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/.

Unsupervised Dimensionality Reduction for Transfer Learning

Patrick Blöbaum, Alexander Schulz and Barbara Hammer

CITEC center of excellence, Bielefeld University, Germany

Abstract. We investigate the suitability of unsupervised dimensionality reduction (DR) for transfer learning in the context of different representations of the source and target domain. Essentially, unsupervised DR establishes a link of source and target domain by representing the data in a common latent space. We consider two settings: a linear DR of source and target data which establishes correspondences of the data and an according transfer, and its combination with a non-linear DR which allows to adapt to more complex data characterised by a global non-linear structure.

1 Introduction

A crucial property of every successful machine learning model is its generalisation ability from the known training data to novel settings, with statistical learning theory offering powerful mathematical tools for establishing formal guarantees for valid generalisation [15]. One core assumption underlying the classical setting is that of data being i.i.d.: the training scenario and future application areas are qualitatively the same, differences result from different sampling from the same distribution only. Transfer learning addresses the setting that source and target data are qualitatively different because they follow a different underlying distribution or they are even contained in different spaces [9].

Models which reliably follow a trend have become increasingly important in the context of big data, distributed systems, and life-long learning, as demonstrated e.g. by quite some recent successful approaches [10]. In this contribution, we will focus on the second problem of data being contained in different spaces. This setting occurs e.g. when the same objects are measured using sensors with different characteristics, a sensor is exchanged in a system (e.g. by a more sensitive one), the same objects are described in different languages, etc. One promise of such transfer consists in a plug-and-play technology for novel sensors or representations, without the need of costly retraining of the underlying models.

A few approaches have been proposed in this context, such as a common feature representation [11, 4], a coupled embedding of data in a low dimensional space [1, 8, 12], or a combination of representation learning and classification [7]. In this contribution, we are interested in the potential of modern unsupervised dimensionality reduction to induce a common representation of data for transfer learning. For this purpose, we will address two problems: How to linearly embed source and target in a common domain such that the resulting characteristics are shared as much as possible? We will rely on a probabilistic modelling of the target domain which induces an explicit embedding mapping via an EM approach. How to extend this framework to non-linear mappings by incorporating modern non-linear DR techniques which are better capable of capturing non-linear characteristics of the data? We will rely on t-SNE as a method which is particularly

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/.

suited to reliably capture cluster structures, and its recent extensions to kernel mappings to allow for an integration into the transfer pipeline [14, 5].

Unlike our approach, most manifold learners rely on explicit correspondences or equivalent information [16]. One rare exception is the approach [17], where local characteristics are directly extracted from the manifold to provide a local fingerprint. However, it is computationally exponential in the neighbourhood size and, further, it does not resolve ambiguities due to local self similarity.

2 Transfer Learning without given correspondences

We assume N source data $\mathbf{x}_i \in X$ and K target data $\mathbf{y}_j \in Y$ with different spaces X and Y but shared underlying information are present. We will model the fact that these two data sets share their structure by embedding both simultaneously in a low dimensional vector space Z where we assume a common distribution of the data sets. This will provide an explicit embedding $\mathbf{x}_i \mapsto \mathbf{z}_i^x \in Z$ of the source data and $\mathbf{y}_j \mapsto \mathbf{z}_j^y \in Z$ of the target data. The question how suitable embeddings can be found will be the subject of sections 2.1 and 2.2. The technical report [2] describes first ideas.

Provided such an embedding is present, knowledge transfer is immediate: Assume source labels $l(\mathbf{x}_i)$ are present. This enables us to learn a classifier on the embedding space based on the training data $(\mathbf{z}_i^x, l(\mathbf{x}_i))$. By means of the mapping $\mathbf{y}_j \mapsto \mathbf{z}_j^y$, this classifier can be directly extended to the target data.

2.1 Shared Linear Embedding

For simplicity, we first assume that data can be embedded linearly into a low dimensional space Z. For the mapping $\mathbf{x}_i \mapsto \mathbf{z}_i^x$ we can simply rely on a PCA embedding which captures the most relevant linear structure of the source data. Note that it is easily possible to exchange this embedding by any other suitable mapping such as LDA in case of auxiliary labels or a non-linear map as we will do in section 2.2. The target embedding should aim for a match with the source distribution in the latent space Z. We consider a parametrised linear mapping

$$f_W : Y \to Z, \mathbf{y} \mapsto \mathbf{z}^y = W\mathbf{y} \tag{1}$$

which induces a mixture of Gaussians $p(\mathbf{z}^x|Y, W) \sim \sum_j \exp\left(-\|\mathbf{z}^x - W\mathbf{y}_j\|^2/(2\sigma^2)\right)$ in the latent space. To enforce a shared distribution of source and target distribution in the latent space, we optimise the log likelihood $\sum_i \log p(\mathbf{z}_i^x|Y, W)$. Its optimisation can rely on an EM approach [3] with hidden variables

$$\gamma_{i,j} := \exp\left(-\|\mathbf{z}_i^x - W\mathbf{y}_j\|^2/(2\sigma^2)\right) / \sum_l \exp\left(-\|\mathbf{z}_i^x - W\mathbf{y}_l\|^2/(2\sigma^2)\right) \tag{2}$$

and a direct minimisation of the following term with respect to W:

$$\sum_{i,j} \gamma_{ij} \|\mathbf{z}_i^x - W\mathbf{y}_j\|^2. \tag{3}$$

It is often useful to apply a standard regularisation in this step. In order to initialise the mapping W, a PCA projection can be utilised. For the bandwidth σ, a deterministic annealing scheme can be employed [13].

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/.

2.2 Shared Non-linear Embedding

It has been emphasised in [14, 6] that linear DR does not allow a reliable characterisation of central data characteristics for many modern data sets; in such cases, a shared linear representation is clearly not sufficient to provide an informative shared representation of source and target domain. We are interested in how far modern non-linear dimensionality reduction techniques allow us to solve this problem. More specifically, we will rely on t-SNE as a particularly powerful embedding technique in case of clustered data [14]. Obviously, it is easily possible to exchange a PCA embedding $\mathbf{x}_i \rightarrow \mathbf{z}_i^x$ by any given non-linear embedding such as t-SNE for the source data. For the target domain, we aim for an explicit mapping and, therefore, rely on the recent extension of t-SNE to a kernel embedding [5]. The linear mapping (1) becomes

$$f_W : Y \rightarrow Z, \mathbf{y} \mapsto \mathbf{z}^y = \sum_j \mathbf{w}_j \cdot k(\mathbf{y}, \mathbf{y}_j) / \sum_l k(\mathbf{y}, \mathbf{y}_l) \tag{4}$$

with Gaussian kernel $k(\mathbf{y}, \mathbf{y}_j) = \exp(-\|\mathbf{y} - \mathbf{y}_j\|^2/\sigma_j^2)$ where σ_j is adjusted according to the effective neighbours of $\mathbf{y}_j$, and $\mathbf{w}_j \in Z$ comprises the parameters of W. For an initialisation, these parameters are adjusted such that $f_W(\mathbf{y}_i)$ approximates the t-SNE projection of the (target) data $\mathbf{y}_i$. The sum in equation (4) can either be over all points or over a subset, only. We use the latter, mainly for regularisation purposes (see [5] for more details). Since, in our setting, we aim for a match of the target and source distribution, we optimise the data likelihood by substituting the M step in equation (3) by the optimisation of

$$\sum_{i,j} \gamma_{ij} \|\mathbf{z}_i^x - f_W(\mathbf{y}_j)\|^2 + \lambda C_{\text{tSNE}} \left(\{\mathbf{y}_i\}_{i=1}^K, \{f_W(\mathbf{y}_i)\}_{i=1}^K \right), \tag{5}$$

with respect to parameters W. To better deal with the non-linearity, we add a regularisation term which enforces structure preservation of the target data during optimisation. We utilise the t-SNE cost function to measure the latter.

3 Experiments

In the following, we evaluate the linear and non-linear transfer learning techniques exemplarily with two data sets. For transfer learning, we will always assume that only the source data are accompanied by labels while this is not the case for the target data. This means, that we cannot use class information for the transfer learning. However, we will use this information in order to evaluate the quality of the transfer learning: In the embedding space we utilise the source data $(\mathbf{z}_i^x, l(\mathbf{x}_i))$ to train a linear Support Vector Machine (employing the one versus one scheme for data with more than two classes). Subsequently, we classify the projected target data $\mathbf{z}_i^y$ and compute the accuracy by comparing to the labels $l(\mathbf{y}_i)$. Note that the latter labels are not used for the transfer learning but for evaluation purposes, only. The classification accuracy for the embedded target data allows to judge in how far the transfer of information was successful.

We employ the following two benchmark data sets in our experiments.

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/.

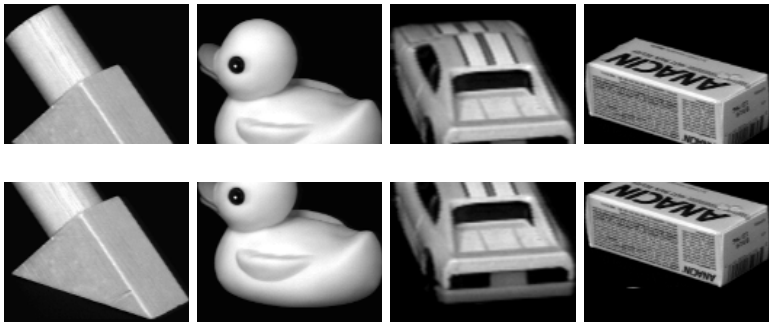

Fig. 1: Examples of images from the Coil data set: the top row contains images from the source data while the bottom row shows the according target images.

- The Iris data set utilises four features to describe three classes of iris plants. 150 instances are available in this data set.
- The Coil data set consists of images of objects that are rotated around their own axes. We employ four items from this data set in our experiments.

We create a transfer learning scenario for the Iris data set utilising the following scheme: We split the data set randomly into two parts, using one for the source and the other for the target data. The latter are additionally mapped with a random matrix to ten dimensions. Note that for this data set, the source and target data don't have any common instances.

For the Coil data set, we cut each image in order to obtain source and target data: We utilise the top 3/4 of each image for the source data and the lower 3/4 for the target data. Such, 1/2 of the information overlaps for both sets. One example object of each class is shown in Fig. 1.

Evaluation of TL with linear embeddings: We apply our approach to the Iris data set. We use a two-dimensional embedding space for the source data created by the PCA mapping. Fig. 2 shows an alignment result of the method. The left image depicts the source and target data in the embedding space while the middle and right image show the source and target data individually. This procedure was iterated ten times yielding the mean classification accuracy of 91% for the source data and of 83% for the target data.

We also apply our approach to the Coil data set. As previously, we iterate the procedure ten times yielding the accuracies 70% and 66% for the source and target domain (see Table 1 for an overview). An exemplary alignment is shown in Fig. 3 (top). The reason for the drop of the accuracy is visible here: Due to the linear mapping, the classes overlap and, hence, an accurate classification is not possible. This holds also for the classification of the source data.

Evaluation of TL with non-linear embeddings: In order to obtain a non-linear embedding we utilise the non-parametric method t-SNE. Applying the scheme from 2.2 yields the mean accuracy after ten runs of 92% for the source and 83% for target data. An exemplary alignment is shown in Fig. 3 (bottom).

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/.

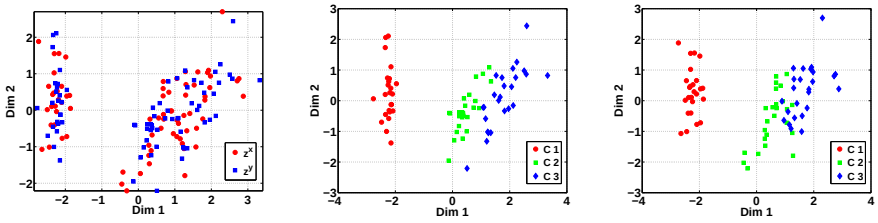

Fig. 2: The linear alignment of source and target data for the Iris data set is shown left. Both data sets are shown individually with their according labelling middle (source) and right (target).

In the middle figure the advantage of non-linear mappings is visible: The classes are well separated which allows a successful consequent transfer learning.

4 Discussion

We have introduced an approach to perform Transfer Learning via mapping source and target data into a common embedding space. For this purpose we have proposed the two possibilities to use linear and non-linear embeddings. The linear embeddings have proven to be very stable but they do not allow an accurate transfer if linear projections cannot embed the class structure adequately. Non-linear methods can improve the transfer of information in this case.

In this paper we have utilised only two dimensionality reduction techniques, i.e. PCA and t-SNE. Other approaches such as supervised methods or manifold embeddings could be particularly useful here and the investigation of their

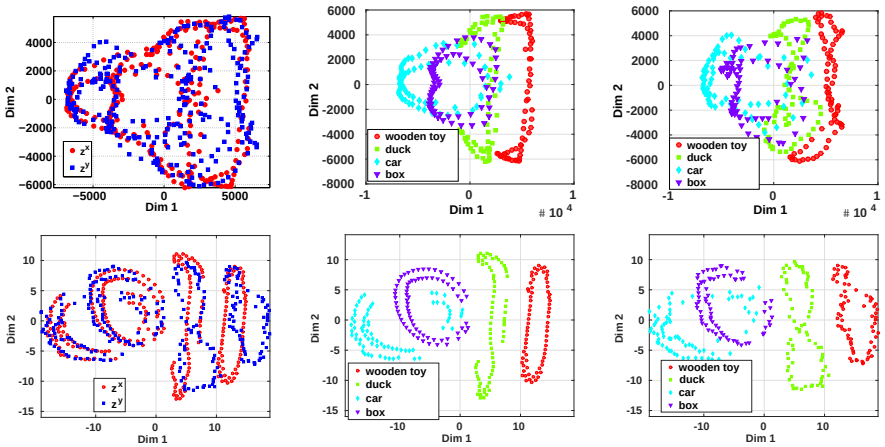

Fig. 3: A linear (top three) and non-linear (bottom three) alignment of source and target data for the Coil data set is shown left. Both data sets are shown individually with their according coloring middle (source) and right (target).

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/.

Table 1: Mean classification accuracies with a linear SVM for the experiments.

Embedding	linear		non-linear
Data sets	Iris	Coil	Coil
Error Source	91%	70%	92%
Error Target	83%	66%	83%

applicability is subject to future work.

Acknowledgements

Funding from DFG under grant numbers HA2719/7-1 and HA2719/6-1 and 6-2 and by the CITEC centre of excellence is gratefully acknowledged.

References

[1] J. Blitzer, S. Kakade, and D. P. Foster. Domain adaptation with coupled subspaces. In *AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 173–181, 2011.

[2] P. Blöbaum and A. Schulz. Transfer learning without given correspondences. In *New Challenges in Neural Computation*, pages 42–51, 2014.

[3] 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):1–38, 1977.

[4] G. Elidan, B. Packer, G. Heitz, and D. Koller. Convex point estimation using undirected bayesian transfer hierarchies. *CoRR*, abs/1206.3252, 2012.

[5] A. Gisbrecht, A. Schulz, and B. Hammer. Parametric nonlinear dimensionality reduction using kernel t-sne. *Neurocomputing*, 147:71–82, 2015.

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

[7] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu. Adaptation regularization: A general framework for transfer learning. *IEEE Trans. Knowl. Data Eng.*, 26(5):1076–1089, 2014.

[8] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[9] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, Oct. 2010.

[10] R. Polikar and C. Alippi. Guest editorial learning in nonstationary and evolving environments. *IEEE Trans. Neural Netw. Learning Syst.*, 25(1):9–11, 2014.

[11] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *(ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pages 759–766, 2007.

[12] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Trans. Knowl. Data Eng.*, 22(7):929–942, 2010.

[13] N. Ueda and R. Nakano. Deterministic annealing em algorithm. *Neural Netw.*, 11(2):271–282, Mar. 1998.

[14] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[15] M. Vidyasagar. *A Theory of Learning and Generalization*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2nd edition, 2002.

[16] C. Wang and S. Mahadevan. Manifold alignment using procrustes analysis. In *ICML '08*, pages 1120–1127, New York, NY, USA, 2008. ACM.

[17] C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *IJCAI'09*, pages 1273–1278, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.