$d_{RBF,\sigma}(x,y) = 1 - \exp\left(\frac{-d_E^2(x,y)}{\sigma}\right)$
Pearson correlation	$d_{Pears}(x,y) = 1 - \frac{\sum_{i=1}^{n}([x]_i - \mu_{\mathbf{x}})([y]_i - \mu_{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n}([x]_i - \mu_{\mathbf{x}})^2}\sqrt{\sum_{i=1}^{n}([y]_i - \mu_{\mathbf{y}})^2}}$
Spearmann correlation	$d_{spear}(x,y) = 1 - \left(\frac{cov(rg(x),rg(y))}{\sigma_{rg(\mathbf{x})}\sigma_{rg(\mathbf{y})}}\right)^2$ with $rg(z)$ is the Euclidean dissimilarity rank of z
Cauchy-Schwarz-divergence	$d_{CS}(x,y) = d_{\gamma=1}(x,y)$

Table I
DISSIMILARITIES USED FOR THE SUGAR DATASET. HERE $[x]_i$ DENOTES THE iTH DIMENSION OF THE n-DIMENSIONAL VECTOR x.

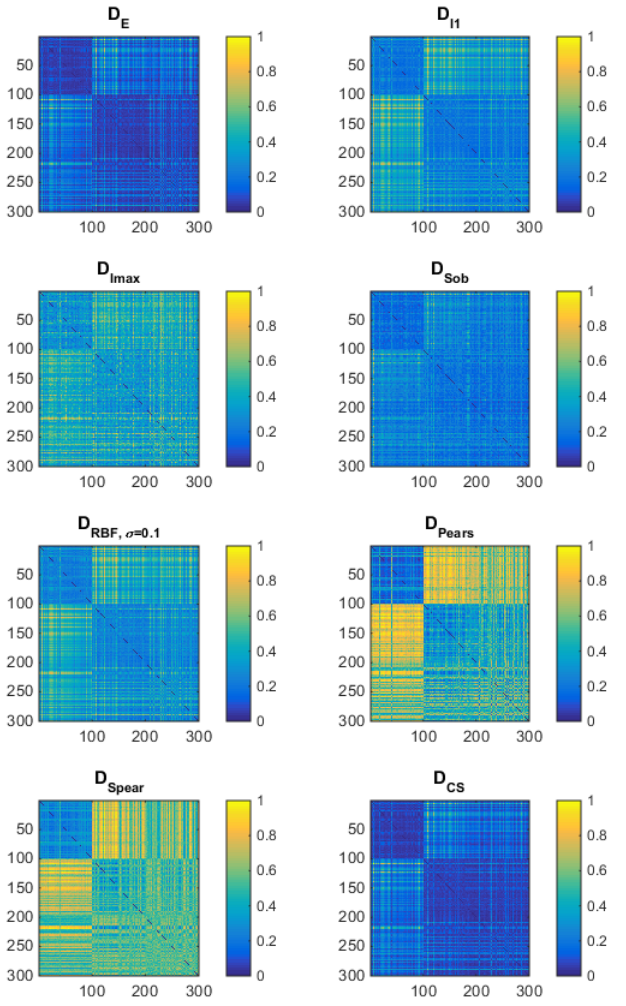


Figure 2. Visualization of the dissimilarity matrices for the sugar dataset using 100 data points per class.

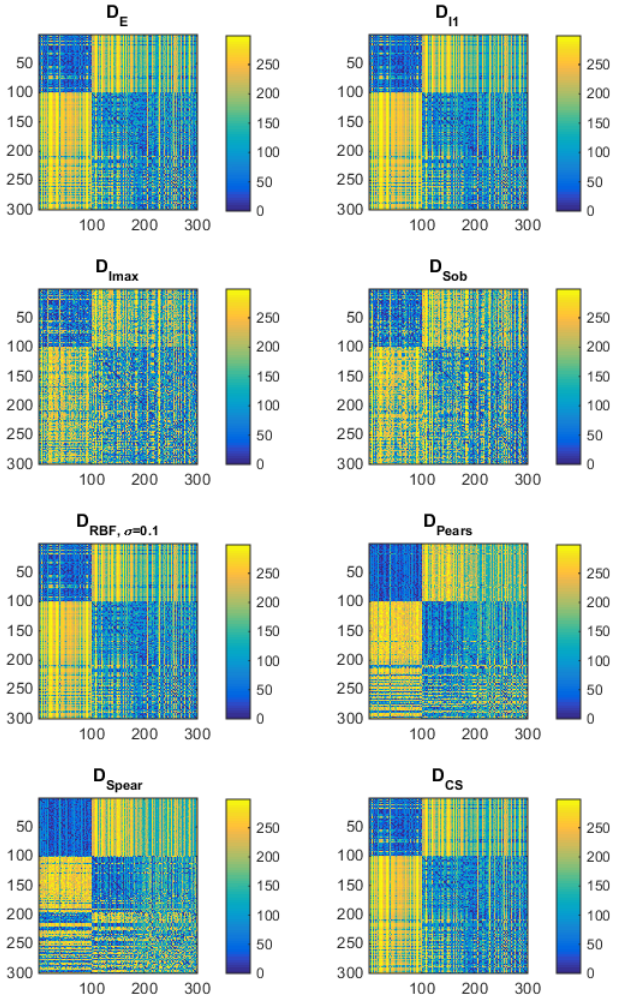


Figure 3. Visualization of the rank matrices for the sugar dataset using 100 data points per class.

	k-NN	stdv	MGLVQ	stdv	$\hat{\lambda}_j$	λ_j
$d_1 = d_E$	81.2	±0.6	81.1	±2.4	0.120	0.315
$d_2 = d_{l_1}$	81.6	±0.3	82.0	±1.6	0.104	0.293
$d_3 = d_{l_{max}}$	74.9	±1.2	71.3	±3.9	0.018	0.002
$d_4 = d_{Sob}$	74.6	±2.2	72.0	±3.5	0.066	0.126
$d_5 = d_{RBF}$	81.2	±0.6	80.5	±2.1	0.073	–
$d_6 = d_{Pears}$	79.4	±1.0	76.8	±3.0	0.033	0.100
$d_7 = d_{Spear}$	72.7	±1.1	74.8	±2.6	0.066	0.168
$d_8 = d_{CS}$	81.2	±0.6	80.8	±2.0	0.521	–
$d_{\hat{\lambda}}$	80.1	±1.1	82.6	±2.0		
d_{λ}	79.6	±1.8	82.5	±1.9		

Table II
SUGAR DATASET: CLASSIFICATION ACCURACIES IN % OBTAINED BY k-NN WITH $k = 3$ AND MGLVQ FOR THE USED DISSIMILARITIES AS WELL AS FOR THE LINEAR COMBINATIONS OF ALL DISSIMILARITIES $d_{\hat{\lambda}}(x,y) = \sum_j \hat{\lambda}_j d_j(x,y)$ AND OF THE REMAINING DISSIMILARITIES $d_{\lambda}(x,y) = \sum_j \lambda_j d_j(x,y)$. THE NORMALIZED LINEAR COMBINATION COEFFICIENTS $\lambda_j \geq 0$ AND $\hat{\lambda}_j \geq 0$ WERE DETERMINED BY GRADIENT DESCENT LEARNING, SEE TEXT.

in Fig.4. Hence, the RBF-distance can be dropped in this application.

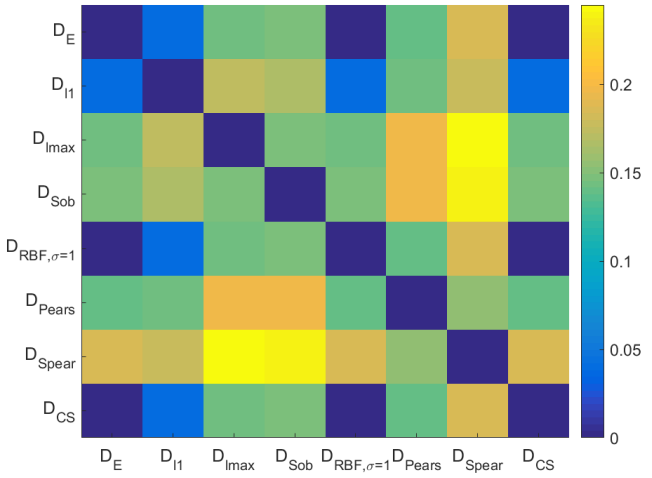


Figure 4. Visualization of ARE measures of the sugar dataset.

In the next step we compared the remaining dissimilarities regarding their behavior when applied in a k-nearest neighbor classifier and a median generalized learning vector quantizer (MGLVQ) [19]. The results achieved for 10 runs with random $50\% : 50\%$-splits are given in Tab.II. For the MGLVQ only one prototype per class was used. Further, we trained a MGLVQ using a linear combination $d_{\lambda}(x,y) = \sum_j \lambda_j d_j(x,y)$ of all remaining dissimilarities as dissimilarity measure where the coefficients $\lambda_j \geq 0$ were determined by (stochastic) gradient descent learning of the combined measure in MGLVQ according to [20] and normalized such that $\sum_j \lambda_j = 1$. The resulting combined dissimilarity $d_{\lambda}(x,y)$ was also applied for the k-NN. The results are depicted in Tab.II. We can see that the linear combination of the remaining different dissimilarities leads to an improvement of the classification accuracy. A linear combination of all dissimilarities does not yield a further improvement as we learn from Tab.II. Thus, the preselection of dissimilarities using the ARE-measure delivers a set of dissimilarities with different characteristics to be applied for best classification performance.

B. Investigation of a Parametrized Dissimilarity for the Analysis of Salmonella Serovars Images

The set of *Laser-Diffuse-Light-Images* (LDL-images) of Salmonella serovars colonies was This dataset was already investigated and described in detail in [19]. In particular, the set provides images from three salmonella types *Salmonella Brandenburg* (SB), *Salmonella Enteritidis* (SE), and *Salmonella Typhimurium* (ST). The task is to distinguish serovars based on the structure of the colonies, see Fig.5.

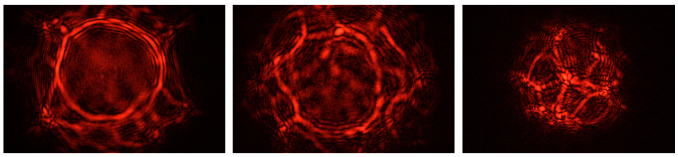


Figure 5. LDL-images of bacteria colonies of Salmonella serovars: Salmonella Brandenburg, Salmonella Enteritidis, and Salmonella Typhimurium (from left to right).

The raw image data of the dataset were manually centered. After normalization and calibration the gray-scale-images are of size 128×128, which are taken as matrices. Overall, the dataset consists of 65 SB-samples, 48 SE-samples and 50 ST-samples. Here we used the images as grey-scale images, which are compared using the γ-divergence

$$d_{\gamma}(x,y) = \frac{\log\left(\frac{\left(\sum_{i=1}^{n}[x]_i^{\gamma+1}\right)\left(\sum_{i=1}^{n}[y]_i^{\gamma+1}\right)^{\gamma}}{\sum_{i=1}^{n}[x]_i[y]_i^{\gamma}}\right)}{\gamma(\gamma+1)} \tag{9}$$

with the parameter $\gamma \geq 0$ [21], [7]. In the limit $\gamma \searrow 0$ the γ-divergence becomes the Kullback-Leibler-divergence whereas for $\gamma = 1$ the Cauchy-Schwarz-divergence is obtained [5]. Moreover, the γ-divergence is not symmetric. In the experiment we varied the γ-parameter between 1 and 5 with $\Delta\gamma = 0.1$. The resulting ARE-matrix is visualized in Fig.6. We can detect in this visualization several blocks indicating similar behavior for the respective γ-divergences as dissimilarity measure. Therefore we selected for the classification task a representative γ-value for each block to be used for γ-divergences and obtain 9 different γ values instead of 41. As an additional dissimilarity we also used the Euclidean distance d_E as reference. Again we applied a MGLVQ and a k-NN for classification learning regarding each dissimilarity. As before, we also learned a linear combination d_{λ} of all 9 divergences and d_E by means of stochastic gradient descent optimization in MGLVQ. The results are depicted in Tab.III. Here the k-NN can not benefit from the linear combination of the γ-divergences whereas MGLVQ again performs better when the linear combination of the divergences is applied.

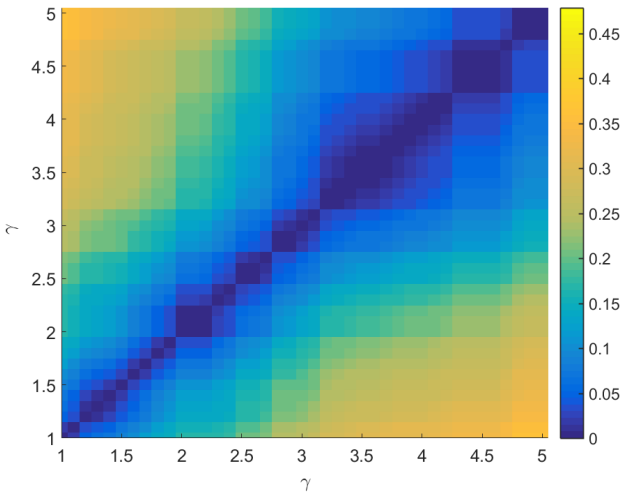


Figure 6. Visualization of the k-ARE measure matrix ($k = 1$) for different γ-divergences of the Salmonella dataset. It is similar to the class dependent k-ARE measure.

	k-NN	stdv	MGLVQ	stdv	λ_j
$d_1 = d_E$	76.5	±4.0	83.3	±3.4	0.014
$d_2 = d_{\gamma=1}$	76.5	±4.0	83.6	±3.3	0.186
$d_3 = d_{\gamma=1.25}$	76.4	±3.8	82.5	±5.2	0.016
$d_4 = d_{\gamma=1.5}$	76.3	±4.1	82.3	±5.3	0.017
$d_5 = d_{\gamma=1.75}$	76.2	±4.2	83.4	±5.3	0.017
$d_6 = d_{\gamma=2.0}$	76.5	±4.2	83.1	±2.4	0.079
$d_7 = d_{\gamma=3.0}$	77.0	±4.3	82.9	±5.3	0.139
$d_8 = d_{\gamma=3.5}$	77.1	±4.3	83.6	±3.2	0.154
$d_9 = d_{\gamma=4.5}$	77.1	±4.3	81.2	±5.9	0.378
d_λ	76.5	±4.0	86.0	±2.7	

Table III
SALMONELLA DATASET: CLASSIFICATION ACCURACIES IN % OBTAINED BY k-NN WITH $k = 3$ AND MGLVQ FOR THE USED DISSIMILARITIES AS WELL AS FOR THE LINEAR COMBINATION $d_\lambda(x, y) = \sum_j \lambda_j d_j(x, y)$ OF ALL DISSIMILARITIES.

IV. CONCLUSIONS

In this paper we introduced a measure to compare (dis-) similarities regarding their topological properties for a given dataset. This allows to identify relevant dissimilarity measures for classification and cluster learning algorithms providing different information. Thus, time consuming learning runs with different dissimilarity measures can be avoided in advance. The approach is based on the definition of rank equivalence of dissimilarities. This rank equivalence can be seen as an degree of equivalence of the generated topological structures in the data by means of the compared dissimilarities.

REFERENCES

[1] E. Pekalska and R.P.W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, 2006.

[2] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

[3] T. Villmann, A. Bohnsack, and M. Kaden. Can learning vector quantization be an alternative to SVM and deep learning? *Journal of Artificial Intelligence and Soft Computing Research*, 7(1):65–81, 2017.

[4] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[5] J.C. Principe. *Information Theoretic Learning*. Springer, Heidelberg, 2010.

[6] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[7] T. Villmann and S. Haase. Divergence based vector quantization. *Neural Computation*, 23(5):1343–1392, 2011.

[8] D. Nebel, M. Kaden, A. Bohnsack, and T. Villmann. Types of (dis-)similarities and adaptive mixtures thereof for improved classification learning. *Neurocomputing*, page in press, 2017.

[9] G. Lumer. Semi-inner-product spaces. *Transactions of the American Mathematical Society*, 100:29–43, 1961.

[10] K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[11] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[12] P. Schneider, B. Hammer, and M. Biehl. Distance learning in discriminative vector quantization. *Neural Computation*, 21:2942–2969, 2009.

[13] U. Knauer, A. Backhaus, and U. Seiffert. Fusion trees for fast and accurate classification of hyperspectral data with ensembles of γ-divergence-based RBF networks. *Neural Computing and Applications*, 26(2):253–262, 2015.

[14] U. Knauer, A. Backhaus, and U. Seiffert. Beyond standard metrics - on the selection and combination of distance metrics for an improved classification of hyperspectral data. In T. Villmann, F.-M. Schleif, M. Kaden, and M. Lange, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of 10th International Workshop WSOM 2014, Mittweida*, volume 295 of *Advances in Intelligent Systems and Computing*, pages 167–177, Berlin, 2014. Springer.

[15] E. Mwebaze, G. Bearda, M. Biehl, and D. Zühlke. Combining dissimilarity measures for prototype-based classification. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2015)*, pages 31–36, Louvain-La-Neuve, Belgium, 2015. i6doc.com.

[16] T. Villmann, M. Kaden, D. Nebel, and A. Bohnsack. Similarities, dissimilarities and types of inner products for data analysis in the context of machine learning - a mathematical characterization. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, and J.M. Zurada, editors, *Proceedings of the 15th International Conference Artificial Intelligence and Soft Computing - ICAISC, Zakopane*, volume 2 of *LNAI 9693*, pages 125–133, Berlin Heidelberg, 2016. Springer International Publishing, Switzerland.

[17] Thomas M. Martinetz, Stanislav G. Berkovich, and Klaus J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[18] F. Melchert, A. Matros, M. Biehl, and U. Seiffert. The sugar dataset: A multimodal hyperspectral dataset for classification and research. *Machine Learning Reports*, 10(3):15–18, 2016.

[19] D. Nebel, B. Hammer, K. Frohberg, and T. Villmann. Median variants of learning vector quantization for learning of dissimilarity data. *Neurocomputing*, 169:295–305, 2015.

[20] M. Kaden, D. Nebel, and T. Villmann. Adaptive dissimilarity weighting for prototype-based classification optimizing mixtures of dissimilarities. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2016)*, pages 135–140, Louvain-La-Neuve, Belgium, 2016. i6doc.com.

[21] A. Cichocki and S.-I. Amari. Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12:1532–1568, 2010.

REFERENCES

[1] E. Pekalska and R.P.W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, 2006.

[2] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

[3] T. Villmann, A. Bohnsack, and M. Kaden. Can learning vector quantization be an alternative to SVM and deep learning? *Journal of Artificial Intelligence and Soft Computing Research*, 7(1):65–81, 2017.

[4] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[5] J.C. Principe. *Information Theoretic Learning*. Springer, Heidelberg, 2010.

[6] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[7] T. Villmann and S. Haase. Divergence based vector quantization. *Neural Computation*, 23(5):1343–1392, 2011.

[8] D. Nebel, M. Kaden, A. Bohnsack, and T. Villmann. Types of (dis-)similarities and adaptive mixtures thereof for improved classification learning. *Neurocomputing*, page in press, 2017.

[9] G. Lumer. Semi-inner-product spaces. *Transactions of the American Mathematical Society*, 100:29–43, 1961.

[10] K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[11] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[12] P. Schneider, B. Hammer, and M. Biehl. Distance learning in discriminative vector quantization. *Neural Computation*, 21:2942–2969, 2009.

[13] U. Knauer, A. Backhaus, and U. Seiffert. Fusion trees for fast and accurate classification of hyperspectral data with ensembles of γ-divergence-based RBF networks. *Neural Computing and Applications*, 26(2):253–262, 2015.

[14] U. Knauer, A. Backhaus, and U. Seiffert. Beyond standard metrics - on the selection and combination of distance metrics for an improved classification of hyperspectral data. In T. Villmann, F.-M. Schleif, M. Kaden, and M. Lange, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of 10th International Workshop WSOM 2014, Mittweida*, volume 295 of *Advances in Intelligent Systems and Computing*, pages 167–177, Berlin, 2014. Springer.

[15] E. Mwebaze, G. Bearda, M. Biehl, and D. Zühlke. Combining dissimilarity measures for prototype-based classification. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2015)*, pages 31–36, Louvain-La-Neuve, Belgium, 2015. i6doc.com.

[16] T. Villmann, M. Kaden, D. Nebel, and A. Bohnsack. Similarities, dissimilarities and types of inner products for data analysis in the context of machine learning - a mathematical characterization. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, and J.M. Zurada, editors, *Proceedings of the 15th International Conference Artificial Intelligence and Soft Computing - ICAISC, Zakopane*, volume 2 of *LNAI 9693*, pages 125–133, Berlin Heidelberg, 2016. Springer International Publishing, Switzerland.

[17] Thomas M. Martinetz, Stanislav G. Berkovich, and Klaus J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[18] F. Melchert, A. Matros, M. Biehl, and U. Seiffert. The sugar dataset: A multimodal hyperspectral dataset for classification and research. *Machine Learning Reports*, 10(3):15–18, 2016.

[19] D. Nebel, B. Hammer, K. Frohberg, and T. Villmann. Median variants of learning vector quantization for learning of dissimilarity data. *Neurocomputing*, 169:295–305, 2015.

[20] M. Kaden, D. Nebel, and T. Villmann. Adaptive dissimilarity weighting for prototype-based classification optimizing mixtures of dissimilarities. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2016)*, pages 135–140, Louvain-La-Neuve, Belgium, 2016. i6doc.com.

[21] A. Cichocki and S.-I. Amari. Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12:1532–1568, 2010.