Semi-supervised Dimensionality Reduction via Multimodal Matrix Factorization

Viviana Beltrán, Jorge A. Vanegas(✉), and Fabio A. González

Mindlab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
{lvbeltranb,javanegasr,fagonzalezo}@unal.edu.co

Abstract. This paper presents a matrix factorization method for dimensionality reduction, semi-supervised two-way multimodal online matrix factorization (STWOMF). This method performs a semantic embedding by finding a linear mapping to a low dimensional semantic space modeled by the original high dimensional feature representation and the label space. An important characteristic of the proposed algorithm is that the new representation can be learned in a semi-supervised fashion. So, annotated instances are used to maximize the discrimination between classes, but also, non-annotated instances can be exploited to estimate the intrinsic manifold structure of the data. Another important advantage of this algorithm is its online formulation that allows to deal with large-scale collections by keeping low computational requirements. According with the experimental evaluation, the proposed STWOMF in comparison with several linear supervised, unsupervised and semi-supervised dimensionality reduction methods, presents a competitive performance in classification while having a lower computational cost.

1 Introduction

Multimedia information presents many opportunities due to the richness of its high-dimensional information, but also implies many computational challenges mainly related with the well-known "curse of dimensionality" [3] that dramatically affects the speed of machine learning algorithms. Dimensionality reduction allows to eliminate the redundancy and the noise present in the manifold structure of the original high dimensional feature representation and tackles the curse of dimensionality by compressing the representation in a more expressive reduced set of variables that preserve the most important characteristics of the initial set. This is done by finding a transformation that does not alter the information presented by the initial data set. Dimensionality reduction is a technique widely used today in many machine learning tasks such as regression, annotation, classification, clustering, pattern recognition, information retrieval among others [1]. This technique would be used in unsupervised as well as supervised approaches. Unsupervised dimensionality reduction is mainly used with the aim of exploring the data structure and extracting meaningful information from data without any prior information. In contrast, in supervised dimensionality reduction specific targets (labeled instances) of interest are used to guide the process

A. Pardo and J. Kittler (Eds.): CIARP 2015, LNCS 9423, pp. 676–682, 2015.
DOI: 10.1007/978-3-319-25751-8_81

of dimensionality reduction. Even though supervised approaches can exploit the labeled data in order to improve classification performance, they require every training instance to be labeled. But a proper annotation of a whole dataset is an arduous process, and for large-scale real-world collections is infeasible to ensure a reliable annotation for each instance. So, in many cases we are in a situation where we have a big quantity of potential data for training our algorithms but only a small fraction with annotations can be used. Even so, non annotated data present valuable information about the manifold structure of the data that should be exploited in some way. This paper presents a semi-supervised dimensionality reduction method based on matrix factorization that can be used in training datasets that are not fully annotated by using the information from annotated instances to preserve the separability between elements from different classes, but also using the non-annotated elements to estimate the intrinsic manifold structure of the data.

The rest of this paper is organized as follows: Section 2, presents a comprehensive revision of related works in linear dimensionality reduction; in Section 3, details about of the proposed method are explained; Section 4, presents an evaluation of the proposed method in comparison with several state-of-the-art linear methods in dimensionality reduction; and finally, Section 5 presents some concluding remarks.

2 Related Work

There are a high number of linear techniques that perform dimensionality reduction by embedding the data to a lower semantic space, among the unsupervised approaches stand out principal component analysis (PCA) [10], factor analysis (FA) and independent component analysis (ICA) [13]. Other approaches like locality preserving projection (LPP) [11] and neighborhood preserving embedding (NPE) [9] try to preserve the local neighborhood structure. Some dimensionality reduction techniques can take into account domain knowledge. This domain knowledge can be expressed in different forms, such as, class labels, pairwise constraints or another kind of prior information. Fisher's linear discriminant analysis (LDA) [8] was one of the first techniques to take advantage of class observation to preserve the separability of the original classes. Also, there are semi-supervised alternatives that learn from a combination of both labeled and unlabeled data. For instance, semi-supervised discriminant analysis (SDA) [5] and the soft label based linear discriminant analysis SL-LDA [16] use the labeled data to maximize the separability between classes and uses the unlabeled data to estimate the intrinsic manifold structure of the data. Also, there are some non-linear alternatives (isometric feature mapping [14], locally linear embedding [12] and Laplacian Eigenmaps [2], among others). Unfortunately the modeling of these non-linearities leads to high computational complexities that make them prohibitive to use in large-scale collections. The method introduced in this paper, presents two characteristics that make it highly scalable: first, it is based on linear transformations, and second, its algorithm is formulated as an online-learning approach, which only needs to keep small portions of the training data in main memory and requires little time to reach a predefined expected risk.

3 Semi-supervised Two-Way Multimodal Online Matrix Factorization

We can represent an entire collection by a matrix $X \in \mathbb{R}^{n \times k}$, where k is the total number of instances in a training set and n is the number of features that represent each instance. In a similar way, we can represent the associated classes by a binary matrix $T \in \mathbb{R}^{m \times k}$, where m is the total number of classes in the collection, and a 1 in the $j-$th position $(1 \leq j \leq m)$ of the i-th column defines the membership of the i-th instance in the $j-$th class.This paper presents a semi-supervised dimensionality reduction framework based on TWOMF (Two-way Multimodal Online Matrix Factorization) [15], which simultaneously finds a mapping from the feature representation and from the class representation to an r-dimensional common semantic space, where $n \gg r$, and additionally, back-projection functions that reconstruct from this low r-dimensional space to the original feature and class representations are learned. These mappings are modeled for encoder and decoder matrices that perform linear transformations to and from the semantic space. So, the feature representation can be projected to the semantic space by an encoder matrix $W_x \in \mathbb{R}^{r \times n}$ and reconstructed back by a decoder matrix $W_x^{'} \in \mathbb{R}^{n \times r}$ such that $H \approx W_x X$ and $X \approx W_x^{'} H$. And, in a similar way, a reconstruction for the label representation is defined by $H \approx W_t T$ and $T \approx W_t^{'} H$, where, $W_t \in \mathbb{R}^{r \times m}$, and $W_t^{'} \in \mathbb{R}^{m \times r}$ are the encoder and decoder matrices for the label representation.Finally, a mapping between the original features and label representation, forcing an alignment of the semantic projections, is expressed by: $T \approx W_t^{'} W_x X$. All these previous conditions are put together and the problem is solved as an optimization problem by minimizing the following loss function:

$$\begin{aligned} L = \alpha \sum_{i=1}^{k} \left\| x_i - W_x^{'} W_x x_i \right\|_F^2 + (1-\alpha) \sum_{i=1}^{l} \left\| t_i - W_t^{'} W_t t_i \right\|_F^2 \\ + \delta \sum_{i=1}^{l} \left\| t_i - W_t^{'} W_x x_i \right\|_F^2 + \beta \left(\| W_v \|_F^2 + \left\| W_v^{'} \right\|_F^2 + \| W_t \|_F^2 + \left\| W_t^{'} \right\|_F^2 \right) \end{aligned} \quad (1)$$

where, x_i is the feature vector of the i-th instance in the data collection X and t_i is the corresponding binary label vector, α controls the relative importance between the reconstruction of the instance representation and the label representation, δ controls the relative importance of the mapping between instance features and label information and β controls the relative importance of the regularization terms, which penalize large values for the Frobenius norm of the transformation matrices. In this paper, we are interested in scenarios where we have a large number of instances for training (k instances), but only a restricted l number of them are properly labeled. The loss function (Eq. 1) takes advantage of both annotated and non-annotated instances. The first term in the loss function uses all the instances to model the low semantic space and the second and third terms use only the annotated instates to model the semantic space and the mapping between features and label information. The final algorithm uses stochastic gradient descent learning [4], by updating the transformation matrices at each iteration with a mini-batch of instances with their corresponding features and label representation that are randomly sampled from the training set, due to the fact that samples in a minibatch are discarded after the minibatch is processed, it is possible to scan large datasets without memory restrictions.The algorithm ends when a predefined maximum number of epochs is reached. Once the learning process is completed, the projection to the low-rank semantic representation can be performed by multiplying the original high-dimensional feature representation by the coding W_x matrix ($h_i = W_x x_i$).

4 Experiments and Results

In this section, we evaluate our algorithm in comparison with several widely-used datasets for dimensionality reduction, manifold learning and classification tasks (the details of each dataset are shown in Table 1). We evaluate the performance of our algorithm by calculating classification accuracy in each one of these datasets. We compare our method with other linear supervised, semi-supervised and unsupervised dimensionality reduction methods. These methods include SVM (Support Vector Machines) with a linear kernel [7], LDA [8], SRDA (spectral regression discriminant analysis) [6], SDA [5] and PCA [10]. For determining the parameters of each method, we perform an exploration by using 5-fold cross-validation. For our method, we need to determine five parameters, including, the learning rate, the mini-batch size and the α, β and δ parameters present in the cost function.

Table 1. Dataset information and data partition for each dataset

Dataset	Original dataset partitions		Low-scale partitions		Large-scale evaluation		#Dim	#Class
	Train	Test	Train	Test	Train	Test		
Covtype	581012		8000	8000	100000	2000	54	7
MNIST	60000	10000	8000	8000	60000	10000	784	10
Letters	20000		8000	8000	–		16	26
USPS	4649	4649	4649	4649	–		256	10

For all algorithms, except for the supervised, i.e, SVM, LDA and SRDA, we use the projected training set to construct a nearest neighborhood classifier (1NN) for evaluating the classification accuracy of the projected test set, in a similar setup as in [16]. In this evaluation, we explore the performance for different percentage of randomly selected annotated instances in training set. Table 2 reports the average accuracies for 10 runs in each configuration in the four datasets using the low-scale partitions (see Table 1). As we can see, the

STWOMF presents competitive results in comparison with all other algorithms when the dimensionality of the semantic representation coincides with the number of classes (r=C). Furthermore, when the dimensionality increases (r=C+10), STWOMF over performs the other algorithms (in our experiments, a further increase of the dimensionality did not contribute to improve the performance of the algorithm).

An evaluation with the two largest datasets using different sizes of training set was performed in order to verify the capability of the proposed method to deal with large-scale collections. Figure 1 presents the average classification accuracies and times for different sizes of the training set (the reported results are the average of 10 runs for each configuration). The STWOMF is compared against the SDA which is another semi-supervised method that also uses the unlabeled data to estimate the manifold structure of the data. For all training sizes only 30% of instances are annotated, so we can see that both methods are able to learn from labeled and unlabeled instances and both can improve their performance as more training instances are available. However, STWOMF presents two advantages: first, unlike SDA, in STWOMF we can increase the dimensionality of the semantic space resulting in an improvement in the performance. For instance, in the MNIST dataset, the STWOMF using 17 latent factor (STWOMF-r17) presents a gain in accuracy of about 6 points over the same STWOMF using only 7 latent factor (STWOMF-r7) and the SDA; and second, STWOMF presents a little increase in the time required for training as more training instances are used, leading to a speedup of about 3.5x-7x over SDA in MNIST and about 8x in CovType. The main reason for the short time used in training phase by STWOMF is that, thanks to its online formulation for large datasets, a few number of epochs are required until the algorithm converges (convergence in all algorithms is verified by means of a minimum threshold required to improve the reconstruction error in each epoch). In fact, for both datasets MNIST and CovType only two epochs are required to achieve convergence.

Table 2. Classification accuracy for different percentages of annotated instances in training set using low-scale partitions. Reported results are the average of 10 runs for each configuration (r = number of latent factors, C = number of classes in the dataset).

METHOD		STWOMF	STWOMF	SDA	LDA	SVM	SRDA	PCA	PCA
		r=C	r=C+10					r=C	r=C+10
COVTYPE	100%	0.725 1.0e-2	0.770 1.0e-2	0.735 0.0	0.708 3.5e-3	0.674 3.3e-16	0.698 3.3e-16	0.707 3.3e-16	0.763 3.3e-16
	60%	0.720 1.9e-2	0.755 1.0e-2	0.719 3.3e-16	0.704 7.6e-3	0.679 3.3e-16	0.685 3.3e-16	0.683 3.3e-16	0.724 0.0
	30%	0.686 1.7e-2	0.712 1.0e-2	0.687 3.3e-16	0.707 7.6e-3	0.667 3.3e-16	0.653 0.0	0.639 3.3e-16	0.679 0.0
MNIST	100%	0.882 0.0	0.939 0.0	0.870 0.0	0.897 0.0	0.839 0.0	0.856 0.0	0.874 0.0	0.938 0.0
	60%	0.864 0.0	0.930 0.0	0.870 0.0	0.890 0.0	0.817 0.0	0.833 0.0	0.863 0.0	0.929 0.0
	30%	0.848 0.0	0.916 0.0	0.850 0.0	0.881 0.0	0.780 0.0	0.786 0.0	0.842 0.0	0.910 0.0
LETTERS	100%	0.946 1.5e-2	0.946 1.6e-3	0.950 3.3e-16	0.699 0.0	0.701 3.3e-16	0.936 0.0	0.940 0.0	0.940 0.0
	60%	0.933 1.9e-3	0.923 0.0	0.940 3.0e-4	0.694 3.3e-16	0.699 0.0	0.919 3.3e-16	0.913 3.8e-3	0.914 0.0
	30%	0.905 3.5e-3	0.885 6.1e-3	0.917 4.4e-4	0.680 3.3e-016	0.697 3.3e-16	0.893 0.0	0.872 2.5e-3	0.872 3.1e-3
USPS	100%	0.936 9.2e-4	0.966 3.3e-3	0.925 6.7e-4	0.943 3.3e-16	0.914 6.6e-16	0.921 6.6e-16	0.930 0.0	0.963 0.0
	60%	0.927 3.4e-3	0.957 1.0e-3	0.917 0.0	0.939 0.0	0.901 0.0	0.906 6.6e-16	0.921 6.6e-16	0.953 0.0
	30%	0.910 4.9e-3	0.942 2.4e-3	0.903 3.3e-16	0.926 6.6e-16	0.883 3.3e-16	0.884 0.0	0.903 3.3e-16	0.938 3.3e-16

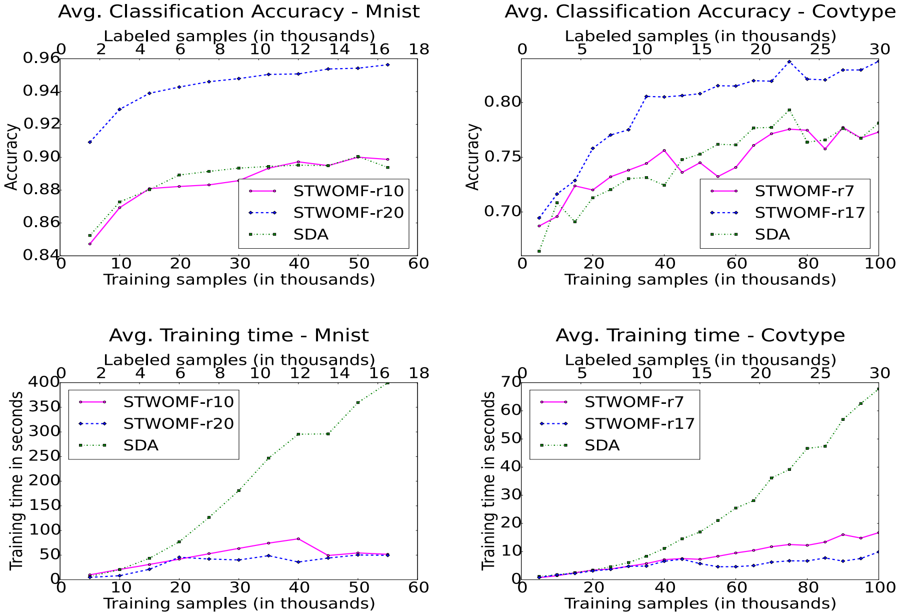

Fig. 1. Average classification accuracy (top) and average required time for training (bottom) in MNIST (left) and CovType (right) datasets using different number of training instances. For all training sizes only 30% of instances are annotated.

5 Conclusions

We presented an approach for dimensionality reduction that takes advantage of annotated data to model a semantic low-space representation that preserves the separability of the original classes. Furthermore, this method has the ability to exploit unlabeled instances for modeling the manifold structure of the data and use it to improve its performance in classification. The experimental evaluation shows that the proposed method presents competitive results in terms of classification accuracy in comparison with several unsupervised, semi-supervised and supervised linear dimensionality reduction methods, but with the advantage of its online learning formulation that allows it to deal with large collections of data by achieving a significantly reduction in computational requirements, in terms of memory consumption and required time for training.

Acknowledgments. This work was partially funded by projects *Diseño e implementación de un sistema de cómputo sobre recursos heterogéneos para la identificación de estructuras atmosféricas en predicción climatológica* number 1225-569-34920 through Colciencias contract number 0213-2013 and Multimodal Image Retrieval to Support Medical CaseBased Scientific Literature Search, ID R1212LAC006 by Microsoft Research/LACCIR. Jorge A. Vanegas thanks for doctoral grant supports Colciencias 617/2013 and Viviana Beltrán also thanks to Colciencias grant 645 *Jóvenes Investigadores e Innovadores de Colciencias* 2014.

References

1. Aggarwal, C.C., Zhai, C.X.: Mining text data. Springer Science & Business Media (2012)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: NIPS, vol. 14, pp. 585–591 (2001)
3. Bellman, R., Kalaba, R.: On adaptive control processes. IRE Transactions on Automatic Control **4**(2), 1–9 (1959)
4. Bottou, L., LeCun, Y.: Large scale online learning. In: Advances in Neural Information Processing Systems 16, NIPS 2003 (2003)
5. Cai, D., He, X., Han, J.: Semi-supervised discriminant analysis. In: IEEE 11th ICCV 2007, pp. 1–7. IEEE (2007)
6. Cai, D., He, X., Han, J.: Srda: An efficient algorithm for large-scale discriminant analysis. IEEE TKDE **20**(1), 1–12 (2008)
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
8. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. John Wiley & Sons (2012)
9. He, X., Cai, D., Yan, S., Zhang, H.-J.: Neighborhood preserving embedding. In: CICCV 2005, vol. 2, pp. 1208–1213. IEEE (2005)
10. Jolliffe, I.: Principal component analysis. Wiley Online Library (2002)
11. Niyogi, X.: Locality preserving projections. In: Neural Information Processing Systems, vol. 16, p. 153. MIT (2004)
12. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
13. Stone, J.V.: Independent component analysis: an introduction. Trends in Cognitive Sciences **6**(2), 59–64 (2002)
14. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000)
15. Vanegas, J.A., Beltran, V., González, F.A.: Two-way multimodal online matrix factorization for multi-label annotation. In: ICPRAM, pp. 279–285, January 2015
16. Zhao, M., Zhang, Z., Chow, T.W., Li, B.: A general soft label based linear discriminant analysis for semi-supervised dimensionality reduction. Neural Networks **55**, 83–97 (2014)