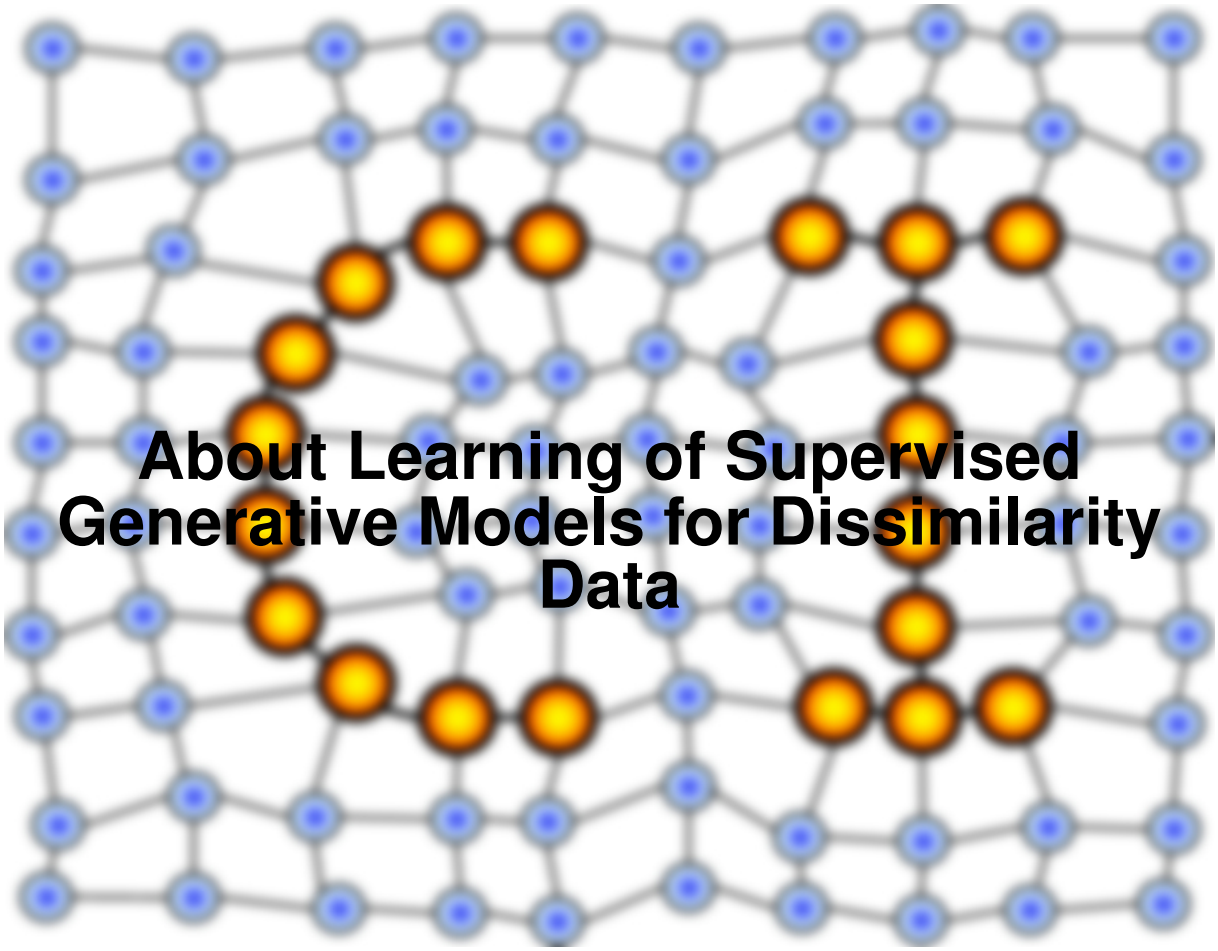
MACHINE LEARNING REPORTS



About Learning of Supervised Generative Models for Dissimilarity Data

Report 06/2013
Submitted: 01.11.2013
Published: 07.11.2013

David Nebel[1], Barbara Hammer[2], and Thomas Villmann[1]
(1)University of Applied Sciences Mittweida, Fac. of Mathematics/Natural and Computer Sciences, Computational Intelligence Group, Mittweida, Technikumplatz 17, 09648 Mittweida, Germany
(2)CITEC Centre of Excellence, Bielefeld University, Theoretical Computer Science Group, Universitätsstraße 21-23,D-33594 Bielefeld, Germany

Abstract

Exemplar based techniques such as affinity propagation represent data in terms of typical exemplars. This has two benefits: (i) the resulting models are directly interpretable by humans since representative exemplars can be inspected in the same way as data points, (ii) the model can be applied to any dissimilarity measure including non-Euclidean or non-metric settings. Most exemplar based techniques have been proposed in the unsupervised setting only, such that their performance in supervised learning tasks can be weak depending on the given data. We address the problem of learning exemplar-based models for general dissimilarity data in a discriminative framework in this contribution. For this purpose, we extend a generative model to an exemplar based scenario using a generalized EM framework for its optimization. The resulting classifiers represent data in terms of sparse models thereby reaching state-of-the art results in benchmarks.

About Learning of Supervised Generative Models for Dissimilarity Data

David Nebel[1], Barbara Hammer[2], and Thomas Villmann[1]

[1]University of Applied Sciences Mittweida,
Fac. of Mathematics/Natural and Computer Sciences,
Computational Intelligence Group

[2]CITEC Centre of Excellence,
Bielefeld University
Theoretical Computer Science Group

Abstract

Exemplar based techniques such as affinity propagation [13] represent data in terms of typical exemplars. This has two benefits: (i) the resulting models are directly interpretable by humans since representative exemplars can be inspected in the same way as data points, (ii) the model can be applied to any dissimilarity measure including non-Euclidean or non-metric settings. Most exemplar based techniques have been proposed in the unsupervised setting only, such that their performance in supervised learning tasks can be weak depending on the given data. We address the problem of learning exemplar-based models for general dissimilarity data in a discriminative framework in this contribution. For this purpose, we extend a generative model proposed in [36] to an exemplar based scenario using a generalized EM framework for its optimization. The resulting classifiers represent data in terms of sparse models thereby reaching state-of-the art results in benchmarks.

1 Introduction

Machine learning has revolutionized the possibility to deal with large electronic data sets. Nevertheless, rapid technological developments continue to pose challenges to the field, such as the big data challenge, or the problem of complex non-vectorial structures, which are increasingly common. Examples of the latter include biological sequences, mass spectra, or metabolic networks, where complex alignment techniques, background information, or general information theoretical principles, for example, drive the comparison of data points [29, 23, 18]. These data cannot be embedded in Euclidean space, and they often do not even fulfill the properties of a metric. Further, for dissimilarities such as used in social network analysis even pseudo-Euclidean embedding such as proposed in [28] might fail due to asymmetric dissimilarities. These developments have caused the need for non-vectorial machine learning tools such as e.g. structure kernels, recursive models, relational models, or quotient embeddings [19, 16, 10].

Since learning tasks become more and more complex, the specific objectives are often not clear a priori. This challenge requires increasingly interactive systems which allow humans to shape the problems according to human insights and expert knowledge at hand [37]. A vital property of machine learning models in this context is their interpretability by means of semantically meaningful interfaces [31]. While interpretable models enable the change of their functionality by experts, popular black box techniques such as the SVM often only provide an excellent classification performance, but no insight on why this is the case. It is hardly possible to visualize its decisions to domain experts in such a way that relevant information can be inferred based thereon by a human observer. The same argument, although to a lesser degree, is valid for alternatives such as the relevance vector machine or sparse models which typically still rely on complex nonlinear combinations [38].

The demand of interpretability can be met with quite diverse technologies, such as sparsity, relevance learning, or enhancement by visualization [3]. One example is offered by dissimilarity based learning: this relies on comparisons of given data to known labeled data points. Hence it is usually easy to interpret the decision: a small number of closest neighbors accounts for the observed classification. These

neighbors can directly be inspected by experts in the same way as data points. Because of this fact, similarity based techniques enjoy a large popularity in application domains such as biomedical applicatons, whereby the methods range from simple k-nearest neighbor classifiers and learning vector quantization up to advanced techniques such as affinity propagation which represents a clustering in terms of typical exemplars [21, 13, 1].

Dissimilarity based techniques can be distinguished according to different criteria: (i) The number of data points used to represent the classifier ranging from dense models such as k-nearest neighbor to sparse representations such as prototype based methods. To arrive at easily interpretable models, a sparse representation in terms of few data points is necessary. (ii) The degree of supervision ranging from clustering techniques such as affinity propagation to supervised learning. Here we are interested in classification techniques, i.e. supervised learning. (iii) The complexity of the dissimilarity measure the methods can deal with ranging from vectorial techniques restricted to Euclidean spaces, adaptive techniques which learn the underlying metrics, up to tools which can deal with arbitrary similarities or dissimilarities [27, 33, 30]. Typically, Euclidean techniques are well suited for simple classification scenarios, but they fail if complex structures are encountered.

Learning vector quantization (LVQ) constitutes one of the few methods to infer a sparse representation in terms of prototypes from a given data set in a supervised way [21], such that it offers a good starting point as an intuitive classification technique which decisions can directly be inspected by humans. Albeit original LVQ has been introduced on somewhat heuristic grounds [21], recent developments in this context provide a solid mathematical derivation of its generalization ability and learning dynamics: LVQ classifiers can be substantiated by large margin generalization bounds [9, 33]; the dynamics of LVQ type algorithms can be derived from explicit cost functions [33, 36, 35]. Interestingly, already the dynamics of classical LVQ provably leads to very good generalization ability in typical model situation as investigated in the framework of online learning [4].

A severe drawback of LVQ type classifiers is their dependency on the Euclidean metric. This problem can partially be avoided by appropriate metric learning, see e.g. [33], or by kernel variants, see e.g. [30], which turn LVQ classifiers into state-of-the-art techniques e.g. in connection to humanoid robotics or computer vision

[12, 20]. However, if data are inherently non-Euclidean, these techniques cannot be applied. Recently, an extension of LVQ type learning by means of an implicit embedding in pseudo-euclidean space has been proposed [17]. Albeit yielding state-of-the-art results, this technique faces two problems: it cannot be used for asymmetric dissimilarities where no pseudo-euclidean embedding exists; by representing prototypes in terms of distributed coefficient vectors, interpretability, one of LVQ's main benefits, is lost. In this contribution, we address the problem by taking an alternative point of view: we address LVQ algorithms derived from generative statistical models, and we extend these techniques to exemplar based learners suitable for arbitrary dissimilarities, similar to the unsupervised setting as proposed in [13].

Unlike unsupervised generative models for density estimation such as classical mixtures of Gaussians, extensions towards more general data structures, or extensions towards richer functionality [5, 6, 14] supervised generative models face two partially contradictory objectives: parameterized generative models describe the data distribution using few, preferably interpretable model parameters; at the same time, the discriminative power of the model depends on its ability to represent a suitably nonlinear and possibly complex decision boundary. Often, these two objectives are taken into account in separate steps only, e.g. training generative models individually on every given class, or incorporating supervised label information as side information for the adaptation of some model parameters only, such as e.g. metric parameters [5, 15].

One interesting approach which transfers classical generative mixture models to a discriminative setting by means of an explicit discriminative cost function has been proposed in [36]: prototypes are equipped with labels, which determine the classification of a given data point. Training takes place by means of a likelihood ratio maximization. In the limit of small bandwidths, a learning rule similar to a classical heuristic LVQ variant results; however, in this limit, the performance of the classifier is often worse as compared to the full probabilistic model [4]. Here, we extend this approach towards general dissimilarity data by taking an exemplar based point of view. A training algorithm can be derived thereof by means of a generalized EM scheme, yielding a state-of-the-art classifier with superior performance as opposed to unsupervised exemplar-based approaches [13].

2 Supervised generative models in Euclidean space

Assume the data space $\mathbb{X}$ is a standard Euclidean vector space. Assume data points $x_1, \ldots, x_N$ together with labels $y_1, \ldots, y_N \in \{1, \ldots, C\}$ are given. Robust soft learning vector quantization (RSLVQ) as proposed in [36] represents data in terms of a mixture model with model parameters $\Theta = \{\theta_1, \ldots, \theta_M\} \in \mathbb{X}$ which induce the probability

$$p(x_i|\Theta) = \sum_{j=1}^{M} p(\theta_j) p(x_i|\theta_j)$$

where, typically, the prior probabilities $p(\theta_j)$ are chosen as constant and $p(x_i|\theta_j)$ is given by a Gaussian distribution in Euclidean space. In [36], the correlation matrix is taken as unit matrix, a generalization towards a general form has been proposed in [34]. For such a mixture of Gaussian, the model parameters θ_i take the role of prototypes and they can serve as an interface towards an interpretation of the model.

For the supervised setting, every prototype is equipped with a class label $c_i \in \{1, \ldots, C\}$, yielding the joint distribution

$$p(x_i, y_i|\Theta) = \sum_{j=1}^{M} \delta_{c_j}^{y_i} \cdot p(\theta_j) p(x_i|\theta_j)$$

with Kronecker δ. Marginalization gives $p(y_i|\Theta) = \sum_{j=1}^{M} \delta_{c_j}^{y_i} \cdot p(\theta_j)$. Thus

$$p(y_i|x_i, \Theta) = \frac{p(x_i, y_i|\Theta)}{p(x_i|\Theta)} = \frac{\sum_{j=1}^{M} \delta_{c_j}^{y_i} \cdot p(\theta_j) p(x_i|\theta_j)}{\sum_{j=1}^{M} p(\theta_j) p(x_i|\theta_j)}.$$

Trainings takes place by an optimization of the log likelihood ratio, assuming i.i.d. data:

$$K(\mathbb{X}, \Theta) = \sum_{i=1}^{N} \ln p(y_i|x_i, \Theta)$$

In Euclidean space, a standard gradient technique can be used for optimization.

3 Supervised generative models for dissimilarity data

Assume $\mathbb{X}$ is a possibly non-Euclidean measurable space equipped with a probability distribution p. The cost function of RSLVQ can be transferred to this setting provided a suitable probability measure $p(x_i|\theta_j)$ is given. There remain, however, two problems:

- In the absence of an underlying vector space, how to define a suitable probability $p(x_i|\theta_j)$ and a suitable space of parameters θ_j for this model, which still yields interpretable representations?
- How to train the model? Optimization by means of gradient techniques is usually impossible unless $\mathbb{X}$ is embedded in a real-vector space.

Here, we are particularly interested in settings where data are characterized by pairwise dissimilarities only, as discussed e.g. in [28, 7]. This corresponds to data x_i being represented by terms $d(x_i, x_j)$ for all pairs of points where $d(x_i, x_j) \geq 0$ is some reasonable measure for the dissimilarity of two objects. Thereby, we do not assume Euclideanity of d, in which case kernel techniques can be used. Nor do we assume symmetry, which would enable the embedding into pseudo-Euclidean spaces [28].

Extension of the objective to exemplars

Since the underlying space $\mathbb{X}$ is unknown, we take an exemplar based approach similar to [13]: model parameters θ_j are restricted to data points $\{x_1, \ldots, x_N\}$, such that the dissimilarity $d(x_i, \theta_j)$ is always well defined. If d is measurable and non negative, we can define a probability in analogy to Gaussians as

$$p(x_i|\theta_j) = \frac{1}{K_j} \cdot \exp(-d(x_i, \theta_j)/\sigma^2)$$

with normalizing constant $K_j = \int_{\mathbb{X}} \exp(-d(x_i, \theta_j)/\sigma^2) d_p(x)$. Thereby, K_j is usually not known and it has to be estimated from data; for simplicity, isotropy is

often assumed, i.e. K_j is constant. Note that this choice preserves interpretability of the model parameters θ_j provided d constitutes a reasonable dissimilarity measure, since decisions are based on the dissimilarity compared to the closest exemplar.

Optimization

For optimization of the model parameters, instead of gradient techniques as used in the vectorial case, a generalized EM strategy is possible, as we will show in the following. The objective

$$K(\mathbb{X}, \Theta) = \sum_{i=1}^{N} \ln p(y_i|x_i, \Theta) = \sum_{i=1}^{N} \ln \sum_{j=1}^{M} \delta_{c_j}^{y_i} \cdot \frac{p(\theta_j)p(x_i|\theta_j)}{\sum_{j=1}^{M} p(\theta_j)p(x_i|\theta_j)}$$

decomposes into a sum of nonnegative functions

$$g(x_i, y_i, \theta_j) = \delta_{c_j}^{y_i} \cdot \frac{p(\theta_j)p(x_i|\theta_j)}{\sum_{j=1}^{M} p(\theta_j)p(x_i|\theta_j)}$$

Set

$$p(\theta_j|x_i, y_i) = \frac{\delta_{c_j}^{y_i} \cdot p(\theta_j)p(x_i|\theta_j)}{\sum_{j=1}^{M} \delta_{c_j}^{y_i} \cdot p(\theta_j)p(x_i|\theta_j)} = \frac{g(x_i, y_i, \theta_j)}{\sum_{j=1}^{M} g(x_i, y_i, \theta_j)}$$

as probability of mode number j. Assume that $\gamma(\theta_j|x_i, y_i)$ is any probability distribution of the mode θ_j conditioned on the point x_i with label y_i. Then, the

objective can be decomposed as

$$
\begin{aligned}
K(\mathbb{X}, \Theta) &= \sum_{i=1}^{N} \ln \left(\sum_{j=1}^{M} g(x_i, y_i, \theta_j) \right) \\
&= -\sum_{i=1}^{N} \left(\sum_{j=1}^{M} \gamma(\theta_j | x_i, y_i) \right) \ln \left(\frac{1}{\sum_{k=1}^{M} g(x_i, y_i, \theta_k)} \right) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{M} \gamma(\theta_j | x_i, y_i) \ln \left(\frac{g(x_i, y_i, \theta_j)}{\gamma(\theta_j | x_i, y_i)} \right) - \sum_{i=1}^{N} \sum_{j=1}^{M} \gamma(\theta_j | x_i, y_i) \ln \left(\frac{g(x_i, y_i, \theta_j)}{\gamma(\theta_j | x_i, y_i)} \right) \\
&\quad - \sum_{i=1}^{N} \sum_{j=1}^{M} \gamma(\theta_j | x_i, y_i) \ln \left(\frac{1}{\sum_{k=1}^{M} g(x_i, y_i, \theta_k)} \right) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{M} \gamma(\theta_j | x_i, y_i) \ln \left(\frac{g(x_i, y_i, \theta_j)}{\gamma(\theta_j | x_i, y_i)} \right) \\
&\quad - \sum_{i=1}^{N} \sum_{j=1}^{M} \gamma(\theta_j | x_i, y_i) \ln \left(\frac{g(x_i, y_i, \theta_j)}{\left(\sum_{k=1}^{M} g(x_i, y_i, \theta_k) \right) \gamma(\theta_j | x_i, y_i)} \right) \\
&= \sum_{i=1}^{N} \mathcal{L}_i (\gamma, \Theta) + \sum_{i=1}^{N} \mathcal{K}_i (\gamma || p)
\end{aligned}
$$

where

$$
\mathcal{L}_i (\gamma, \Theta) = \sum_{j=1}^{M} \gamma(\theta_j \mid x_i, y_i) \ln \left(\frac{g(x_i, y_i, \theta_j)}{\gamma(\theta_j \mid x_i, y_i)} \right)
$$

and

$$
\mathcal{K}_i(\gamma || p) = -\sum_{j=1}^{M} \gamma(\theta_j \mid x_i, y_i) \ln \left(\frac{p(\theta_j \mid x_i, y_i)}{\gamma(\theta_j \mid x_i, y_i)} \right)
$$

denotes the Kullback-Leibler divergence of the two probabilities. Since the latter is non-negative, the function $\sum_{i=1}^{N} \mathcal{L}_i (\gamma, \Theta)$ constitutes a lower bound for the objective $K(\mathbb{X}, \Theta)$.

Within a generalized EM scheme, starting from a random initialization of the model parameters θ_j as random data points x_i with suitable label, an iterative improvement of the objective is possible as shown in Algorithm 1, similar to a classical EM scheme as introduced in [11, 25]. Note that the objective $K(\mathbb{X}, \Theta)$ is

improved in every adaptation cycle, since step 2 sets the Kullback-Leibler divergence to 0 such that, for this choice of γ_{ji}, the objective coincides with its lower bound $\sum_{i=1}^{N} \mathcal{L}_i(\gamma, \Theta)$. Step 3 improves this function per definition. Since only a finite number of different model parameters θ_j are available, stemming from the given exemplars, the algorithm converges in a finite number of iterations.
Note that step 2 can easily be realized by setting

$$\gamma_{ji} \leftarrow \frac{\delta_{c_j}^{y_i} \cdot p(\theta_j) p(x_i|\theta_j)}{\sum_j \delta_{c_j}^{y_i} \cdot p(\theta_j) p(x_i|\theta_j)}$$

The objective $\sum_{i=1}^{N} \mathcal{L}_i(\gamma, \Theta)$ in step 3 decomposes for every j into summands of the form

$$\sum_{i=1}^{N} \gamma_{ji} \ln \frac{g(x_i, y_i, \theta_j)}{\gamma_{ji}}$$

hence we can optimize the terms by setting a parameter $\theta_j \leftarrow x_l$ where

$$\sum_{i=1}^{N} \gamma_{ji} \ln g(x_i, y_i, \theta_j) < \sum_{i=1}^{N} \gamma_{ji} \ln g(x_i, y_i, x_l)$$

This can be evaluated in $\mathcal{O}(N^2)$ for every mode j and all x_l .

Alternative objectives

This generalized EM scheme allows us to transfer the method to alternative costs which are not explicitly formulated as likelihood optimization, but which can be

---

Algorithm 1 Generalized EM algorithm for the optimization of the likelihood cost function

---

1. Initialize Θ^{old}
2. **E Step:** $\gamma_{ji} := \gamma(\theta_j \mid x_i, y_i) \leftarrow p(\theta_j^{\text{old}} \mid x_i, y_i)\ \forall j, i$
3. **M Step:** for fixed γ_{ji}, determine Θ^{new} which improves the function $\sum_{i=1}^{N} \mathcal{L}_i(\gamma, \Theta)$
4. If $\Theta^{\text{new}} = \Theta^{\text{old}}$ then stop, else set: $\Theta^{\text{old}} \leftarrow \Theta^{\text{new}}$ and go to step 2.

---

decomposed into a sum of non-negative contributions. Alternative optimization schemes for LVQ type classifiers have been proposed in [32] based on the objective of margin maximization and in [35] based on the misclassification error. These objectives can be decomposed in a similar way, offering the possibility to introduce exemplar based counterparts suited for general dissimilarity data. Note that, albeit the training objective is different in these settings, the form of the resulting classifier is the same as for RSLVQ, being an exemplar based nearest neighbor classifier. We exemplarily evaluate median generalized LVQ (mGLVQ) as extension of [32].

4 Experiments

We evaluate the proposed model in comparison to alternatives using the benchmark scenarios as proposed in [7]. These benchmarks contain dissimilarity data represented in terms of pairwise dissimilarities only. In general, these data are non-Euclidean, such that SVM techniques cannot directly be applied. The approach [7] investigates a preprocessing of the data by diverse techniques to enforce a positive semidefinite kernel for SVM. In addition to SVM, we compare to an exemplar-based unsupervised clustering with posterior labeling obtained by affinity propagation (AP) [13], and kernel LVQ variants and relational LVQ, which implicitly embed data in Euclidean or pseudo-Euclidean space [17]. Note that only the exemplar based techniques AP and the LVQ variant as developed in this contribution represent data in terms of a small number of exemplars suitable for a direct inspection. Both, kernel and relational LVQ, represent prototypes in terms of distributed coefficients only. For SVM and kernel variants, preprocessing of non-Euclidean data is necessary; for this purpose the best results obtained by clip, flip, or shift are reported [7].
The data sets are as follows:

1. Voting contains 435 samples in 2 classes, representing categorical data compared based on the value difference metric [7].

2. Aural Sonar consists of 100 signals with two classes (target of inter-

est/clutter), representing sonar signals with dissimilarity measures according to an ad hoc classification of humans [7].

3. Protein consists of 213 data from 4 classes, representing globin proteins compared by an evolutionary measure [7].

4. Face Recognition consists of 945 samples with 139 classes, representing faces of people, compared by the cosine similarity [7].

5. The sonatas data set contains complex symbolic data similar to [24]. It contains dissimilarities between 1,068 sonatas from the classical period (Beethoven, Mozart and Haydn) and the baroque era (Scarlatti and Bach). The data are in the MIDI file format, taken from the online MIDI collection Kunst der Fuge[1]. Their mutual dissimilarities are measured with the normalized compression distance (NCD), see [8], which is applied to a a specific preprocessing, which integrates invariances for music information retrieval, see [24]. The musical pieces are classified according to their composer.

6. The Copenhagen Chromosomes data set constitutes a benchmark from cytogenetics [22]. A set of 4,200 human chromosomes from 21 classes (the autosomal chromosomes) are represented by grey-valued images. These are transferred to strings measuring the thickness of their silhouettes. These strings are compared using edit distance [26].

7. The Vibrio data set consists of 1,100 samples of vibrio bacteria populations characterized by mass spectra. The spectra contain approx. 42,000 mass positions. The full data set consists of 49 classes of vibrio-sub-species. The mass spectra are preprocessed with a standard workflow using the BioTyper software [23]. As usual, mass spectra display strong functional characteristics due to the dependency of subsequent masses, such that problem adapted similarities such as described in [2, 23] are beneficial. In our case, similarities are calculated using a specific similarity measure as provided by

[1] www.kunstderfuge.com

the BioTyper software[23]. The Vibrio similarity matrix S has a maximum score of 3. The corresponding dissimilarity matrix is obtained as $D = 3 - S$.

All data sets are characterized by dissimilarity matrices only, which are symmetric, but not Euclidean. One can characterize the non-Euclideanity of the data by a reference to the signature, which corresponds to the triplet formed by the number of positive eigenvalues, the number of negative eigenvalues, and the number of (numerically) zero eigenvalues of a pseudo-Euclidean embedding of the data [28]. Obviously, data are pdf iff the second entry is zero. For the data as described above, we obtain the following signature values:

Voting	Aural	Protein	FaceRec	Sonatas	Chromosom	Vibrio
(16,1,418)	(61,38,1)	(169,38,6)	(45,0,900)	(1063,4,1)	(1951,2206,43)	(573,527,0)

This indicates, that Voting, FaceRec, and Sonatas are almost Euclidean while all other data contain a significant contribution of non-Euclidean nature.

For all experiments, the setup as described in [7] is used, i.e. results are obtained by a repeated ten-fold cross-validation with ten repeats. Parameters are optimized by a cross-validation within this scheme. The number of prototypes is chosen as a small multiple of the number of classes. We report the result of median RSLVQ, as described in this contribution, and median GLVQ, which can be derived in an analogous way based on the GLVQ cost function, the latter implicitly formalizing the objective to optimize the hypothesis margin of the classifier [32, 33]. To avoid local optima while iterative optimization of the M step, we use 10 random restarts for this step.

Interestingly, the median variants based on a probabilistic framework (mRSLVQ) and a large hypothesis margin approach (mGLVQ) provide almost identical results. In all but one case, the two discriminative exemplar-based techniques improve the performance of the exemplar based unsupervised method AP, clearly indicating that taking label information into account while training has beneficial effects for clustering tasks. In all but one case, the results obtained by median LVQ variants are comparable to best results obtained by relational or kernel LVQ variants, the latter implicitly embedding data in a high dimensional Hilbert space (possibly after preprocessing a non-Euclidean data matrix), or pseudo-Euclidean

	mRSLVQ	mGLVQ	Relational/Kernel RSLVQ/GLVQ	AP	SVM	# Prototypes
Voting	**0.956**	**0.956**	0.9466	0.935	0.9511	20 (20)
Aural	**0.91**	0.907	0.8875	0.685	0.88	6 (10)
Protein	0.912	0.904	**0.986**	0.771	0.9802	4 (20)
Face Rec	0.986	**0.987**	0.9665	0.951	0.9627	139 (139)
Sonatas	0.799	0.808	0.8493	0.7087	**0.8914**	5 (5)
Chromosom	0.854	0.889	0.9571	0.895	**0.9755**	105 (21)
Vibrio	1	1	1	0.99	1	49 (49)

Table 1: Results of Median RSLVQ (mRSLVQ) and Median GLVQ (mGLVQ) in comparison with the best results for Relational and Kernel variants of LVQ, with Affinity Propagation (AP) and Support Vector Machines (SVM) taking the best data preprocessing from clip/flip/shift for SVM and kernel LVQ variants. The classification accuracy was produced by repeated 10-fold cross-validation with 10 repeats. The last column contains the number of prototypes used for mRSLVQ/mGLVQ and in brackets the number of prototypes which was used for the kernel / relational variants.

case, respectively. Unlike the latter which represent prototypes in a distributed way, median LVQ represents prototypes in the form of a single exemplar, i.e. a data point, which can be directly inspected by a human observer in the same form as data points. In three cases, the results obtained by median LVQ are better than SVM, whereby the former represent data in terms of a small number of representative exemplars and not by points lying at the class boundaries, and the former do not require preprocessing of the data in case of a non-Euclidean signature.
For two data sets, Sonatas and Chromosomes, the classification accuracy is worse than SVM results by 10%. These data sets are the two largest data sets each containing more than 1000 data points. It can be expected that SVM benefits from the possibility to fine tune the decision boundaries in these cases, which is not possible for LVQ variants with a small number of prototypes per class. Interestingly, Chromosomes is the only data set where the unsupervised exemplar based technique AP and relational variants obtain a classification accuracy which is better by 4% resp. 10% accuracy, indicating that the choice of exemplars seems tricky in this case, giving rise to local optima of the algorithm.
However, in general, the results show that discriminative exemplar based technique are able to improve unsupervised exemplar based techniques, reaching state-of-the-art performance in all but two data sets. Since the methods can be used without data preprocessing also for non-Euclidean settings, and since they offer interpretable models in terms of few data exemplars, median LVQ variants seem a good alternative in settings where interpretability constitutes a crucial requirement.

5 Conclusions

The supervised generative model RSLVQ has been extended to general dissimilarity data by means of an exemplar based approach. Optimization of the cost function could be done based on a generalized EM scheme, which provably converges towards a local optimum in a finite number of steps in this setting. Unlike relational or kernel LVQ variants, the model preserves the intuitive interpretability of classical LVQ for the non-Euclidean case by restricting prototypes to data

positions. Unlike kernel techniques, preprocessing of non-Euclidean data to enforce pdf is superfluous. As demonstrated in experiments, this approach can lead to sparse models with state-of-the-art performance.

Acknowledgment

The authors gratefully acknowledge very helpful discussions with A. GISBRECHT, BIELEFELD UNIVERSITY, CITEC. D. Nebel was supported by a grant of the European Social Fund, Saxony (ESF).

References

[1] W. Arlt, M. Biehl, A. E. Taylor, S. Hahner, R. Libé, B. A. Hughes, P. Schneider, D. J. Smith, H. Stiekema, N. Krone, E. Porfiri, G. Opocher, J. Bertherat, F. Mantero, B. Allolio, M. Terzolo, P. Nightingale, C. H. L. Shackleton, X. Bertagna, M. Fassnacht, and P. M. Stewart. Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors. *Journal of Clinical Endocrinology & Metabolism*, 96:3775–3784, 2011.

[2] S. B. Barbuddhe, T. Maier, G. Schwarz, M. Kostrzewa, H. Hof, E. Domann, T. Chakraborty, and T. Hain. Rapid identification and typing of listeria species by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Applied and Environmental Microbiology*, 74(17):5402–5407, 2008.

[3] V. V. Belle and P. Lisboa. Research directions in interpretable machine learning models. In *European Symposium on Artificial Neuronal Networks, Computational Intelligence and Machiene Learning*, 2013.

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

[5] C. M. Bishop. *Pattern recognition and Machine Learning*. Springer, 2006.

[6] C. M. Bishop and M. Svensen. gtm: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[7] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10:747–776, 2009.

[8] R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *CoRR*, cs.CV/0312044, 2003.

[9] K. Crammer, R. Gilad-Bachrach, A. Navot, and N. Tishby. Margin analysis of the lvq algorithm. *NIPS*, pages 462–469, 2002.

[10] G. Da San Martino and A. Sperduti. Mining structured data. *Computational Intelligence Magazine, IEEE*, 5:42–49, 2010.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

[12] A. Denecke, H. Wersing, J. Steil, and E. Koerner. Online figure-ground segmentation with adaptive metrics in generalized lvq. *Neurocomputing*, 72(7-9):1470–1482, 2009.

[13] B. J. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, pages 972–976, January 2007.

[14] N. Gianniotis and P. Tiňo. Visualization of tree-structured data through generative topographic mapping. *IEEE Transactions on Neural Networks*, 19:1468–1493, 2008.

[15] A. Gisbrecht and B. Hammer. Relevance learning in generative topographic mapping. *Neurocomputing*, 74(9):1359–1371, 2011.

[16] T. Gärtner and G. C. Garriga. Guest editors' introduction: special issue on mining and learning with graphs. *Machine Learning*, 75(1):1–2, 2009.

[17] B. Hammer, D. Hofmann, F.-M. Schleif, and X. Zhu. Learning vector quantization for (dis-)similarities. Neurocomputing (in press).

[18] P. J. Ingram, M. P. Stumpf, and J. Stark. Network motifs: structure does not determine function. *BMC Genomics*, 7:108, 2006.

[19] B. J. Jain and K. Obermayer. Structure spaces. *The Journal of Machine Learning Research*, 10:2667–2714, 2009.

[20] T. Kietzmann, S. Lange, and M. Riedmiller. Incremental grlvq: Learning relevant features for 3d object recognition. *Neurocomputing*, 71 (13-15):2868–2879, 2008.

[21] T. Kohonen. *Self-Organizing Maps*. Springer, 2001.

[22] C. Lundsteen, J. Phillip, and E. Granum. Quantitative analysis of 6985 digitized trypsin g-banded human metaphase chromosomes. *Clinical Genetics*, 18:355–370, 1980.

[23] T. Maier, S. Klepel, U. Renner, and M. Kostrzewa. Fast and reliable maldi-tof ms–based microorganism identification. *Nature Methods*, 3, 2006.

[24] B. Mokbel, A. Hasenfuss, and B. Hammer. Graph-based representation of symbolic musical data. In A. Torsello, F. Escolano, and L. Brun, editors, *GbRPR*, volume 5534 of *Lecture Notes in Computer Science*, pages 42–51. Springer, 2009.

[25] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. MIT Press, 1999.

[26] M. Neuhaus and H. Bunke. Edit distance-based kernel functions for structural pattern classification. *Pattern Recognition*, 39(10):1852–1863, 2006.

[27] S. Parameswaran and K. Q. Weinberger. Large margin multi-task metric learning. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *NIPS*, pages 1867–1875. Curran Associates, Inc., 2010.

[28] E. Pekalska and R. P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications (Series in Machine Perception and Artifical Intelligence)*. 2005.

[29] O. Penner, P. Grassberger, and M. Paczuski. Sequence alignment, mutual information, and dissimilarity measures for constructing phylogenies. *PLOS ONE*, 6(1), 2011.

[30] A. K. Qin and P. N. Suganthan. A novel kernel prototype-based learning algorithm. In *ICPR (4)*, pages 621–624, 2004.

[31] H. Ruiz, I. H. Jarman, J. D. Martín, S. Ortega-Martorell, A. Vellido, E. Romero, and P. J. G. Lisboa. Towards interpretable classifiers with blind signal separation. In *IJCNN*, pages 1–7. IEEE, 2012.

[32] A. Sato and K. Yamada. Generalized learning vector quantization. *NIPS*, 1995.

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

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

[35] S. Seo, M. Bode, and K. Obermayer. Soft nearest prototype classification. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 14(2):390–398, March 2003.

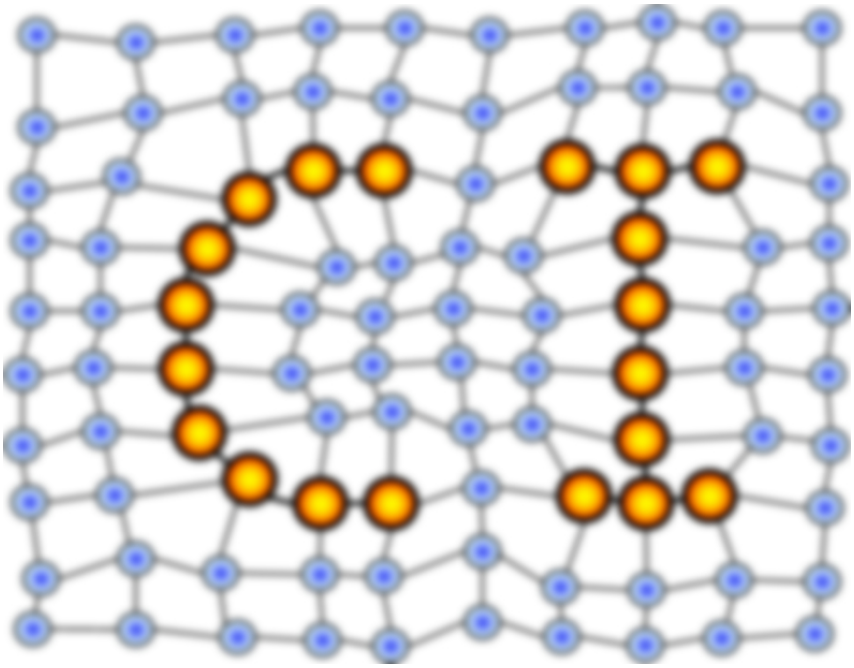
[36] S. Seo and K. Obermayer. Soft learning vector quantization. *Neural Computation*, 15(7):1589–1604, 2003.

[37] J. J. Thomas and K. A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[38] M. E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

MACHINE LEARNING REPORTS

Report 06/2013



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.