ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Supervised Manifold Learning with Incremental Stochastic Embeddings

Oliver Kramer

Computational Intelligence Group,
Carl von Ossietzky University, 26111 Oldenburg, Germany
oliver.kramer@uni-oldenburg.de

Abstract. In this paper, we introduce an incremental dimensionality reduction approach for labeled data. The algorithm incrementally samples in latent space and chooses a solution that minimizes the nearest neighbor classification error taking into account label information. We introduce and compare two optimization approaches to generate supervised embeddings, i.e., an incremental solution construction method and a re-embedding approach. Both methods have in common that the objective is to minimize the nearest neighbor classification error computed in the low-dimensional space. The resulting embedding is a surrogate of the high-dimensional labeled set. The set allows conclusions about the data set structure and can be used as preprocessing step for classification of labeled patterns.

1 Introduction

Dimensionality reduction is the task to find a mapping from a high-dimensional space $\mathbb{R}^d$ to a low-dimensional space $\mathbb{R}^q$ maintaining most of the high-dimensional characteristics. Often, not an explicit mapping $f : \mathbb{R}^d \to \mathbb{R}^q$ is computed, but low-dimensional counterparts $\mathbf{Z} = [\mathbf{z}]_{i=1}^N$ for high-dimensional patterns $\mathbf{X} = [\mathbf{x}]_{i=1}^N$. Based on unlabeled data, dimensionality reduction (DR) methods try to represent the intrinsic structure of the data space. Supervised embeddings can be used for visualization and as pre-processing method for supervised learning, i.e., classification and regression.

In this paper, an approach is presented that places low-dimensional patterns based on label information. Label information is an excellent indicator for classification and regression methods. The idea of supervised embeddings is to find a low-dimensional set of latent representations for a high-dimensional set of patterns that has similar characteristics with respect to an employed supervised learning method. With such an approach, the characteristics of the data space are mapped to the latent space w.r.t. the properties the supervised method is able to represent. We will use k-nearest neighbor (kNN) classifiers to map from latent space to label space for maintaining neighborhod information. The optimization criterion to minimize the kNN error induces kNN neighborhoods with correct label information.

The incremental approach will be introduced in Section 2, while the re-embedding approach is presented in Section 3. After discussion of related work in Section 4, the approaches are experimentally analyzed in Section 5. Conclusions are drawn in Section 6.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

2 Incremental Solution Construction

In this section, we introduce an incremental DR method based on the classification accuracy criterion for labeled data sets. If patterns $\mathbf{x}_i$ carry a label y_i, the classification error in latent space $\mathbb{R}^q$ can be used for the embedding process. Optimizing w.r.t. the classification error is further motivated by the fact that DR methods are often employed as preprocessing methods in classification setups. The incremental supervised embedding (incSE) algorithm constructs a solution pattern by pattern. The mechanism can be described inductively, also see Figure 1 for the pseudocode of the incSE. We have given patterns $\mathbf{X} = [\mathbf{x}]_{i=1}^N$ with assigned labels $\mathbf{y} = (y_1, \ldots, y_N)$. The first pattern $\mathbf{x}_1$ is embedded at an arbitrary latent position, e.g., at the origin $\mathbf{z}_1 = \mathbf{0}$. This results in a latent matrix $\overline{\mathbf{Z}} = [\mathbf{z}_1]$ and a corresponding pattern matrix $\overline{\mathbf{X}} = [\mathbf{x}_1]$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the sequence of embedded patterns with corresponding latent positions $\mathbf{z}_1, \ldots, \mathbf{z}_n$. Pattern $\mathbf{x}_{n+1}$ with $n+1 \leq N$ is embedded by first searching for the closest embedded pattern

$$\mathbf{x}^* = \arg \min_{\mathbf{x}=\mathbf{x}_1,\ldots,\mathbf{x}_n} \|\mathbf{x}_{n+1} - \mathbf{x}\|^2 \tag{1}$$

of pattern matrix $\overline{\mathbf{X}} = [\mathbf{x}_j]_{j=1}^n$. Based on its latent position $\mathbf{z}^*$, κ candidate latent positions $\mathbf{z}_1^*, \ldots, \mathbf{z}_\kappa^*$ are sampled using the Gaussian distribution

$$\mathbf{z}_l^* = \mathbf{z}^* + \hat{\mathbf{z}}_l \tag{2}$$

for $l = 1, \ldots, \kappa$ with $\hat{\mathbf{z}}_l \sim \mathcal{N}(0, \sigma)$. The distance $\sigma = \|\mathbf{x}_{n+1} - \mathbf{x}^*\|$ between pattern $\mathbf{x}_{n+1}$ that has to be embedded, and the closest embedded pattern $\mathbf{x}^*$ is employed as standard deviation of the Gaussian sampling. This allows the preservation of distances between patterns in latent space. From the κ candidate latent positions, the one leading to the lowest classification $E(\cdot)$ error is selected

$$\mathbf{z}_{n+1} = \arg \min_{\mathbf{z}=\mathbf{z}_1^*,\ldots,\mathbf{z}_\kappa^*} E(\overline{\mathbf{Z}}_{n+1}), \tag{3}$$

which is defined as

$$E(\overline{\mathbf{Z}}_{n+1}) = \sum_{i=1}^{n+1} \|f_q(\mathbf{z}_i) - y_i\|^2, \tag{4}$$

where $f_q(\cdot)$ is the classifier mapping from latent space $\mathbb{R}^q$ to label space. Alternatively, the objective can be to maintain the classification behavior of the high-dimensional patterns $E'(\overline{\mathbf{Z}}_{n+1}) = \sum_{i=1}^N \|f_q(\mathbf{z}_i) - f_d(\mathbf{x}_i)\|^2$, where $f_d(\cdot)$ maps from the original data space to the label space. The stochastic embedding with Gaussian sampling is similar to unsupervised nearest neighbors [8], which consideres the data space reconstruction error of unsupervised regression [5].

Algorithm 1: incSE

Require: $\mathbf{X}$, $\mathbf{y}$, k, κ
1: $\overline{\mathbf{Z}} = [\mathbf{0}]$, $\overline{\mathbf{X}} = [\mathbf{x}_1]$
2: **for** $i = 2$ **to** N **do**
3: choose $\mathbf{x}_i$
4: select closest pattern $\mathbf{x}^*$ with latent position $\mathbf{z}^*$
5: **for** $l = 1$ **to** κ **do**
6: $\mathbf{z}_l^* \sim \sigma \cdot \mathcal{N}(\mathbf{z}^*, 1)$ with $\sigma = \|\mathbf{x}_i - \mathbf{x}^*\|^2$
7: **end for**
8: choose $\mathbf{z}_i = \arg\min_{\mathbf{z}=\mathbf{z}_1^*,\ldots,\mathbf{z}_\kappa^*} E(\overline{\mathbf{Z}}_{n+1})$
9: $\overline{\mathbf{Z}} = [\overline{\mathbf{Z}}, \mathbf{z}_i]$, $\overline{\mathbf{X}} = [\overline{\mathbf{X}}, \mathbf{x}_i]$
10: **end for**
11: **return** $\mathbf{Z}$

Fig. 1: Pseudo-code of incremental approach incSE

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

3 Re-embedding

As closely related alternative to incSE, we compare to a supervised re-embedding approach (reSE) that improves an existing embedding w.r.t. the low-dimensional classification error, see Figure 2. The re-embedding approach randomly selects one or more points and embeds them at a better latent position leading to a reduced error $E(\cdot)$. First, a complete embedding $\mathbf{Z}$ must be available. The initial $\mathbf{Z}$ can be generated randomly or can be a result of incSE and other embedding algorithms such as principal component analysis (PCA), locally linear embedding (LLE) [10] or isometric mapping (ISOMAP) [12]. A pattern $\mathbf{z}^*$ is randomly chosen from $\mathbf{Z}$. With Gaussian sampling, see Equation 2, the latent position of the pattern is randomly changed. The change is accepted, if the error $E(\mathbf{Z}')$ of the re-embedding manifold $\mathbf{Z}'$ is lower, i.e., if $E(\mathbf{Z}')$ with $\mathbf{Z}'$ containing $\mathbf{z}^* = \mathbf{z}'$. The algorithm proceeds until a termination condition is fulfilled.

Algorithm 2: reSE

Require: $\mathbf{X}$, $\mathbf{y}$, $\mathbf{Z}$
repeat
  Randomly choose $\mathbf{z}^*$ from $\mathbf{Z}$
  $\mathbf{z}' = \mathbf{z}^* + \mathcal{N}(0, \sigma)$
  if $E(\mathbf{Z}') < E(\mathbf{Z})$ **then**
    $\mathbf{z}^* = \mathbf{z}'$ for $\mathbf{Z}$
  end if
until termination condition
return $\mathbf{Z}$

Fig. 2: Pseudocode of re-embedding approach reSE

4 Related Work

A method for supervised dimensionality reduction is linear discriminant analysis (LDA) [1], which seeks for a transformation matrix such that the between-class scatter is maximized and the within-class scatter is minimized. Sugiyama [11] combines LDA with locality-preserving projection, which is well appropriate on multimodal data and requires the solution of a generalized eigenvalue problem. Related methods for embedding labeled data are discriminant adaptive nearest neighbor (DANN) [3], mixture discriminant analysis (MDA) [4], and neighborhood component analysis (NCA) [2]. MDA is based on a mixture discriminant analysis that extends LDA to maximum likelihood estimation of Gaussian mixture distributions. Related to incSE is unsupervised nearest neighbors (UNN) [6]. UNN is based on the unsupervised regression framework that maps from the low-dimensional space to data space. Nearest neighbor regression is used for this mapping. The variant that is able to map into latent spaces of arbitrary dimensionalities $1 \leq q < d$ is based on stochastic embeddings, similar to the Gaussian sampling of incSE. A kernel variant [7] extends the flexibility of UNN.

5 Experimental Analysis

In the following, we experimentally analyze and compare incSE and reSE on a benchmark data sets w.r.t. the kNN classification error, when mapping from latent space to label space. incSE is incrementally constructing the manifold,

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

while reSE starts from a random initialization and reSEiso from an initial embedding computed by ISOMAP. Each experiment is repeated 50 times and the mean and the corresponding standard deviation are shown in Table 1. Both reSE variants employ 10,000 re-embedding steps. On the MakeClass data set,

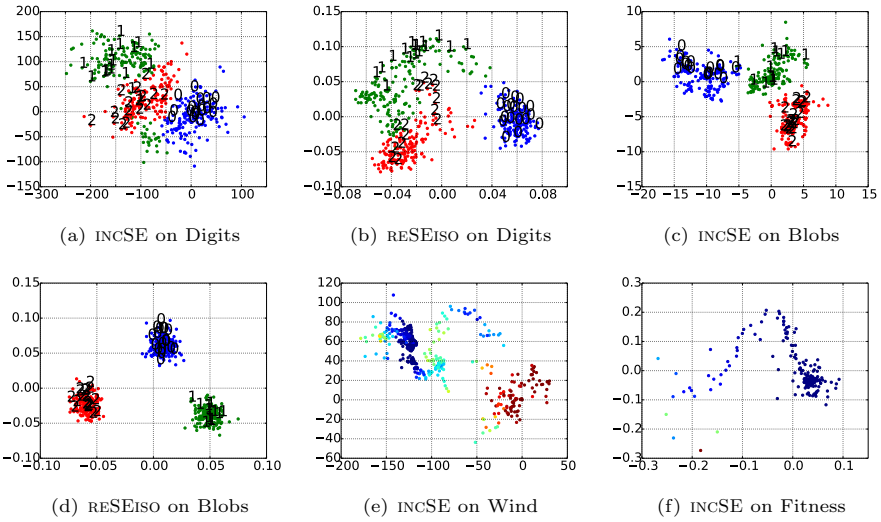
Table 1: Comparison of kNN classification error E when mapping from latent space to label space with incSE, reSE with random initialization, reSEiso with ISOMAP initialization, and LDA with $q = 2$ to native kNN in the original data space. Each stochastic experiment is repeated 50 times.

data	kNN	incSE	reSE	reSEiso	LDA
MakeClass	0.050	0.059 ± 0.012	0.088 ± 0.009	0.089 ± 0.008	0.121
Digits	0.012	0.102 ± 0.026	0.413 ± 0.011	0.136 ± 0.007	0.400
Faces	0.341	0.300 ± 0.012	0.301 ± 0.010	0.301 ± 0.010	0.259
Blobs	0.050	0.086 ± 0.028	0.085 ± 0.010	0.012 ± 0.002	0.045
Friedman 1	19.800	7.604 ± 1.022	4.022 ± 0.278	4.240 ± 0.454	7.437
Friedman 2	18.029	6.534 ± 0.672	3.715 ± 0.237	4.202 ± 0.389	6.647
Wind	3.503	3.460 ± 0.662	11.633 ± 0.824	1.729 ± 0.125	3.755
Housing	0.050	0.086 ± 0.028	0.085 ± 0.010	0.012 ± 0.002	0.032
Fitness	0.361	0.427 ± 0.029	0.278 ± 0.035	0.283 ± 0.012	0.599

incSE performs better than reSE and reSEiso. Here, reSE is even slightly better than reSEiso, but not statistically significant. On Digits, incSE again outperforms both reSE variants, but reSEiso is better than reSE. All three variants show approximately the same behavior on the Faces data set, while reSEiso outperforms reSE and incSE on the Blobs data set. The same situation can be observed on the Housing data set. In case of regression problem Friedman 1, it is interesting to observe that the achieved errors from the low-dimensional spaces are lower than the kNN error in data space. reSE achieves the best embedding, while reSEiso is competitive and incSE is significantly worse. This also holds for Friedman 2. A different behavior can be observed on the NREL wind data. While the method incSE achieves a slightly smaller error than the data space error, it is outperformed by reSEiso. But reSE fails with a large MSE. Both incSE variants show better results than incSE on the Fitness data set and achieve values that are better than the original regression error. The comparison to LDA shows that LDA is outperformed on MakeClass and Digits by all other methods, while showing superior results on the Faces data set. On the Blobs data set, LDA achieves better results than kNN and incSE, but worse than reSE.

In the following, the embeddings of incSE and reSE are compared visually. For the tests, we employ the data set Digits with $N = 500$ patterns and the Blobs data set based on generating Gaussian distributed patterns. For classification in latent space, we employ kNN with neighborhood size $k = 20$. Figure 3(a) shows the embedded patterns of Digits with incSE. Patterns from different classes are clearly separated and are neighboring to each other. An initialization with ISOMAP and post-processing with reSE also yields an accurate latent

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.



(a) INCSE on Digits (b) RESEISO on Digits (c) INCSE on Blobs

(d) RESEISO on Blobs (e) INCSE on Wind (f) INCSE on Fitness

Fig. 3: INCSE and RESE variants on the benchmark problems Digits, Blobs, Wind, and Fitness.

neighborhood with low kNN errors, see Figure 3(b) and Table 1.

Figure 3(c) shows the embedding of the Blobs data set with INCSE, Figure 3(d) shows the corresponding results of RESEISO. Similar to the results on Digits, INCSE computes smooth embeddings with reasonable neighborhoods between patterns of same classes. The high accuracies, see Table 1, confirm the results. The initialization with ISOMAP with post-processing via re-embedding well separates all three classes. The visualization of INCSE on the wind data, see Figure 3(e) shows that wind situations with similar labels (wind in the future) form neighborhoods that have a cluster-like shape. Figure 3(f) shows the result of INCSE embedding on the Fitness data set. The small path of points from left to right with the agglomeration of points in the right part of the figure illustrate the optimization process that converges towards the optimum. Such visualizations of optimization processes can help the practitioner to analyze the fitness landscape and the corresponding search process.

6 Conclusions

The supervised embedding method INCSE allows the embedding of high-dimensional patterns into low-dimensional latent spaces. Using the label information for the embeddings allows an efficient computation of the low-dimensional representations. Two algorithmic variants have been introduced. INCSE and RESEISO have been the most successful variants with respect to numerical results and visual inspection of exemplary embeddings. Due to the nearest neighbors search for each pattern, the complexity of INCSE is $O(N^2)$. The proposed

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

algorithms can be used for visualization and as preprocessing of classification methods. Various extensions are possible to use the low-dimensional representations for classification. One promising approach might be to determine the closest pattern $\mathbf{x}^*$ of $\mathbf{x}'$ in data space, and to employ kNN in latent space to determine the label of $\mathbf{x}'$. Another idea is to embed the pattern $\mathbf{x}'$ based on the data space reconstruction error, oriented to unsupervised regression [5].

References

[1] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Journal of Machine Learning Research*, 7(2):179–188, 1936.

[2] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17 (NIPS)*, pages 513–520, 2004.

[3] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(6):607–616, 1996.

[4] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society*, 58(1):155–176, 1996.

[5] S. Klanke and H. Ritter. Variants of unsupervised kernel regression: General cost functions. *Neurocomputing*, 70(7-9):1289–1303, 2007.

[6] O. Kramer. Dimensionality reduction by unsupervised k-nearest neighbor regression. In *International Conference on Machine Learning and Applications and Workshops (ICMLA)*, pages 275–278, 2011.

[7] O. Kramer. Unsupervised nearest neighbors with kernels. In *Advances in Artificial Intelligence - 35th Annual German Conference on AI (KI)*, pages 97–106, 2012.

[8] O. Kramer. Unsupervised nearest neighbor regression for dimensionality reduction. *Soft Computing*, page online first, 2014.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[11] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007.

[12] J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

A Benchmark Problems

MakeClass is a classification data set generated with the SCIKIT-LEARN [9] method `make_classification` with d dimensions and two centers. The UCI Digits data set comprises handwritten digits with $d = 64$. The Faces data set is called *Labeled Faces in the Wild* and has been introduced for studying face recognition problems, see `http://vis-www.cs.umass.edu/lfw/`. The Blobs data set is generated with the SCIKIT-LEARN [9] method `make_blobs` with two classes and a standard deviation of $\sigma = 10.0$. Friedman 1 and Friedman 2 are regression problems generated with SCIKIT-LEARN. The Wind data set is based on spatio-temporal time series data from the National Renewable Energy Laboratory (NREL) western wind data set comprising 32,043 wind turbines, each holding ten 3 MW turbines over a timespan of three years in a 10-minute resolution. The dimensionality is $d = 22$. The Housing data set, also known as California housing from the STATLIB repository comprises 20640 8-dimensional patterns and one label. Fitness is data set based on an optimization run of a (15+100)-evolution strategy on the Sphere function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$ with $d = 20$ dimensions and 21000 fitness function evaluations.