Adaptive Dissimilarity Measures, Dimension Reduction and Visualization

Kerstin Bunte

Book cover: Isosurfaces of different dissimilarity measures on a glass surface:
Front (right to left): Minkowski metric with $p = 1$ and isolevel 0.3 and Itakura-Saito divergence with isolevel 0.2.
Back (right to left): Gamma-divergence with $\gamma = 1.5$ and isolevel 0.02 and Minkowski metric with $p = 3$ and isolevel 0.3.
Alternative and adaptive similarity measures are discussed in this thesis.

Published by *Atto Producties Europe* - www.attoproducties.nl - Groningen

supported by the Netherlands Organisation for Scientific Research (NWO)
under project number 612.066.620

Netherlands Organisation for Scientific Research

RIJKSUNIVERSITEIT GRONINGEN

Adaptive Dissimilarity Measures, Dimension Reduction and Visualization

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
vrijdag 16 december 2011
om 12.45 uur

door

Kerstin Bunte

geboren op 6 september 1981
te Bielefeld, Duitsland

Promotores:	Prof. dr. M. Biehl
	Prof. dr. N. Petkov
Beoordelingscommissie:	Prof. dr. E. Merényi
	Prof. dr. M. Opper
	Prof. dr. M. Verleysen

ISBN: 978-90-367-5186-5

Contents

Acknowledgments

Kerstin Bunte
Groningen
March 23, 2020

List of symbols

$\mathcal{X}$	input space	9
n	number of input vectors	9
$\boldsymbol{x}^i$	i-th example	9
y^i	i-th label	9
N	dimensionality of the data	9
C	number of classes	9
$\boldsymbol{w}^j$	j-th prototype	9
$c(\boldsymbol{w}^j)$	class of j-th prototype	9
$n_{\boldsymbol{w}}$	number of prototypes	9
d	dissimilarity measure	9
R^i	receptive field of prototype $\boldsymbol{w}^i$	10
M	target dimension of dimension reduction	14

List of Figures

List of Algorithms

Chapter 1

Introduction

Due to advanced sensor technology, rapidly increasing digitalization capabilities and the availability of less and less expensive storage volume the amount of data has grown tremendously in the last decades. In the years between 1999 and 2002 an increase of stored information about 30% each year was estimated (Lyman and Varian 2003). Usually this data consists of a variety of measured features leading to also very high dimensional data sets. Manually inspection of the data becomes more costly and automatic methods to help humans to quickly scan through massive data amounts are desirable. This gave rise to many applications in computer science to process the available data: advanced techniques including data mining (Han and Kamber 2005), pattern recognition (Duda et al. 2000) and machine learning (Mitchell 1997, Ripley 1996, Bishop 2006), among others. Even with great progress in those fields the optimization of existing methods and development of novel schemes is highly desirable to perform faster and more efficient data analysis.

The field of machine learning concerns the design of algorithms, which aim at the optimization of adaptive systems on the basis of example data. A model is adapted to learn complex patterns and process new data coming from the same domain better regarding the specified objective. The analysis of patterns involves a number of tasks including data representation, classification, clustering, density estimation, regression, feature extraction and dimension reduction, just to name a few. A lot of data visualization tools have been developed to use cognitive capabilities of humans for structure detection in visual images. Structural characteristics of the data can be captured almost instantly by humans despite the amount of data points which are represented in the visualization. Hence, dimension reduction and visualization are commonly used modern data mining techniques (Lee and Verleysen 2007). Machine learning is broadly categorized into reinforcement, supervised and unsupervised learning. Reinforcement learning is inspired by behaviorist psychology and concerns the finding of suitable actions to maximize some notion of reward (Sutton and Barto 1998). Supervised techniques involve external supervision, which provides correct responses to the given inputs. The aim is usually the discrimination of the categories and to maximize the generalization for novel data. Unsupervised methods, on the other hand, do not need supervision and their goal is the discov-

ery of underlying structures and regularities based on the definition of some basic properties of the data. An elaborate description concerning the history of machine learning can be found in, e. g. (Bishop 1995, Ripley 1996, Mitchell 1997, Duda et al. 2000, Bishop 2006).

A very intuitive supervised technique called k-Nearest Neighbor (k-NN) classifier compares the unknown data to all known examples with respect to some dissimilarity measure (Duda et al. 2000). Obviously the computational effort and memory usage scales with the number of known samples. Therefore prototype-based techniques were developed, which employ representations of data subsets. The prototypes are vector locations in the feature space. They usually serve as typical representatives and reflect the characteristics of the data in their direct neighborhood. Some prominent unsupervised examples are the Self-organizing Map (SOM) (Kohonen et al. 2001) and Neural Gas (NG) (Martinetz and Schulten 1991). And a popular supervised family of such prototype-based classification methods is Learning Vector Quantization (LVQ) (Kohonen et al. 2001). All these methods crucially depend on the distance measure, which is used to adapt the prototype positions and performs the nearest prototype classification. Therefore the learning of adaptive metrics with respect to the given problem at hand was investigated (Xing et al. 2002, Chopra et al. 2005, Frome et al. 2007, Schneider et al. 2009b, Schneider et al. 2009a).

This thesis investigates adaptive dissimilarities and applications varying from classification up to supervised and unsupervised dimension reduction.

1.1 Scope of this thesis

The objective of this thesis is manifold, it contains:

- the introduction of prototype-based adaptive dissimilarity learning with limited rank matrices,
- a new method based on that principle for learning in complex valued data domains and
- a general view and new algorithms for unsupervised as well as supervised dimension reduction and visualization.

Adaptive dissimilarities are a powerful tool, which are shown to improve the performance of supervised methods, such as for example LVQ and the k-NN classifiers. These classification algorithms crucially depend on the distance measure used. Metric adaptation techniques allow the learning of discriminative dissimilarity mea-

sures from a given set of representative example data. Restrictions in adaptive matrix learning, e. g. the limitation of the rank, enables the learning of discriminative global or local linear transformations. These transformations can then be used for supervised dimension reduction and visualization. It also reduces the number of the effective learning parameters, which might be interesting from the computational point of view.

In the first part of this contribution previously proposed methods for metric learning in LVQ are extended to limited rank matrices. Several practical applications are investigated including Content Based Image Retrieval (CBIR), dimension reduction and visualization. Furthermore we provide an extension which can be used on complex valued data shown on an example for texture classification in images.

The second part of this thesis focuses on dimension reduction and visualization. We provide a general view on existing dimension reduction methods, which originally provide just an implicit mapping of the given data points itself. Based on this general principle we extend these methods to learn the parameters of explicit mapping functions instead. This provides direct out-of-sample extensions, reduces computational effort by restricting the learning process just on a small subset of the possible large data set and enables the formal investigation of the generalization ability. Furthermore we provide an unsupervised dimension reduction method, which in contrast to other techniques exhibit a complexity which scales linear with the number of data points in every step. It aims in the combination of fast online learning with the high quality of direct divergence optimization, successfully used by state-of-the-art techniques.

1.2 Outline

Part I

Adaptive Dissimilarity Measures

Chapter 2

Distance Based Classification

Everything has its beauty but not everyone sees it.

Confucius

Abstract

This chapter introduces the basic Learning Vector Quantization (LVQ) algorithms and notations used throughout the thesis. We discuss nearest prototype classification and a set of LVQ learning schemes, which are relevant in the context of this work. Furthermore we explain the concept of parameterized dissimilarity and metric adaptation proposed in the literature.

2.1 Introduction

Machine learning (Mitchell 1997, Bishop 2006) constitutes a huge field in computer science expanding into broad distribution of both, application and theory. The term "learning" comprises the biological point of view by modeling the theory of psychologists of learning in animals and humans. And it also addresses the development of algorithms aiming at the adjustment to a given objective based on empirical data. Thus, from a given set of input/output pairs produced by an complicated unknown process a machine should be able to adjust its internal structure such that the correct output is reproduced for a large number of samples. This part of the thesis concentrates a subfield usually referred to as supervised learning: Samples are given for which the output is (sometimes only approximately) known. The aim is to find a hypothesis that closely agrees with these given data and generalizes well, i.e. produces the desired output also for new samples.

Learning Vector Quantization (LVQ) and its variants constitute a popular family of supervised prototype-based classifiers. The basic algorithm introduced by (Kohonen 1986) is parameterized by a set of labeled prototypes representing the

classes in the input space in combination with a dissimilarity measure. The classification takes places by a nearest prototype scheme, i.e. a new sample is assigned to the class represented by the closest prototype with respect to the given metric. These algorithms are naturally suitable for multi-class problems without changing the learning rules and the complexity is usually dependent on the number of prototypes and only indirect on the number of classes. This classification procedure is closely related to the popular k-Nearest Neighbor (k-NN) approach (Cover and Hart 1967), which keeps the given labeled data set as a reference set and classifies every new data point to the class given by the majority among its k nearest neighbors. Although the k-NN approach is one of the most intuitive and simplest classification algorithms it shows often very good performance. Nevertheless, it might become very expensive in memory usage and computation for very large reference sets. Prototype methods overcome those problems by defining a clustering on the data. Another advantage of LVQ is the interpretability of the resulting parameters: It does not suffer from a "black box" character like an Artificial Neural Network (ANN) or a Support Vector Machine (SVM). The prototypes reflect the characteristic class-specific attributes of the input samples.

The basic heuristic algorithm, called LVQ1 (Kohonen 1986), adapts a set of prototypes from labeled training data by implementing Hebbian learning steps. Additionally, Kohonen introduced two alternative learning schemes: optimized learning-rate LVQ (OLVQ1) and LVQ2.1, aiming at faster convergence and better approximation of Bayesian decision boundaries, respectively. Furthermore, several LVQ variants were proposed, which are derived from an explicit cost function (Sato and Yamada 1996, Seo and Obermayer 2002, Seo et al. 2003). Cost function based approaches are easily extended to a larger number of adaptive parameters. And methods of theoretical learning theory can be used to investigate risk bounds and convergence behavior. A mathematical analysis with respect to the cost function is performed in (Sato and Yamada 1998) and the authors of (Crammer et al. 2002) showed that LVQ aims at margin optimization and therefore good generalization ability can be expected. Further theoretical analysis of different LVQ variants and statistical physics investigations on simplified model situations can be found in (Ghosh et al. 2006, Biehl et al. 2007). Further extensions of the LVQ classification scheme includes the combination with other prototype-based learning schemes. For example the comprehension of the neighborhood cooperation known from Self-organizing Map (SOM) or Neural Gas (NG) into the learning process (Kohonen 2002, Hammer et al. 2005).

Particularly interesting for distance-based machine learning methods like mentioned before is the employed dissimilarity measure. A very common choice is the Euclidean distance, which is a special case of the Minkowski metric. Recently, also

divergences known from information theory were used as dissimilarity measure in vector quantization schemes (Mwebaze et al. 2011, Villmann and Haase 2011). In supervised settings where auxiliary information, such as labels, is available the adaptation of the distance by means of metric learning became popular. Some LVQ variants have been proposed, which aim at the optimization of the distance measure for a specific application (Bojer et al. 2001, Hammer and Villmann 2002, Schneider et al. 2009b, Schneider et al. 2009a). Also methods which aim at the optimization of the k-NN classification scheme have been developed using adaptive dissimilarities (Goldberger et al. 2004, Weinberger et al. 2006). Usually a big improvement of the classification performance can be observed when metric learning is incorporated in the algorithms. In the following section we will review some machine learning techniques used throughout the thesis, especially, existing metric adaptation schemes are presented.

2.2 Nearest prototype classification

We assume that the input data $\mathcal{X}$ consists of n examples $\{\boldsymbol{x}^i\}_{i=1}^n \in \mathbb{R}^N$ together with their corresponding labels $y^i \in \{1, \ldots, C\}$, where N denotes the dimension and C the number of classes or categories. A nearest prototype classifier is parameterized by a set of labeled prototype vectors $\boldsymbol{w}^j$, also called *codebook*, and a distance measure d. The protoytpes $\boldsymbol{w}^j$ are defined on the same feature space as the input data and they carry the label $c(\boldsymbol{w}^j)$ of the class they aim to represent. This implies the definition

$$\mathbf{W} = \{(\boldsymbol{w}^j, c(\boldsymbol{w}^j)) \in \mathbb{R}^N \times \{1, \ldots, C\}\}_{j=1}^{n_{\boldsymbol{w}}} \ , \tag{2.1}$$

where the number of prototypes $n_{\boldsymbol{w}} \geqslant C$, which means that at least one prototype per class is needed. A popular distance measure is the Euclidean distance, which is a special case of the general Minkowski metric

$$d^p(\boldsymbol{x}, \boldsymbol{w}) = \left(\sum_{i=1}^{N} |x_i - w_i|^p \right)^{\frac{1}{p}} \tag{2.2}$$

with $p = 2$. Examples of the equidistance lines using the Minkowski metric and different values for p are shown in Figure 2.1. The classification takes places by a winner-takes-all scheme, i.e. a new data point $\boldsymbol{x}$ is assigned to the class represented by the closest prototype:

$$\boldsymbol{x} \leftarrow c(\boldsymbol{w}^i), \text{with } \boldsymbol{w}^i = \arg\min_j d(\boldsymbol{x}, \boldsymbol{w}^j), \tag{2.3}$$

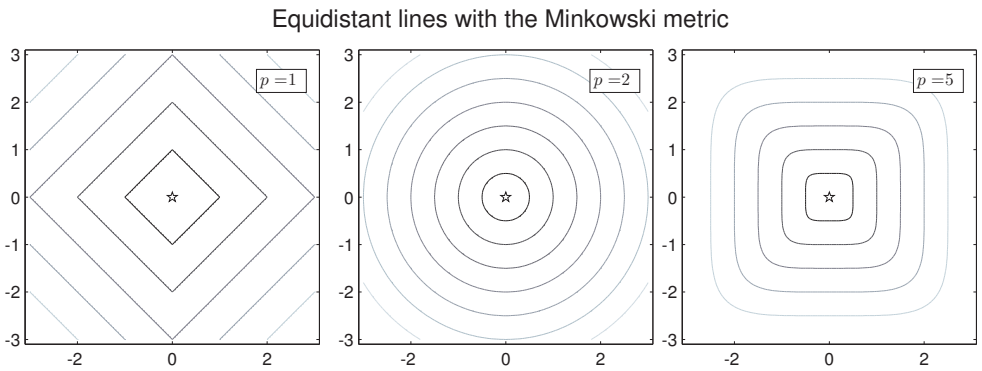

Figure 2.1: Visualization of the equidistance lines from the origin using the Minkowski metric with different values of p.

braking ties arbitrary. The set of protoytpes and the metric is partitioning the input data space. Each prototype $\boldsymbol{w}^i$ has a receptive field R^i, which is a region in the feature space where $\boldsymbol{w}^i$ is closer to the data than any other prototype:

$$R^i = \{\boldsymbol{x} \in \mathcal{X} |\ d(\boldsymbol{x}, \boldsymbol{w}^i) < d(\boldsymbol{x}, \boldsymbol{w}^j), \forall i \neq j\} \ . \tag{2.4}$$

Figure 2.2 shows two examples of nearest prototype classification on a three class problem using different distance measures. The Euclidean distance leads piecewise linear decision boundaries and receptive fields. For different values of p in the Minkowski metric more general decision boundaries can be realized.

The number of protoytpes is a hyper-parameter of the model and has to be optimized by means of a validation procedure. Too few prototypes may not represent the data structure sufficiently, which yields poor classification performance and too many prototypes may cause overfitting leading to poor generalization ability of the classifier. Many machine learning techniques have been proposed based on the nearest prototype classification scheme. Some of them used in the thesis will be addressed in the next sections.

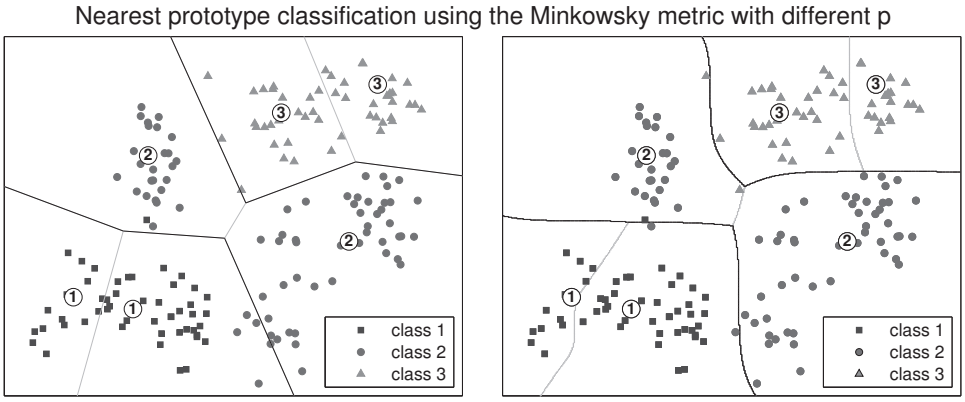

Figure 2.2: Visualization of the decision bounds of a nearest prototype classification scheme using different distances. The data is consisting of 3 classes and each class is represented by two prototypes. The Euclidean distance (left panel) shows piecewise linear boundaries where the gray lines denote the receptive fields of each prototype. In the right panel the Minkowski metric of order $p = 5$ is used.

Published as:

K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann and M. Biehl – *"Discriminative Visualization by Limited Rank Matrix Learning,"* Leipzig University, Machine Learning Reports (2:3), pp. 37–51, 2008.

K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann and M. Biehl – *"Limited Rank Matrix Learning Discriminative Dimension Reduction and Visualization,"* accepted for publication in Neural Networks 2011.

Chapter 3

Limited Rank Matrix LVQ

Projection makes it possible.

The impossible triangle. Shigeo Fukuda

Abstract

We present an extension of the Generalized Matrix Learning Vector Quantization algorithm. In the original scheme, adaptive square matrices of relevance factors parameterize a discriminative distance measure. We extend the scheme to matrices of limited rank corresponding to low-dimensional representations of the data. This allows to incorporate prior knowledge of the intrinsic dimension and to reduce the number of adaptive parameters efficiently. In particular, for very high dimensional data, the limitation of the rank can reduce computation time and memory requirements significantly. Furthermore, two- or three-dimensional representations constitute an efficient visualization method for labeled data sets. The identification of a suitable projection is not treated as a pre-processing step but as an integral part of the supervised training. Several real world data sets serve as illustration and demonstrate the usefulness of the suggested method.

3.1 Introduction

In (Schneider et al. 2009b, Schneider et al. 2009a) the concept of Generalized Matrix LVQ (GMLVQ) is introduced. It uses the quadratic form Eq. (**??**) as distance including a full matrix of relevances, which can account for correlations between different features. An adaptive self-affine transformation Ω (see Eq. (**??**)) of feature space identifies the coordinate system which is most suitable for the given classification task. The original formulation of GMLVQ employs symmetric squared matrices $\Omega \in \mathbb{R}^{N \times N}$ and is summarized in Algorithm **??**. In the simplest case, one matrix is taken to define a global distance measure. Extensions to class-wise or local matrices, attached to individual prototypes Eq. (**??**), are technically straightforward and allow for the parameterization of more complex decision boundaries.

In this chapter we present and discuss an important modification: the use of rectangular transformation matrices $\Omega \in \mathbb{R}^{M \times N}$ with $M \leqslant N$ (**?**, **?**). The corresponding relevance matrices Λ are of bounded rank M or, in other words, distances are evaluated in a space with reduced dimension, see Eq. (**??**). The motivation for considering this variation of GMLVQ is at least two-fold: (a) prior knowledge about the intrinsic dimension of the data can be incorporated efficiently and (b) the number of free parameters in the learning problem may be reduced significantly.

Although unrestricted GMLVQ displays a tendency to reduce the rank of the relevance matrices in the training process, the advantages of restricting the rank explicitly are obvious. In particular for nominally very high-dimensional data, e.g. in image analysis or bioinformatics, unrestricted relevance matrices become intractable. In addition, optimization results can be poor when the search is performed in an unnecessarily large parameter space. Furthermore, the exact control of the rank allows for pre-defining the dimension of the intrinsic representation and is, for instance, suitable for the discriminative visualization of labeled data sets. In contrast with many other schemes that consider dimension reduction as a pre-processing step, our method performs the training of prototypes and the identification of a suitable transformation simultaneously. Hence, both sub-tasks are guided by the ultimate goal of implementing the desired classification scheme.

Appropriate projections into two- or three-dimensional spaces can furthermore be used for efficient visualization of labeled data. Visualization enables to use the astonishing cognitive capabilities of humans for visual perception when extracting information from large data volumes. Structural characteristics can be captured almost instantly by humans, independent of the number of displayed points. Classical unsupervised dimension reduction techniques represent data points contained in a high dimensional data manifold by low dimensional counterparts in, for instance, two or three dimensions, while preserving as much information as possible. Since it

is not clear in advance which parts of the data are relevant to the user, this problem is inherently ill-posed: depending on the specific data domain and the situation at hand, different aspects can be in the focus of attention. Prior knowledge, in form of label information, can be used to formulate a well-defined objective in terms of the classification performance.

There exist a few classical dimensionality reduction tools which take class labels into account: e.g. Classical Fisher Linear Discriminant Analysis (LDA), the recently introduced local Fisher discriminant analysis (LFDA) (Sugiyama and Roweis 2007), Neighborhood Component Analysis (NCA) (Goldberger et al. 2004), as well as partial Least Squares regression (PLS). These methods can be extended to nonlinear projections by kernel methods (Ma et al. 2007, Baudat and Anouar 2000). Adaptive dissimilarity measures which modify the metric according to the given auxiliary information have been introduced e.g. in (Kaski et al. 2001, Peltonen et al. 2004, **?**, **?**, **?**).The resulting metric can be integrated into various techniques such as SOM, Multidimensional Scaling (MDS), or a recent information theoretic model for data visualization (Kaski et al. 2001, Peltonen et al. 2004, Venna et al. 2010). An ad hoc metric adaptation is used in (Geng et al. 2005) to extend Isomap (Tenenbaum et al. 2000) to class labels. Alternative approaches change the cost function of dimensionality reduction, for instance by using conditional probabilities, class-wise similarity matrices or introducing a covariance-based coloring matrix for the side information as proposed in (Iwata et al. 2007, Memisevic and Hinton 2005, Song et al. 2008). The detailed explanation of the most important supervised and unsupervised dimension reduction techniques is given in Part **??** of this thesis.

In the next section we describe the Limited Rank Matrix LVQ (LiRaM LVQ) as extension of the original GMLVQ formulation. Afterwards we apply the novel approach to a benchmark problem and study the influence of the dimension reduction on the classification performance. We also compare the limited rank version to the naive approach of taking the first components of the full rank GMLVQ. We show that reducing the rank after training not only requires more memory and CPU time, but also yields inferior classification performance compared to LiRaM LVQ. In Sec. **??** we present example applications of our algorithm in the visualization of labeled data. We also compare with visualizations obtained by LFDA and NCA. We conclude by summarizing our findings and providing an outlook on perspective investigations.

3.A Derivatives of GMLVQ and LiRaM LVQ

Here we show the derivatives of the GMLVQ costfunction E_{GMLVQ} for one presented training example $\boldsymbol{x}^i$, see Eq. (**??**), with respect to the prototypes $\boldsymbol{w}^L$ with $L \in \{J, K\}$ and the transformation matrix $\Omega \in \mathbb{R}^{M \times N}$. The derivative with respect to the prototypes can be formulated like following:

$$d_L^\Lambda = \sum_r^N \sum_m^M \sum_n^N (x_r^i - w_r^L)\Omega_{mr}\Omega_{mn}(x_n^i - w_n^L) \tag{3.1}$$

$$\frac{\partial E_{\mathrm{GMLVQ}}}{\partial \boldsymbol{w}^L} = \frac{\Phi(\mu^i)}{\partial \mu^i} \cdot \frac{\partial \mu^i}{\partial d_L^\Lambda} \cdot \frac{\partial d_L^\Lambda}{\partial \boldsymbol{w}^L} \tag{3.2}$$

$$\frac{\partial \mu^i}{\partial d_J^\Lambda} = \gamma^J = \frac{(d_J^\Lambda + d_K^\Lambda) - (d_J^\Lambda - d_K^\Lambda)}{(d_J^\Lambda + d_K^\Lambda)^2} = \frac{2d_K^\Lambda}{(d_J^\Lambda + d_K^\Lambda)^2} \tag{3.3}$$

$$\frac{\partial \mu^i}{\partial d_K^\Lambda} = \gamma^K = \frac{-(d_J^\Lambda + d_K^\Lambda) - (d_J^\Lambda - d_K^\Lambda)}{(d_J^\Lambda + d_K^\Lambda)^2} = \frac{-2d_J^\Lambda}{(d_J^\Lambda + d_K^\Lambda)^2} \tag{3.4}$$

$$\frac{\partial d_L^\Lambda}{\partial w_r^L} = -2 \cdot \sum_n^N \sum_m^M \Omega_{mr}\Omega_{mn}(x_n^i - w_n^L) = -2\left[\Omega^\top \Omega\right]_r (\boldsymbol{x}^i - \boldsymbol{w}^L) \tag{3.5}$$

$$\frac{\partial d_L^\Lambda}{\partial \boldsymbol{w}^L} = -2 \cdot \Omega^\top \Omega(\boldsymbol{x}^i - \boldsymbol{w}^L) \ . \tag{3.6}$$

The corresponding matrix update reads:

$$\frac{\partial E_{\mathrm{GMLVQ}}}{\partial \Omega_{mn}} = \frac{\Phi(\mu^i)}{\partial \mu^i} \cdot \frac{\partial \mu^i}{\partial \Omega_{mn}} \tag{3.7}$$

$$\begin{aligned}
\frac{\partial \mu^i}{\partial \Omega_{mn}} &= \frac{\left(\frac{\partial d_J^\Lambda}{\partial \Omega_{mn}} - \frac{\partial d_K^\Lambda}{\partial \Omega_{mn}}\right)(d_J^\Lambda + d_K^\Lambda) - (d_J^\Lambda - d_K^\Lambda)\left(\frac{\partial d_J^\Lambda}{\partial \Omega_{mn}} + \frac{\partial d_K^\Lambda}{\partial \Omega_{mn}}\right)}{(d_J^\Lambda + d_K^\Lambda)^2} \\
&= \frac{2d_K^\Lambda}{(d_J^\Lambda + d_K^\Lambda)^2} \cdot \frac{\partial d_J^\Lambda}{\partial \Omega_{mn}} + \frac{-2d_J^\Lambda}{(d_J^\Lambda + d_K^\Lambda)^2} \cdot \frac{\partial d_K^\Lambda}{\partial \Omega_{mn}} \\
&= \gamma^J \frac{\partial d_J^\Lambda}{\partial \Omega_{mn}} + \gamma^K \frac{\partial d_K^\Lambda}{\partial \Omega_{mn}}
\end{aligned} \tag{3.8}$$

$$\begin{aligned}
\frac{\partial d_L^\Lambda}{\partial \Omega_{mn}} &= 2\sum_r^N (x_n^i - w_n^L)\Omega_{mr}(x_r^i - w_r^L) \\
&= 2\left[\Omega(\boldsymbol{x}^i - \boldsymbol{w}^L)\right]_m \cdot (x_n^i - w_n^L) \ .
\end{aligned} \tag{3.9}$$

3.B Derivatives of Localized LiRaM LVQ

Now we describe the derivatives of the Localized LiRaM LVQ (LLiRaM LVQ) scheme for one presented training example $\boldsymbol{x}^i$ with respect to the prototypes $\boldsymbol{w}^L$, the transformation matrix $\Omega \in \mathbb{R}^{M\times N}$ and the localized dissimilarities denoted by $\Psi^L \in \mathbb{R}^{M\times M}$ with $L \in \{J, K\}$. We assume the quantities of the cost function Eq. (**??**) correspond to $d_J^\Lambda = d_J^{\Psi^J}(\boldsymbol{x}^i, \boldsymbol{w}^J)$ and $d_K^\Lambda = d_K^{\Psi^K}(\boldsymbol{x}^i, \boldsymbol{w}^K)$ using the distance measure defined in Eq. (**??**). The derivative with respect to the prototypes is given by:

$$d_L^{\Psi^L}(\boldsymbol{x}^i, \boldsymbol{w}^L) = \sum_j^N \sum_k^M \sum_l^M \sum_m^M \sum_n^N (x_j^i - w_j^L)\Omega_{kj}\Psi_{lk}^L \Psi_{lm}^L \Omega_{mn}(x_n^i - w_n^L) \tag{3.10}$$

$$\frac{\partial E_{\mathrm{GMLVQ}}}{\partial \boldsymbol{w}^L} = \frac{\Phi(\mu^i)}{\partial \mu^i} \cdot \frac{\partial \mu^i}{\partial d_L^{\Psi^L}} \cdot \frac{\partial d_L^{\Psi^L}}{\partial \boldsymbol{w}^L} \tag{3.11}$$

$$\frac{\partial \mu^i}{\partial d_J^{\Psi^J}} = \gamma_\Psi^J = \frac{2 d_K^{\Psi^K}}{(d_J^{\Psi^J} + d_K^{\Psi^K})^2} \tag{3.12}$$

$$\frac{\partial \mu^i}{\partial d_K^{\Psi^K}} = \gamma_\Psi^K = \frac{-2 d_J^{\Psi^J}}{(d_J^{\Psi^J} + d_K^{\Psi^K})^2} \tag{3.13}$$

$$\frac{\partial d_L^{\Psi^L}}{\partial w_r^L} = -2 \sum_k^M \sum_l^M \sum_m^M \sum_n^N \Omega_{kr}\Psi_{lk}^L \Psi_{lm}^L \Omega_{mn}(x_n^i - w_n^L)$$

$$\frac{\partial d_L^{\Psi^L}}{\partial \boldsymbol{w}^L} = -2\Omega^\top \Psi^{L\top} \Psi^L \Omega(\boldsymbol{x}^i - \boldsymbol{w}^L) \tag{3.14}$$

The derivative with respect to the matrices is given by:

$$\frac{\partial E_{\mathrm{GMLVQ}}}{\partial \Omega} = \frac{\Phi(\mu^i)}{\partial \mu^i} \cdot \frac{\partial \mu^i}{\partial \Omega} = \Phi' \cdot \left(\gamma_\Psi^J \cdot \frac{\partial d_J^{\Psi^J}}{\partial \Omega} + \gamma_\Psi^K \cdot \frac{\partial d_K^{\Psi^K}}{\partial \Omega} \right) \tag{3.15}$$

$$\frac{\partial d_L^{\Psi^L}}{\partial \Omega_{mn}} = 2 \sum_j^N \sum_k^M \sum_l^M (x_n - w_n^L)\Psi_{kl}^L \Psi_{km}^L \Omega_{lj}(x_j - w_j^L) \tag{3.16}$$

$$\frac{\partial d_L^{\Psi^L}}{\partial \Omega} = 2 \cdot \Psi^{L\top} \Psi^L \Omega(\boldsymbol{x} - \boldsymbol{w}^L)(\boldsymbol{x} - \boldsymbol{w}^L)^\top \tag{3.17}$$

$$\frac{\partial E_{\mathrm{GMLVQ}}}{\partial \Psi^L} = \frac{\Phi(\mu^i)}{\partial \mu^i} \cdot \frac{\partial \mu^i}{\partial \Psi^L} = \Phi' \cdot \gamma_\Psi^L \cdot \frac{\partial d_L^{\Psi^L}}{\partial \Psi^L} \tag{3.18}$$

$$\frac{\partial d_L^{\Psi^L}}{\partial \Psi_{mn}^L} = 2 \sum_j^N \sum_k^N \sum_l^M (x_k - w_k^L)\Omega_{nk}(x_j - w_j^L)\Psi_{ml}^L \Omega_{lj} \tag{3.19}$$

$$\frac{\partial d_L^{\Psi^L}}{\partial \Psi^L} = 2 \cdot \Psi^L \left(\Omega(\boldsymbol{x}^i - \boldsymbol{w}^L)(\boldsymbol{x}^i - \boldsymbol{w}^L)^\top \right) \Omega^\top \tag{3.20}$$

Bibliography

Baudat, G. and Anouar, F.: 2000, Generalized discriminant analysis using a kernel approach, *Neural Computation* **12**(10), 2385–2404.

Biehl, M., Ghosh, A. and Hammer, B.: 2007, Dynamics and generalization ability of LVQ algorithms, *Journal of Machine Learning Research* **8**, 323–360.

Bishop, C. M.: 1995, *Neural networks for pattern recognition*, Oxford University Press, USA.

Bishop, C. M.: 2006, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Bojer, T., Hammer, B., Schunk, D. and von Toschanowitz, K. T.: 2001, Relevance determination in learning vector quantization, *in* M. Verleysen (ed.), *Proc. of the 9th European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium, pp. 271–276.

Chopra, S., Hadsell, R. and Lecun, Y.: 2005, Learning a similarity metric discriminatively, with application to face verification, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Press, San Diego, CA, pp. 539–546.

Cover, T. M. and Hart, P. E.: 1967, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* **13**(1), 21–27.

Crammer, K., Gilad-bachrach, R., Navot, A. and Tishby, N.: 2002, Margin analysis of the LVQ algorithm, *Advances in Neural Information Processing Systems (NIPS) 2002*, Vol. 15, MIT press, Cambridge, MA, USA, pp. 462–469.

Duda, R. O., Hart, P. E. and Stork, D. G.: 2000, *Pattern Classification*, Wiley-Interscience Publication.

Frome, A., Sha, F., Singer, Y. and Malik, J.: 2007, Learning globally-consistent local distance functions for shape-based image retrieval and classification, *Proc. of the 11th IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil.

Geng, X., Zhan, D.-C. and Zhou, Z.-H.: 2005, Supervised nonlinear dimensionality reduction for visualization and classification, *IEEE Transactions on Systems, Man, and Cybernetics Part B* **35**(6), 1098–1107.

Ghosh, A., Biehl, M. and Hammer, B.: 2006, Performance analysis of lvq algorithms: A statistical physics approach, *Neural Networks* **19**(6-7), 817–829.

Goldberger, J., Roweis, S., Hinton, G. and Salakhutdinov, R.: 2004, Neighborhood Component Analysis, *Advances in Neural Information Processing Systems (NIPS)*.

Hammer, B., Strickert, M. and Villmann, T.: 2005, Supervised neural gas with general similarity measure, *Neural Processing Letters* **21**(1), 21–44.

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

Han, J. and Kamber, M.: 2005, *Data Mining: Concepts and Techniques*, Morgen Kaufmann Publishers Inc.

Iwata, T., Saito, K., Ueda, N., Stromsten, S., Griffiths, T. L. and Tenenbaum, J. B.: 2007, Parametric embedding for class visualization, *Neural Computation* **19**(9), 2536–2556.

Kaski, S., Sinkkonen, J. and Peltonen, J.: 2001, Bankruptcy analysis with self-organizing maps in learning metrics, *IEEE Transactions on Neural Networks* **12**, 936–947.

Kohonen, T.: 1986, Learning vector quantization for pattern recognition, *Technical Report TKK-F-A601*, Helsinki University of Technology, Espoo, Finland.

Kohonen, T. K.: 2002, *The handbook of brain theory and neural networks*, MIT press, Cambridge, MA, chapter Learning vector quantization, pp. 631–635.

Kohonen, T., Schroeder, M. R. and Huang, T. S. (eds): 2001, *Self-Organizing Maps*, 3rd edn, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Lee, J. and Verleysen, M.: 2007, *Nonlinear dimensionality reduction*, 1st edn, Springer.

Lyman, P. and Varian, H. R.: 2003, *How Much Information*. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on March 23, 2020.

Ma, B., Qu, H. and Wong, H.: 2007, Kernel clustering-based discriminant analysis, *Pattern Recognition* **40**(1), 324–327.

Martinetz, T. and Schulten, K.: 1991, A "neural-gas" network learns topologies, *Artificial Neural Networks* **I**, 397–402.

Memisevic, R. and Hinton, G.: 2005, Multiple relational embedding, *in* L. K. Saul, Y. Weiss and L. Bottou (eds), *Advances in Neural Information Processing Systems (NIPS) 17*, MIT Press, Cambridge, MA, pp. 913–920.

Mitchell, T. M.: 1997, *Machine Learning*, McGraw-Hill Series in Computer Science, WCB/McGraw-Hill, Boston, MA.

Mwebaze, E., Schneider, P., Schleif, F.-M., Aduwo, J., Quinn, J., Haase, S., Villmann, T. and Biehl, M.: 2011, Divergence based classification in learning vector quantization, *Neurocomputing* **74**(9), 1429–1435.

Peltonen, J., Klami, A. and Kaski, S.: 2004, Improved learning of Riemannian metrics for exploratory analysis, *Neural Networks* **17**, 1087–1100.

Ripley, B. D.: 1996, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

Sato, A. S. and Yamada, K.: 1996, Generalized learning vector quantization, *in* M. C. M. D. S. Touretzky and M. E. Hasselmo (eds), *Advances in Neural Information Processing Systems (NIPS)*, Vol. 8, MIT Press, Cambridge, MA, USA, pp. 423–429.

Sato, A. and Yamada, K.: 1998, An analysis of convergence in generalized LVQ, *in* L. Niklasson, M. Bodén and T. Ziemke (eds), *Proc. of the 8th International Conference on Artificial Neural Networks*, Vol. 1, Springer, London, pp. 170–176.

Schneider, P., Biehl, M. and Hammer, B.: 2009a, Adaptive relevance matrices in learning vector quantization, *Neural Computation* **21**(12), 3532–3561.

Schneider, P., Biehl, M. and Hammer, B.: 2009b, Distance learning in discriminative vector quantization, *Neural Computation* **21**(10), 2942–2969.

Seo, S., Bode, M. and Obermayer, K.: 2003, Soft nearest prototype classification, *IEEE Transactions on Neural Networks* **14**, 390–398.

Seo, S. and Obermayer, K.: 2002, Soft learning vector quantization, *Neural Computation* **15**, 1589–1604.

Song, L., Smola, A. J., Borgwardt, K. M. and Gretton, A.: 2008, Colored maximum variance unfolding, *in* J. C. Platt, D. Koller, Y. Singer and S. T. Roweis (eds), *Advances in Neural Information Processing Systems (NIPS)*, MIT Press, Cambridge, MA.

Sugiyama, M. and Roweis, S.: 2007, Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis, *Journal of Machine Learning Research* **8**, 1027–1061.

Sutton, R. S. and Barto, A. G.: 1998, Reinforcement learning: An introduction, *IEEE Transactions on Neural Networks* **9**(5), 1054–1054.

Tenenbaum, J. B., Silva, V. d. and Langford, J. C.: 2000, A global geometric framework for nonlinear dimensionality reduction, *Science* **290**(5500), 2319–2323.

Venna, J., Peltonen, J., Nybo, K., Aidos, H. and Kaski, S.: 2010, Information retrieval perspective to nonlinear dimensionality reduction for data visualization, *Journal of Machine Learning Research* **11**, 451–490.

Villmann, T. and Haase, S.: 2011, Divergence based vector quantization using Fréchet-derivatives, *Neural Computation* **23**(5), 1343–1392. accepted for publication.

Weinberger, K. Q., Blitzer, J. and Saul, L. K.: 2006, Distance metric learning for large margin nearest neighbor classification, *Advances in Neural Information Processing Systems (NIPS)* **18**, 1473–1480.

Xing, E. P., Ng, A. Y., Jordan, M. I. and Russell, S.: 2002, Distance metric learning, with application to clustering with side-information, *Advances in Neural Information Processing Systems (NIPS) 15*, MIT Press, Cambridge, pp. 505–512.

Samenvatting

Deze thesis presenteert een aantal extensies van het Generalized LVQ (GLVQ) algoritme gebaseerd op het concept van adaptive similarity measures. Deze metric learning kan worden gebruikt in een grote verscheidenheid aan applicaties, waaronder Content Based Image Retrieval (CBIR), supervised dimension reduction en advanced texture learning bij image analysis, om een paar te noemen. Het gedetailleerde onderzoek naar dimensionality reduction komt uitgebreid aan bod in de tweede helft van de thesis. Dit omvat onderzoek naar generalized explicit dimension reduction mappings voor unsupervised en supervised dimension reduction. Een nieuwe techniek voor efficient unsupervised non-linear dimension reduction wordt voorgesteld die de concepten van fast online learning en optimalisatie van divergenties combineert. Tot slot worden drie op divergentie gebaseerde algoritmes gegeneraliseerd en onderzocht op het gebruik van willekeurige divergenties.

In Chapter 2 wordt de benodigde achtergrond voor adaptive metric learning en prototype-based classification gegeven. Vervolgens wordt LiRaM LVQ geïntroduceerd in Chapter 3, een algoritme gericht op efficiënte optimalisatie van classificatie, met name bij zeer hoog-dimensionale datasets. Door de rank van de adaptieve matrix, een onderdeel van de gebruikte afstand, te begrenzen, kan het aantal vrije parameters expliciet worden gereguleerd. We laten zien dat naast computationele efficiëntie, het begrenzen van de rank een hogere kwaliteit laat zien vergeleken met alternatieve methoden gebaseerd op de decompositie van eigenwaarden na training, met name wanneer de target-dimensie lager is dan de intrinsieke dimensionaliteit van de dataset. Daarnaast staat dit concept discriminant linear dimension reduction toe, gericht op het behoud van de classification accuracy bij lagere dimensionaliteit. Door de distance measure in globale en lokale of klasse-specifieke matrices te ontbinden kunnen complexere decision boundaries worden bewerkstelligd in de visualisatie. Dit combineert linear dimension reduction met localized

similarity measures in laag-dimensionale ruimte, wat resulteert in non-linear decision boundaries van de receptieve velden. De dimension reduction met LiRaM LVQ toont vergelijkbare of betere resultaten dan alternatieve state-of-the-art technieken. Bovendien is de methode ook computationeel gezien efficiënt. In contrast met andere high-quality technieken vereist het niet de berekening van pair-wise affinities van de datapunten, maar slechts hun afstand tot het (kleine) aantal prototypes, wat over het algemeen minder berekeningen vereist. Verschillende experimenten op real-world datasets worden gepresenteerd en bevestigen onze claims.

Chapter **??** presenteert een voorbeeldapplicatie van LiRaM LVQ in de context van CBIR. Voor veel medische applicaties is de hoeveelheid data enorm gestegen in de afgelopen jaren. Daarom zijn computer aided diagnosis systems, die geautomatiseerd databases doorzoeken om potentieel interessante data voor een bepaalde taak voor te selecteren, zeer wenselijk. Dit werk behandelt CBIR in de context van dermatologie. In een samenwerkingsverband heeft de afdeling Dermatologie van het Universitair Medisch Centrum Groningen een database met afbeeldingen van verschillende typen huidletsels beschikbaar gesteld. Het doel is om gegeven een afbeelding een bepaald aantal vergelijkbare afbeeldingen op te leveren. Met het gebruik van adaptive metrics waren we in staat om het aandeel correct opgeleverde afbeeldingen aanzienlijk te verhogen, voor willekeurige color spaces. We vergelijken twee technieken voor distance learning: de Large Margin Nearest Neighbor (LMNN) en de LiRaM LVQ methode. Het is opmerkelijk dat LiRaM LVQ hierbij beter presteerde dan LMNN met typische instellingen. Door de complexiteit en het tijdsverbruik van LMNN te laten toenemen konden vergelijkbare resultaten worden behaald.

In Chapter **??** introduceren we een complexe variant op GLVQ voor texture classification, genaamd Color Image Analysis LVQ (CIA LVQ). Deze flexibele methode combineert discriminative local linear projections in het Fourierdomein met linear filtering, e.g. met Gabor filters. Lineaire filteroperaties zijn vaak gedefinieerd op intensiteitswaarden. In het verleden zijn enkele heuristieke methoden voor filteroperaties op kleurenafbeeldingen voorgesteld die de response- of energiewaarden van kleurkanalen op een betekenisvolle manier combineren. Onze methode is van verschillende aard omdat het gebaseerd is op een automatisch lerende procedure gestuurd door supervised training. Hiervoor wordt a priori een Gabor filterbank verzameld met gewichten en oriëntaties passend bij de texture recognition taak. We nemen willekeurige segmenten van kleurenafbeeldingen van bekende klassen en voor elk van deze transformeren we de kleurkanalen afzonderlijk naar het Fourierdomein. De transformaties van kleurwaarden naar intensiteitswaarden worden geleerd door het CIA LVQ systeem om de filterresponses op deze getransformeerde segmenten beter te kunnen onderscheiden. In het bijzonder bij textures die zich

in de natuur voordoen zoals schors en voedselstructuren presteert de voorgestelde techniek beter dan alternatieve methoden waaronder het naïeve gebruik van een RGB naar grijswaarden transformatie, hetgeen in de praktijk vaak gebruikt wordt. Bovendien toont CIA LVQ uitstekende eigenschappen met betrekking tot evaluatie-afbeeldingen die niet eerder aan het systeem getoond zijn.

Deel **??** van deze thesis behandelt verschillende aspecten die betrekking hebben op dimension reduction. In Chapter **??** wordt een nieuwe algemene opvatting voorgesteld die de aanpassing van verschillende methoden voor dimension reduction voor explicit mappings vergemakkelijkt. In plaats van een impliciete optimalisatie van de posities van laag-dimensionale datapunten predefiniëren we de vorm van een mapping-functie f_W geparametriseerd door W, en optimaliseren we de parameters ten behoeve van een specifiek doel. Dit heeft het voordeel dat de training uitgevoerd kan worden op slechts een klein deel van de data en een rechtstreekse out-of-sample extensie voor alle datapunten is direct beschikbaar. Daarnaast wordt een theoretisch onderzoek naar de generalisatie-eigenschappen van dimension reduction mogelijk. We demonstreren het concept van dimension reduction mappings gebaseerd op de t-distributed SNE (t-SNE) kostenfunctie en verschillende alternatieven voor de mapping-functie f_W. Dit omvat zowel unsupervised linear en non-linear mappings gebaseerd op local PCA alsook supervised mappings die gebruikmaken van discriminative local linear projections. We vergelijken de methode met verschillende state-of-the-art technieken, tonen de uitstekende generalisatie-eigenschappen voor verschillende datasets en behandelen tenslotte het theoretische onderzoek naar dimension reduction mappings. In alle gevallen geeft onze methode vergelijkbare of zelfs betere resultaten.

Chapter **??** onderzoekt supervised dimension reduction gebaseerd op adaptieve afstanden en local linear projections verkregen door GMLVQ and LiRaM LVQ. Dit maakt de integratie van dimension reduction in de optimalisatieprocedure gericht op discriminative visualizations mogelijk. We laten zien met behulp van verschillende voorbeelden dat bestaande methoden voor dimension reduction uitgebreid kunnen worden naar een supervised setting gebruikmakend van de geleerde metrics en discriminative transformations van LVQ.

In Chapter **??** wordt een methode voor unsupervised dimension reduction voorgesteld, die fast sequential online learning combineert met direct divergence optimization zoals gebruikt in Stochastic Neighbor Embedding (SNE) en t-SNE. Deze techniek heet Self Organized Neighbor Embedding (SONE) en vertoont enkele interessante eigenschappen: in zijn oorspronkelijke formulering is SONE gebaseerd op een structuurhypothese die de gebruiker in staat stelt om het uiterlijk van de uiteindelijke embedding en de computationele inspanningen aan te passen. Veel technieken voor dimension reduction vereisen de berekening van alle pair-wise affinities

van laag-dimensionale afbeeldingsvectoren in een optimalisatiestap. Dit heeft een computationele complexiteit van $\mathcal{O}(n^2)$ tot gevolg, waarbij n staat voor het aantal datapunten. SONE berekent de afstanden naar één sampling vector uit de gegeven hypothese in iedere iteratie voor de aanpassing van alle punten. Daarmee is de computationele complexiteit lineair afhankelijk van het aantal punten en sampling vectors gegeven door de hypothese. Ondanks het feit dat de methode minder complex is dan SNE en t-SNE, toont het een vergelijkbare kwaliteit zoals gedemonstreerd wordt aan de hand van een aantal voorbeelden.

Chapter **??** behandelt een systematische aanpak voor de wiskundige behandeling van divergence based dimension reduction, zoals SNE, t-SNE en SONE, ten behoeve van de uitwisseling van hun respectievelijke modules. Naast de onafhankelijke behandeling van de verdeling in laag-dimensionale ruimte, e.g. het gebruik van een Gaussian voor SNE en een t-verdeling in t-SNE, concentreren we ons op de divergentie waarmee het verschil tussen verdelingen in de originele en de embeddingruimte gemeten wordt. Daarom bekijken we de divergentie-families en hun eigenschappen. We stellen een algemeen framework voor gebaseerd op het concept van Fréchet-afgeleiden en leiden de expliciete learning rules voor een breed scala aan divergenties af. In de experimenten concentreren we ons op de evaluatie van de Gamma-divergentie voor t-SNE en SONE in een aantal real-world datasets. We zagen dat de Gamma-divergentie de kwaliteit van de embeddings voor small neighborhoods verbetert vergeleken met de originele formulering met behulp van Kullback-Leibler.

Index