Nonlinear Discriminative Data Visualization

Kerstin Bunte[1], Barbara Hammer[2], Petra Schneider[1], Michael Biehl[1]

1- University of Groningen - Institute of Mathematics and Computing Sciences
P.O. Box 407, 9700 AK Groningen - The Netherlands

2- Clausthal University of Technology - Institute of Informatics
Julius Albert Strasse 4, 38678 Clausthal-Zellerfeld - Germany

Abstract. Due to the tremendous increase of electronic information with respect to the size of data sets as well as dimensionality, visualization of high-dimensional data constitutes one of the key problems of data mining. Since embedding in lower dimensions necessarily includes a loss of information, methods to explicitly control the information kept by a specific visualization technique are highly desirable. The incorporation of supervised class information constitutes an important specific case. In this contribution we propose an extension of prototype-based local matrix learning by a charting technique which results in an efficient nonlinear dimension reduction and discriminative visualization of a given labelled data manifold.

1 Introduction

Visualization of high-dimensional data constitutes an active field of research, see e.g. [1, 2, 3] for recent overviews. The embedding of high-dimensional data into lower dimensionality is necessarily linked to loss of information. Since the relevant information depends on the situation at hand, data visualization constitutes an inherently ill-posed problem. In consequence, visualization techniques address different objectives such as distance or topology preservation.

One systematic way to explicitly guide the information conserved in a low-dimensional embedding has been pioneered in [4]: auxiliary information is incorporated and only the aspects which contribute to this information are visualized. The approach [4] proposes a corresponding adaptation of the self-organizing map (SOM, [5]) resulting in an accurate but very costly model. Explicit class labels constitute one relevant special case of this general approach. This setting is addressed by classical linear discriminance analysis (LDA, [6]), which is restricted to a linear visualization in at most $C-1$ dimensions, C being the number of classes [1]. Alternative linear discriminative visualizations include targeted projection pursuit [7] and discriminative component analysis [8]. Nonlinear extensions can be reached by kernelization as proposed e.g. in [6], leading to quadratic complexity w.r.t. the number of data points. Alternatives include an extension of SOM, incorporating class labels into its cost function [9] and other supervised techniques, like model-based visualization [10] and parametric embedding [11] among others.

In this contribution we propose an efficient nonlinear discriminative visualization technique which combines prototype-based classification and recent matrix learning schemes, resulting in local linear views of the data and a charting step which merges the individual projections. The procedure leads to an explicit nonlinear mapping of the data manifold to lower dimensions.

2 Prototype-based matrix learning

Learning vector quantization (LVQ) [5] constitutes a particularly intuitive classification algorithm which represents data by means of prototypes. LVQ itself constitutes a heuristic algorithm, hence extensions have been proposed for which convergence and learnability can be guaranteed [12, 13]. One particularly crucial aspect of LVQ schemes is the dependency of the underlying metric, usually the Euclidean metric. Therefore, general metric adaptation has been introduced into LVQ schemes [13, 14]. Besides an increased capacity, linear data visualization schemes within the receptive fields of the classifier are provided [15]. This scheme will constitute the first step of our visualization pipeline.

Assume labelled training data $\{\boldsymbol{x}_i, c(\boldsymbol{x}_i)\}_{i=1}^N \in \mathbb{R}^n \times \{1, \ldots, C\}$ are given. The aim of LVQ is to find k prototypes $\boldsymbol{w}_j \in \mathbb{R}^n$ with class labels $c(\boldsymbol{w}_j) \in \{1, \ldots, C\}$ such that they represent the classification as accurately as possible. A data point $\boldsymbol{x}_i$ is assigned to the class of its closest prototype $\boldsymbol{w}_j$ where $d(\boldsymbol{x}_i, \boldsymbol{w}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{w}_l)$ for all $j \neq l$. d usually denotes the squared Euclidean distance $d(\boldsymbol{x}_i, \boldsymbol{w}_j) = (\boldsymbol{x}_i - \boldsymbol{w}_j)^\top (\boldsymbol{x}_i - \boldsymbol{w}_j)$. Generalized LVQ (GLVQ) adapts prototype locations by minimizing the cost function

$$E_{\mathrm{GLVQ}} = \sum_{i=1}^{N} \Phi \left(\frac{d(\boldsymbol{w}_J, \boldsymbol{x}_i) - d(\boldsymbol{w}_K, \boldsymbol{x}_i)}{d(\boldsymbol{w}_J, \boldsymbol{x}_i) + d(\boldsymbol{w}_K, \boldsymbol{x}_i)} \right), \tag{1}$$

where $\boldsymbol{w}_J$ denotes the closest prototype with the same class label as $\boldsymbol{x}_i$, and $\boldsymbol{w}_K$ is the closest prototype with a different class label. Φ is a monotonic function, e.g. the logistic function or the identity, which we used in this work. This cost function aims at an adaptation of the prototypes such that a large hypothesis margin is obtained, this way achieving correct classification and, at the same time, robustness of the classification. A learning algorithm can be derived from the cost function E_{GLVQ} by means of a stochastic gradient descent as shown in [13, 12]. Matrix learning in GLVQ (GMLVQ) substitutes the usual squared Euclidean distance d by a more advanced dissimilarity measure which carries adaptive parameters, thus resulting in a more complex and better adaptable classifier. In [15], it was proposed to choose the dissimilarity as

$$d_k(\boldsymbol{w}_k, \boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{w}_k)^\top \Lambda_k (\boldsymbol{x}_i - \boldsymbol{w}_k) \tag{2}$$

with an adaptive local, symmetric and positive semidefinite matrix $\Lambda_k \in \mathbb{R}^{n \times n}$. It corresponds to piecewise quadratic receptive fields. Positive semidefiniteness and symmetry can be guaranteed by setting $\Lambda_k = \Omega_k^\top \Omega_k$ for $\Omega_k \in \mathbb{R}^{a \times n}$ with arbitrary $a \leq n$, so data are transformed locally by Ω_k according to the classification task. Optimization takes place by a gradient descent of the cost function E_{GLVQ} (1) with subsequent normalization $\sum_i [\Lambda_k]_{ii} = 1$. To prevent oversimplification effects we substract a regularization term $\mu/2 \cdot \ln(\det(\Omega_k \Omega_k^\top))$ with $\mu > 0$ from the cost function, see [14] for details. The precise update formulas and definitions of parameters can be found in [15, 14]. Note that Ω_k is not uniquely given by (1), the dissimilarity being invariant under rotation, for example. We choose a unique representation $\widehat{\Omega}_k$ of Ω_k by using a canonical representation, which is explained in [15].

Besides providing local classification rules, matrix adaptation gives rise to local linear transformations of the data space

$$P_k : \boldsymbol{x} \mapsto \widehat{\Omega}_k^\top (\boldsymbol{x} - \boldsymbol{w}_k) \tag{3}$$

which emphasize the class represented by $\boldsymbol{w}_k$. We can restrict the rank of $\Omega_k \in \mathbb{R}^{m \times n}$ to $m \in \{2, 3\}$, such that low-dimensional visualizations are obtained. Alternatively, we can extract the largest eigenvectors from $\widehat{\Omega}_k$. For identical $\widehat{\Omega}_k = \widehat{\Omega}$ for all k, a global linear discriminative visualization is obtained [15].

3 Visualization by matrix charting

For general data sets, a faithful global *linear* visualization will not be possible. Therefore, we apply an alternative which glues together the local linear projections P_k provided by local GMLVQ to a global *nonlinear* discriminative embedding of the data. For this purpose, we use a technique introduced in [16] in the frame of unsupervised data visualization. Local GMLVQ provides k local linear projections $P_k(\boldsymbol{x}_i) \in \mathbb{R}^m$ for every point $\boldsymbol{x}_i$. We assume that responsibilities $p_{ki} = p_k(\boldsymbol{x}_i)$ of prototype $\boldsymbol{w}_k$ for data point $\boldsymbol{x}_i$ are available with $\sum_k p_{ki} = 1$, e. g. chosen as $p_k(\boldsymbol{x}_i) \propto \exp(-(\boldsymbol{x}_i - \boldsymbol{w}_k)^\top \Lambda_k (\boldsymbol{x}_i - \boldsymbol{w}_k)/\sigma_k)$ where $\sigma_k > 0$ is an appropriate bandwith. The bandwith σ_k has to be determined appropriately such that a reasonable overlap of neighbored charts is obtained. One way is to set σ_k as half the mean Euclidean distance of prototype $\boldsymbol{w}_k$ to its closest k_1 prototypes, k_1 being a reasonable fraction of k. Charting as introduced in [16] finds affine transformations $B_k : \mathbb{R}^m \to \mathbb{R}^m$ of the local coordinates P_k such that the resulting points coincide on overlapping parts, i.e. it minimizes the cost function

$$E_{\text{charting}} = \sum_{i,k,j} p_{ki} p_{ji} \| B_k(P_k(\boldsymbol{x}_i))) - B_j(P_j(\boldsymbol{x}_i)) \|^2 . \tag{4}$$

An algebraic solution can be found by reformulating the problem as a generalized eigenvalue problem, see e.g. [2]. This leads to a global embedding in $\mathbb{R}^m$

$$\boldsymbol{x} \mapsto \sum_k p_k(\boldsymbol{x}) \cdot B_k(P_k(\boldsymbol{x})) . \tag{5}$$

4 Experiments

Three tip Star: We create 3000 samples in $\mathbb{R}^4$, consisting of two classes partitioned in three modes with equal prior arranged on a star in the first two dimensions (see Fig.1 left), and two dimensions with random Gaussian noise. The four-dimensional data was then additionally rotated by a random transformation in $\mathbb{R}^4$. For training we take 900 random samples of each class and three prototypes per class, which are initialized by randomly selected samples. Training is done for 300 epochs, matrix learning starts after 30 epochs. We train LVQ schemes with restricted rank 2 matrices and regularization parameter $\mu = 0.1$. The parameters were chosen due to the experience in other problems [15, 14] and their sensitivity will be evaluated in forthcoming projects. For the determination of responsibilities for charting, we set $k_1 = k/2 = 3$.

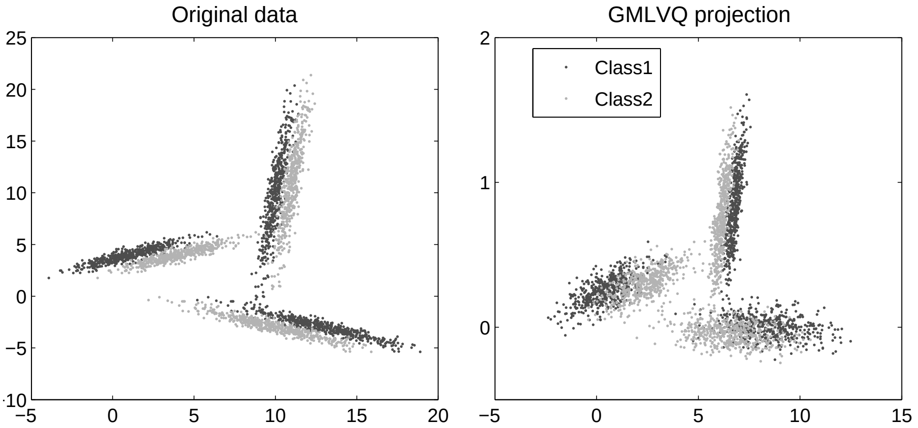

Fig. 1: Left: Visualization of the two informative dimensions of the three tip star data set before adding the 2 noise dimensions and rotation in $\mathbb{R}^4$. Right: The data projected with a global matrix learned by LVQ.

The classification error of GMLVQ reaches 0.21 (train) resp. 0.19 (test) for a global matrix, 0.039 (train) resp. 0.037 (test) for local matrices, and 0.45 (train and test) for LDA, clearly indicating that neither LDA nor global GMLVQ can capture the regularity of this data set due to its multimodality. GMLVQ with local matrix adaptation learns the regularity almost perfectly.

For visualization, we use the canonical representation $\widehat{\Omega}$ resp. $\widehat{\Omega}_k$. Visualization with local GMLVQ and charting is depicted in Fig. 2. The linear projections obtained by global GMLVQ are shown in Fig. 1. Obviously, LDA does not provide a valid projection because of its restriction to one dimension according to the number of classes, and its restriction to unimodal clusters. Compared to LDA, linear projection by global GMLVQ provides a better visualization since it correctly identifies the last two dimensions as irrelevant, and it chooses a projection which gives a good class separation for two of the six clusters corresponding to a single dominant direction in feature space. The other 4 modes overlap as can be seen in Fig. 1. Local GMLVQ and charting yields a perfect visualization of the underlying cluster shapes due to the correct identification of locally separating class boundaries, see Fig.2. The star is only rotated which can be a result from the non unique eigenvalues or from the rotation in the original four-dimensional vector.

Letter: As a second demonstration, we use the letter recognition data set from the UCI repository [17] consisting of 20.000 data points representing the 26 capital letters in 20 different fonts, digitalized by 16 primitive numerical attributes. We use 1 prototype per class and matrices restricted to rank 3. For global matrix learning, we use 500 epochs, initial learning rate 0.1 for the prototypes, and 0.01 for the matrix parameters. For local matrix learning, we use 300 epochs and learning rates 0.001 for prototypes and 0.0001 for matrices see [15] for details.

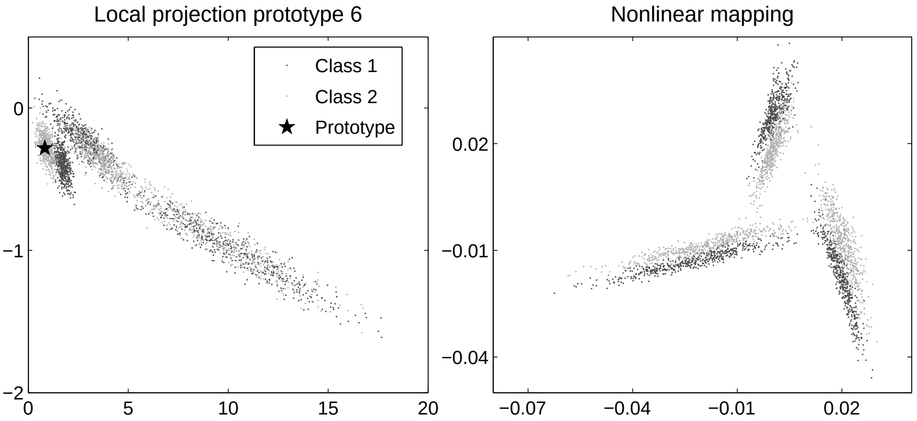

Fig. 2: Left: Star data set projected with one of the 6 local transformations learned from LVQ with local matrix adaptation. Right: Result after charting the locally projected data. Note, that it resembles to a rotated version of Fig. 1(left).

The classification error on the full data set gives 0.53 for global GMLVQ, 0.13 for local GMLVQ, and 0.295 for LDA. Local GMLVQ improves the accuracy by more than factor 2 compared to LDA.

The visualization in 3D resulting from these methods (using $k_1 = 5$) is depicted in Fig. 3. The classes corresponding to 'M', 'W', 'N', and 'A' are highlighted exemplarily. Local GMLVQ well separates the classes and correctly displays their mutual closeness; the overall data set is embedded in a manifold which resembles a saddle and which provides comparably large freedom being a model of the two dimensional hyperbolic space. Both, GMLVQ and LDA show a tendency to overlap the classes, and the overall data set resembles a sphere.

5 Conclusions

We have introduced a new discriminative nonlinear visualization technique based on recent matrix learning schemes for prototype-based classifiers and charting. Compared to alternatives such as LDA and global matrix learning, the method showed very promising results in two examples. The overall technique possesses a couple of benefits: unlike LDA and variants, it provides a nonlinear embedding of data. Its complexity is linear in the number of examples, thus providing a fast alternative to quadratic schemes such as kernel LDA. It gives an explicit smooth embedding function of the data manifold such that out of sample extensions are immediate. The combination of this visualization technique with a prototype based learning scheme offers further possibilities to interactively display the data: prototypes and their embedding in low-dimensions allow to compress the visual information of modes, which is of particular interest for huge data sets.

Acknowledgment This work was supported by the "Nederlandse organisatie voor Wetenschappelijke Onderzoek (NWO)“ under project code 612.066.620.

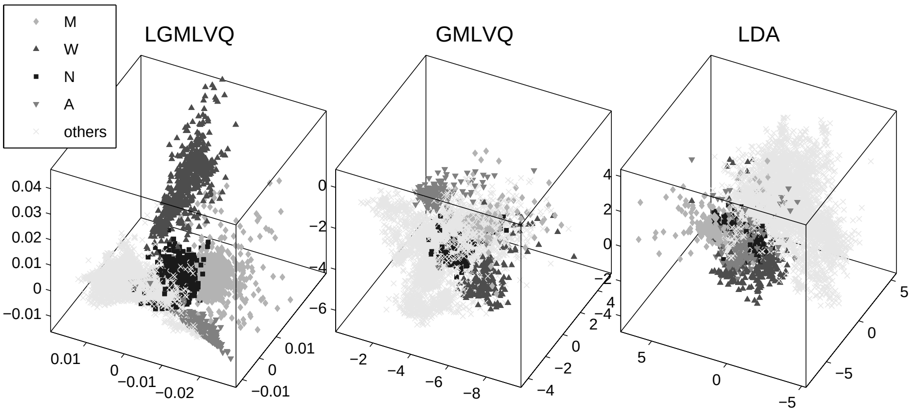

Fig. 3: Different projections of the UCI letter recognition data sets, explicitly displaying the letters M, N, and W.

References

[1] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction.* Springer, 2007.

[2] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. 2007.

[3] D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

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

[5] T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, Heidelberg, 2nd edition, 1997.

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

[7] J. Faith, R. Mintram, and M. Angelova. Targeted projection pursuit for visualising gene expression data classifications. *Bioinformatics*, 22:2667–2673, 2006.

[8] J. Peltonen, J. Goldberger, and S. Kaski. Fast discriminative component analysis for comparing examples. *NIPS*, 2006.

[9] T. Villmann, B. Hammer, F.-M. Schleif, T. Geweniger, and W. Hermann. Fuzzy classification by fuzzy labeled neural gas. *Neural Networks*, 19(6-7):772–779, 2006.

[10] Petri Kontkanen, Jussi Lahtinen, Petri Myllymäki, Tomi Silander, and Henry Tirri. Supervised model-based visualization of high-dimensional data. *Intell. Data Anal.*, 4(3,4):213–227, 2000.

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

[12] B. Hammer, M. Strickert, and T. Villmann. On the generalization ability of GRLVQ networks. *Neural Processing Letters*, 21(2):109–120, 2005.

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

[14] P. Schneider, K. Bunte, B. Hammer, T. Villmann, and M. Biehl. Regularization in matrix relevance learning. Technical Report MLR-02-2008, 2008. ISSN:1865-3960 http://www.uni-leipzig.de/∼compint/mlr/mlr_02_2008.pdf.

[15] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl. Discriminative visualization by limited rank matrix learning. Technical Report MLR-03-2008, 2008. ISSN:1865-3960 http://www.uni-leipzig.de/∼compint/mlr/mlr_03_2008.pdf.

[16] M. Brand. Charting a manifold. Technical Report 15, Mitsubishi Electric Research Laboratories (MERL), 2003. http://www.merl.com/publications/TR2003-013/.

[17] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases. `http://archive.ics.uci.edu/ml/`, last visit 19.04.2008, 1998.