ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Supervised Generative Models for Learning Dissimilarity Data

D. Nebel[1]*, B. Hammer[2], and T. Villmann[1]

1- University of Appl. Sciences Mittweida - Dept. of Mathematics
Mittweida, Saxonia - Germany

2- CITEC Centre of Excellence,Theoretical Computer Science Group
Bielefeld University, Germany

Abstract. Exemplar based techniques such as affinity propagation [1] 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. Here, we address the problem of learning exemplar-based models for general dissimilarity data in a discriminative framework. For this purpose, we extend a generative model proposed in [2] to an exemplar based scenario using a generalized EM framework for its optimization. The resulting classifiers represent data in terms of sparse models while keeping high performance in state-of-the art benchmarks.

1 Introduction

Machine learning has revolutionized the possibility to deal with large 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 [3, 4, 5]. These data cannot be embedded in Euclidean space, and they often do not even fulfill the properties of a metric. Further, dissimilarities might fail due to asymmetry. These developments have caused the need for non-vectorial machine learning tools such as e.g. structure kernels, recursive and relational models, affinity propagation (AP) or quotient embeddings [6, 7, 1, 8]. While learning tasks become more and more complex a vital property of machine learning models in this context is their interpretability. Popular black box techniques such as the SVM often provide an excellent classification performance, but no insight on why this decision is obtained such that relevant information can be inferred based thereon by a human observer.
The demand of interpretability can be met with quite diverse technologies, such as sparsity, relevance learning, or enhancement by visualization [9]. Yet, dissimilarity based learning is usually easy to interpret the decision if a small number of most similar neighbors with known labels accounts for the observed classification . These neighbors can directly be inspected by experts in the same way as data points. Dissimilarity based techniques can be distinguished according to different criteria: **(i) Sparsity**: 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. Usually, sparsity supports interpretability of the models. **(ii) Complexity of the dissimilarity measure** the methods can deal with ranging from vectorial techniques restricted to Euclidean spaces, adaptive models which learn the underlying metrics, up to tools which

*D. Nebel was supported by a grant of the European Social Fund, Saxony (ESF).

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

can deal with arbitrary similarities or dissimilarities [10, 11, 12]. **(ii) Degree of supervision** ranging from clustering techniques such as AP to supervised learning.
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 [13], which decisions can directly be inspected by humans. Albeit original LVQ has been introduced on somewhat heuristic grounds [13], recent developments provide a solid mathematics, its generalization ability and learning dynamics [14, 11, 2, 15]. A severe drawback of LVQ classifiers is their dependency on the Euclidean metric. This problem can partially be avoided by appropriate metric learning, see e.g. [11], or by kernel variants, see e.g. [12, 16]. However, if data are inherently non-Euclidean, these techniques cannot be applied. Recently, an extension of LVQ type learning by means of an implicit pseudo-euclidean embedding has been proposed [17]. Although yielding excellent results, this technique faces two problems: it cannot be applied for asymmetric dissimilarities where no pseudo-euclidean embedding exists; by representing prototypes in terms of distributed coefficient vectors, the easy interpretability of LVQ's is lost.
In this contribution, we take 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 [1]. A training algorithm can be derived as a generalized expectation maximaization (EM) scheme, yielding a state-of-the-art classifier with superior performance as opposed to unsupervised exemplar-based approaches [1].

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 [2] 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 [2], the correlation matrix is taken as unit matrix, a generalization towards a general form has been proposed in [18]. 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

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

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Trainings takes place by an optimization of the log likelihood ratio as cost function

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

assuming i.i.d. data. In Euclidean space, gradient methods 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: *1)* 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? *2)* 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 restrict ourself in settings of pairwise dissimilarities $d(x_i, x_j) \geq 0$ only, as discussed e.g. in [19, 20]. Thereby, we do not suppose 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 [19].
Since the underlying space $\mathbb{X}$ is unknown, we take an exemplar based approach similar to [1]: 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.
For optimization of the model parameters, instead of gradient techniques as used in the vectorial case, a generalized EM strategy is possible. For this purpose, we consider the nonnegative function

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

defining the conditional probability of θ_j. Then, the objective (1) decomposes into

$$K(\mathbb{X}, \Theta) = \sum_{i=1}^{N} \ln \sum_{j=1}^{M} g(x_i, y_i, \theta_j) = \sum_{i=1}^{N} \mathcal{L}_i(\gamma, \Theta) + \sum_{i=1}^{N} \mathcal{K}_i(\gamma || p) \tag{2}$$

with $\gamma(\theta_j|x_i, y_i)$ is an arbitrary probability distribution of the mode θ_j conditioned on the point x_i with label y_i, see [21]. Further, we have

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

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Algorithm 1 Generalized EM algorithm for optimization of the cost function $K(\mathbb{X}, \Theta)$

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.

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 (nonnegative) Kullback-Leibler divergence between $p(\theta_j \mid x_i, y_i)$ and $\gamma(\theta_j \mid x_i, y_i)$ such that $\sum_{i=1}^{N} \mathcal{L}_i(\gamma, \Theta)$ constitutes a lower bound for $K(\mathbb{X}, \Theta)$.
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 [22, 23]. 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. For details we refer to [21].

4 Experiments

We evaluate the proposed model in comparison to alternatives using seven benchmark scenarios as proposed and described in [20]. These benchmarks contain dissimilarity data represented in terms of pairwise symmetric dissimilarities only, which are in general non-Euclidean, such that SVM techniques can only applied after appropriate preprocessing [20]. In addition to SVM, we compare to an exemplar-based unsupervised clustering with posterior labeling obtained by AP [1], and kernel LVQ variants and relational LVQ, which implicitly embed data into the 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 [20].
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 [19]. Obviously, data are pdf iff the second entry is zero. For the used datasets 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)

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

	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: Averaged 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), see text. 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.

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 [20] 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 and median GLVQ (mGLVQ), which can be derived in an analogous way based on the cost function of the generalized LVQ (GLVQ) [24], the latter implicitly formalizing the objective to optimize the hypothesis margin of the classifier [25, 11]. 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 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 outperform the classification accuracy by at least 4%.

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

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

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.

References

[1] Brendan J J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, pages 972–976, January 2007.
[2] Sambu Seo and Klaus Obermayer. Soft learning vector quantization. *Neural Computation*, 15(7):1589–1604, 2003.
[3] Orion Penner, Peter Grassberger, and Maya Paczuski. Sequence alignment, mutual information, and dissimilarity measures for constructing phylogenies. *PLOS ONE*, 6(1), 2011.
[4] Thomas Maier, Stefan Klepel, Uwe Renner, and Markus Kostrzewa. Fast and reliable maldi-tof ms–based microorganism identification. *Nature Methods*, 3, 2006.
[5] Piers J Ingram, Michael PH Stumpf, and Jaroslav Stark. Network motifs: structure does not determine function. *BMC Genomics*, 7:108, 2006.
[6] Brijnesh J. Jain and Klaus Obermayer. Structure spaces. *The Journal of Machine Learning Research*, 10:2667–2714, 2009.
[7] Thomas Gärtner and Gemma C. Garriga. Guest editors' introduction: special issue on mining and learning with graphs. *Machine Learning*, 75(1):1–2, 2009.
[8] G. Da San Martino and A. Sperduti. Mining structured data. *Computational Intelligence Magazine, IEEE*, 5:42–49, 2010.
[9] Vanya Van Belle and Paulo Lisboa. Research directions in interpretable machine learning models. In *European Symposium on Artificial Neuronal Networks, Computational Intelligence and Machiene Learning*, 2013.
[10] Shibin Parameswaran and Kilian Q. Weinberger. Large margin multi-task metric learning. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *NIPS*, pages 1867–1875. Curran Associates, Inc., 2010.
[11] Petra Schneider, Michael Biehl, and Barbara Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532–3561, 2009.
[12] A. Kai Qin and Ponnuthurai N. Suganthan. A novel kernel prototype-based learning algorithm. In *ICPR (4)*, pages 621–624, 2004.
[13] Teuvo Kohonen. *Self-Organizing Maps*. Springer, 2001.
[14] Koby Crammer, Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. Margin analysis of the lvq algorithm. *NIPS*, pages 462–469, 2002.
[15] Sambu Seo, Mathias Bode, and Klaus Obermayer. Soft nearest prototype classification. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 14(2):390–398, March 2003.
[16] T. Villmann, S. Haase, and M. Kaden. Kernelized vector quantization in gradient-descent learning. *Neurocomputing*, page in press, 2014.
[17] Barbara Hammer, Daniela Hofmann, Frank-Michael Schleif, and Xibin Zhu. Learning vector quantization for (dis-)similarities. Neurocomputing (in press).
[18] P. Schneider, M. Biehl, and B. Hammer. Distance learning in discriminative vector quantization. *Neural Computation*, 21:2942–2969, 2009.
[19] Elzbieta Pekalska and Robert P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications (Series in Machine Perception and Artifical Intelligence)*. 2005.
[20] 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.
[21] D. Nebel, B. Hammer, and T. Villmann. About learning of supervised generative models for dissimilarity data. *Machine Learning Reports*, 7(MLR-05-2013):1–19, 2013. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/˜fschleif/mlr/mlr_05_2013.pdf.
[22] 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.
[23] 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.
[24] D. Nebel and T. Villmann. A median variant of generalized learning vector quantization. In M. Lee, A. Hirose, Z.-G. Hou, and R.M. Kil, editors, *Proceedings of International Conference on Neural Information Processing (ICONIP)*, volume II of *LNCS*, pages 19–26, Berlin, 2013. Springer-Verlag.
[25] A. Sato and K. Yamada. Generalized learning vector quantization. *NIPS*, 1995.