Semi-supervised Regularized Coplanar Discriminant Analysis⋆

Rakesh Kumar Sanodiya*[1], Michelle Davies Thalakottur[3], Jimson Mathew[1], and Matloob Khushi[2]

Indian Institute of Technology Patna, India[1]
School of Computer Science, The University of Sydney, Sydney, Australia[2]
MKSSS Cummins College of Engineering, Pune, India[3]
rakesh.pcs16@iitp.ac.in

Abstract. Dimensionality Reduction is a widely used method of removing redundant features and data compression. Dimensionality reduction usually occurs in a supervised setting in which all the samples are labelled. However, aspects of spectral clustering and semi-supervised learning can be used in Dimensionality reduction to ensure minimum loss of important data while projecting the high-dimensional data into lower dimensions. We have proposed a novel framework called Semi-supervised Regularized Co-planar Discriminant Analysis (SRCDA) that creates a graph of labelled and unlabelled data and uses label propagation to predict the classes of the unlabelled data. Additionally, we introduce a regularized term which is used to prevent overfitting. The proposed algorithm is evaluated against several other state-of-the-art algorithms with benchmark datasets including PIE Face, ORL and Yale Dataset. The proposed algorithm shows higher accuracies compared to the other algorithms and can be used in real-life datasets where the unlabelled data is vastly greater than the labelled samples. We have also conducted a statistical significance test to verify the results obtained.

Keywords: Semi-supervised Learning · Dimensionality Reduction · Regularization

1 Introduction

Dimensionality Reduction (DR) is the cornerstone when it comes to analyzing high dimensional data. In the real-world where high dimensional data is being generated at ever increasing rates, the usefulness on Dimensionality Reduction techniques cannot be underestimated. They allow data in high dimensions to be projected to a lower dimensional subspace while attempting to preserve some feature or structure of the data. Dimensionality Reduction can be used for data visualization and data compression and can improve the performance of a model by taking care of multi-collinearity by removing redundant features and make

⋆

it more efficient by reducing the computing time. Often, many features are redundant or are correlated instead of being independent. Hence, ultimately there is only a small subspace of the original representation that has features relevant to the task at hand. In such a case, Dimensionality Reduction enables low dimensional representation while ensuring minimum information loss.

The use of entirely labelled information can be called supervised learning while if only unlabelled information is used to train the model, it is called unsupervised learning [1]. The use of large amounts of unlabelled information in conjunction with few labelled points is referred to as semi-supervised learning [2]. Supervised learning holds the disadvantage that supervised models require large amounts of labelled information that is time-consuming and expensive to annotate. Additionally, unlabelled information is more readily available in the real-world. Dimensionality Reduction has traditionally been a unsupervised learning technique, but can hold the disadvantage that label information is not used at all, when the labels can also provide useful information that can be propagated to the unlabelled data via the geometry of the data distribution.

In this paper, we have proposed Semi-supervised Regularized Coplanar Discriminant Analysis (SRCDA) which adds a regularization term to the original criteria of RCDA which is based on the prior knowledge of the distribution geometry provided by both the labeled and unlabeled data. This is done by constructing a graph using similarity between embeddings and finding its Laplacian. Our proposed algorithm is shown to perform better than most state-of-the-art Dimensionality Reduction algorithms when evaluated on various real-world datasets.

The major contributions of the current work are that we have utilized both labelled and unlabelled information instead of relying on one over another. A graph is constructed which allows for label information propagation which helps in classifying the unlabelled information more accurately. Additionally, we have used learnt knowledge of the manifold in which the data exists to add a regularization term that prevents overfitting of the model to the few labelled samples as well as giving better accuracies.

The rest of the paper is organized as follows: Section 2 explores the relevant related work after going through various literature surveys. In Section 3 we provide a detailed overview of RCDA and the specifications of our proposed method. Experimental Results are discussed in Section 4 and we conclude and provide suggestions for future work in Section 5.

2 Related Work

We shall first discuss some relevant Dimensionality Reduction algorithms postulated from literature surveys [3]. Principal Component Analysis (PCA) was developed by Pearson [4] in 1901. PCA transforms the input data by projecting it to a new coordinate system such that the variance along the principal components is maximized. PCA has been generalized for non-linear data in Kernel PCA [5] by using a kernel function to project the lower-dimensional data to higher dimensions where it is easier to work with it. However, PCA is not without its faults. It fails to capture the intrinsic geometry of the manifold in which the data exists while Kernel PCA can suffer from the curse of dimensionality with the model requiring more data samples.

Linear Discriminant Analysis (LDA) [6] tries to find a linear transformation such that the class structure and differences of the data distribution remain conserved. To this end, it calculates a within-class scatter matrix, which it tries to minimize, and a between-class scatter matrix, which it tries to maximize, in order to find the projection matrix for the data. LDA has a few constraints which include that each data point must be labeled with a single class and the class boundaries are taken to be linear. Hence, LDA does not allow for outliers that do not belong to any class or samples that belong to more than one class.

Locally Linear Embedding (LLE) [7] approaches dimensionality reduction by considering neighbour information instead of the global manifold. It computes a lower dimensional embedding through weights learnt from its neighbours that allow the data point to be reconstructed. While LLE might be favoured by researchers due to its non-iterative approach to finding low-dimensional embeddings, it is sensitive to noise and is unable to deal with outliers.

Like LLE, Locality Preserving Projections (LPP) [8] also considers local neighbour information to calculate the Laplacian of a graph whose edges are assigned weights in order to preserve the local geometry. The projection will then use these weights to preserve the local manifold structure. While LPP is very good at capturing the geometry of the neighbourhood, it is not able to capture variable relationships which can then be lost after projection. Related to LPP, Globality-locality Preserving Projections (GLPP) [9] tries to preserve the sample relationships as well as the local manifold structure by computing a Laplacian on the subject invariant part of the samples and another Lapalacian on the intra-subject part of the sample. It then jointly learns a common Laplacian which is more robust to noise and is able to preserve the geometry better.

Regularized Coplanar Discriminant Analysis (RCDA) [10] uses coplanarity of samples to preserve class information while projecting the data to lower dimensions. It is discussed in more detail in Section 3.1.

Semi-supervised Discriminant Analysis (SDA) [11] is an extension of LDA which uses a graph Laplacian to learn the structure of the data distribution in order to introduce some distribution appropriate regularization term that helps the model to predict labels more accurately.

3 The Proposed Method

3.1 An Overview of RCDA

RCDA is a Dimensionality Reduction algorithm that calculates a projection matrix such that when the high dimension data is projected to lower dimensions, the between-class variance is preserved in the new manifold while the within-class variance is reduced [10]. Thus, the class information is preserved while projecting to a lower dimension.

Linear projection directions are found such that the coplanarity of samples from the same class are maximized while samples from different classes are made noncoplanar. For this, a within-class coplanar compactness and between-class coplanar separability is calculated after projection of the high dimensional feature matrix to lower dimensions. The within-class coplanar compactness is used to measure the error in the within-class linear representation, in order to minimize it, while the between-class coplanar separability is used to measure the error in the between-class linear representation in order to maximize it. Firstly, mean normalization is performed on the projected lower dimension data after which the within-class coplanar compactness is defined as the sum of errors of the within-class linear representation. It is represented by the following term:

$$\begin{aligned}\sum_{i=1}^{n}(\|W^T x_i - W^T X_i \beta_i^W\|_2^2 + \lambda\|\beta_i^W - \tilde{\beta}_i^W\|_2^2) \\ = \mathrm{Tr}(W^T S_W W) + \lambda\|B_w - \tilde{B}_W\|_F^2\end{aligned} \tag{1}$$

Similarly, after mean-normalization, the within-class coplanar compactness is defined by the sum of the errors of within-class linear representation.

$$\begin{aligned}\sum_{i=1}^{n}\sum_{c=1}^{C}(\|W^T x_i - W^T X_c \beta_{i,c}^b\|_2^2 + \lambda\|\beta_{i,c}^b - \tilde{\beta}_i^b\|_2^2) \\ = \mathrm{Tr}(W^T S_b W) + \lambda\sum_{c=1}^{C}\|B_c - \tilde{B}_W\|_F^2\end{aligned} \tag{2}$$

Using the within-class coplanar compactness and the between-class coplanar separability terms, RCDA defines its optimization function as follows,

$$\min_W \frac{\min_{\beta^W}\mathrm{Tr}(W^T S_W W) + \lambda\|B_W - \tilde{B}_W\|_F^2}{\min_{\beta^b}\mathrm{Tr}(W^T S_b W) + \lambda\sum_{c=1}^{C}\|B_c - \tilde{B}_c\|_F^2} \tag{3}$$

A linear projection matrix W, the within-class linear representation coefficient β^W and the between-class linear representation coefficient β^b is learnt simultaneously.

3.2 Semi-Supervised Setting

We have considered the case of RCDA in a semi-supervised setting and have modified the objective function so that the algorithm performs better for datasets with large amounts of unlabelled information and a few labelled data points. We can rewrite Equation 3 and introduce a Regularization term $J(a)$ to form our objective function. Here, a is the eigen vector corresponding to non-zero eigenvalue of the optimization problem in Equation 4. Due to the lack of this regularizer term, RCDA suffers from over-fitting when there is insufficient training examples. Based on the a priori information of the manifold structure, we can determine the value of the term regularization, thereby merging the manifold structure to improve accuracy.

$$\min_{W} \frac{\min_{\beta^W} Tr(W^T S_W W) + \lambda \|B_W - \tilde{B}_W\|_F^2 + J(a)}{\min_{\beta^b} Tr(W^T S_b W) + \lambda \sum_{c=1}^{C} \|B_c - \tilde{B}_c\|_F^2} \tag{4}$$

We use graph based semi-supervised learning, in which labeled data instances information can be propagated to unlabeled instances by using the similarity matrix. In another word, we can say that similar data instances in low dimensional manifold will have same class label because of their similar embedding. Therefore, the relationship between the instances in the given data-set can be represented as a weight graph or matrix G. For example, the relationship between data instances x_i and x_j is represented by an edge if they are the similar else there is no edge between them. Thus, the weight matrix G can be computed as follows,

$$G_{ij} = \begin{cases} 1, & \text{if } x_i \in N_k(x_j) \mid x_j \in N_k(x_i) \\ 0, & \text{otherwise} \end{cases}$$

where, $N_k(x_j)$ is the set of k nearest neighbours of x_j. From G, Laplacian matrix Matrix L can be calculated as $L = D - G$ where, D is the diagonal matrix whose entries are the column sums of G. Thus we can define a regularizer as follows,

$$\begin{aligned} J(a) &= \sum_{ij} (a^T x_i - a^T x_j)^2 S_{ij} = 2 \sum_i a^T x_i D_{ii} x_i^T a - 2 \sum_{ij} a^T x_i G_{ij} x_j^T a \\ &= 2a^T X(D-G)X^T a = 2a^T XLX^T a \end{aligned} \tag{5}$$

Thus, the Objective function for our proposed model can be written as,

$$\min_{W} \frac{\min_{\beta^W} Tr(W^T S_W W) + \lambda \|B_W - \tilde{B}_W\|_F^2 + 2a^T XLX^T a}{\min_{\beta^b} Tr(W^T S_b W) + \lambda \sum_{c=1}^{C} \|B_c - \tilde{B}_c\|_F^2} \tag{6}$$

This optimization problem can be solved by the maximum eigenvalue solution to the generalized eigenvalue problem.

4 Experimental Results

4.1 Experimental Setup and Benchmark Datasets

We have evaluated our proposed algorithm, SRCDA, against 4 other algorithms including, RCDA [10], LDA [6], PCA [4] and a NN (Nearest Neighbour) on the Cambridge ORL Database of Faces [12], Yale Face Database [13], PIE Face

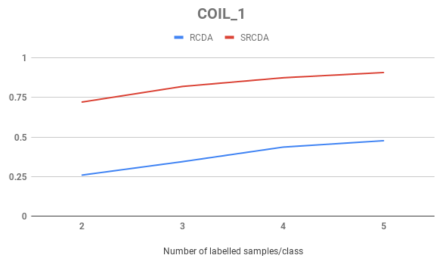
Dataset [14] and the COIL Dataset [15]. For our experiments, we have chosen a varying number (1-5) of samples per class as our labelled dataset and taken the rest of dataset, after discarding their labels as the unlabelled dataset. We have then performed PCA on the dataset to reduce its dimensionality. Datasets like the PIE Face Dataset have 1024 features per sample and performing the experiment on such a large feature matrix would be computationally inefficient. Hence we project the data onto a lower dimension (100) and then evaluate the algorithm on the projected matrix. We use a kNN classifier evaluate the performance of the algorithm. We run the experiment 30 times and compute the average result which has been shown in given in Table 1.

The datasets used in evaluating the algorithm have been discussed below:

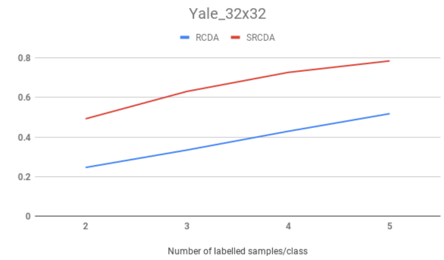
- Cambridge ORL [12]: This database, compiled by AT&T Laboratories Cambridge, consists of 40 subjects against a dark homogeneous background in an upright, frontal position. Each subject has 10 images each with varying lighting, facial expressions and facial details.
- Yale [13]: This database has 10 subjects with images captured with 9 poses, under 64 illumination conditions and 1 image of the subject with an ambient illumination for each pose. There are a total of 5850 images in the database. The size of each image is 640x480.
- PIE Face [14]: The CMU Pose, Illumination, and Expression (PIE) database has 68 subjects captured across 13 poses, under 43 illumination conditions and with 4 different expressions. It has over 40,000 images. The images are all high quality images captured using 13 Sony DXC 9000 camera's.
- COIL [15]: Columbia Object Image Library (COIL-20) is a database of grayscale images of 20 subjects with 72 images each. Images of the subject were taken at pose intervals of 5°over 360°, against a black background. The database has two sets of images, the first of which contains 720 unprocessed images of 10 subjects while the second set contains 1,440 size normalized images of 20 objects.

4.2 Results and Discussions

In Table 1 we have reported the values obtained in our experiments and have plotted the values of a few of the datasets in Fig. 1. In the table the highest accuracies have been shown in bold text. We see that SRCDA shows a marked improvement over RCDA and this trend is seen in every dataset. This is especially true when the number of labelled samples per class is small. As the number of labelled samples per class increases it is seen that the performance of RCDA becomes comparable to that of SRCDA in some datasets. If the dataset has a large number of samples per class, then the performance may not be as high as that as SRCDA until the number of labelled samples per class is equal to the total number of labelled samples per class, i.e., the problem is converted into a supervised learning problem. Hence, SRCDA would be able to generalize well on any type of dataset and perform well.



(a) COIL dataset



(b) Yale dataset

Fig. 1: Comparison of performance of SRCDA vs RCDA on COIL and Yale dataset

Table 1: Comparison of SRCDA with various state-of-the-art algorithms

ORL Dataset, Task: ORL_32x32						**Yale Dataset, Task: Yale_32x32**					
Algorithms	**1**	**2**	**3**	**4**	**5**	**Algorithms**	**1**	**2**	**3**	**4**	**5**
RCDA	0.1417	0.3067	0.8325	0.9327	0.9579	RCDA	0.1352	0.2461	0.3343	0.4295	0.5184
LDA	N/A	0.4439	0.7385	0.9384	**0.9707**	LDA	N/A	0.2489	0.3389	0.4303	0.5269
PCA	0.6036	0.7640	0.8523	0.9018	0.9407	PCA	**0.4189**	0.5541	0.6440	0.7147	0.7699
NN	0.5949	0.7548	0.8492	0.9058	0.9400	NN	0.4145	**0.5569**	**0.6503**	0.7145	0.7683
SRCDA	**0.6123**	**0.8218**	**0.9093**	**0.9458**	0.9632	**SRCDA**	0.3129	0.4925	0.6311	**0.7274**	**0.7855**
PIE Face Dataset, Task: PIE05						**PIE Face Dataset, Task: PIE07**					
Algorithms	**1**	**2**	**3**	**4**	**5**	**Algorithms**	**1**	**2**	**3**	**4**	**5**
RCDA	0.0789	0.6992	0.7368	0.7066	0.6969	RCDA	0.0997	0.7281	0.8196	0.8286	0.8235
LDA	N/A	0.5404	0.7795	0.8350	**0.8727**	LDA	N/A	0.6176	0.8211	0.8782	**0.9067**
PCA	0.1559	0.2550	0.3359	0.4088	0.4881	PCA	0.1835	0.3037	0.4046	0.4854	0.5568
NN	0.1654	0.2705	0.3593	0.4329	0.4947	NN	0.1858	0.3062	0.4047	0.4927	0.5656
SRCDA	**0.5327**	**0.7289**	**0.8021**	**0.8393**	0.8685	**SRCDA**	**0.5452**	**0.7396**	**0.8373**	**0.8812**	0.9044
PIE Face Dataset, Task: PIE09						**PIE Face Dataset, Task: PIE27**					
Algorithms	**1**	**2**	**3**	**4**	**5**	**Algorithms**	**1**	**2**	**3**	**4**	**5**
RCDA	0.0984	0.7285	0.8053	0.8189	0.8292	RCDA	0.0762	0.7065	0.7644	0.7571	0.7525
LDA	N/A	0.6043	0.8181	0.8701	**0.9054**	LDA	N/A	0.5452	0.7844	0.8389	**0.8751**
PCA	0.1925	0.3176	0.4204	0.5096	0.5826	PCA	0.1559	0.2550	0.3359	0.4088	0.4678
NN	0.1938	0.3234	0.4276	0.5154	0.5963	NN	0.1579	0.2576	0.34289	0.4161	0.4778
SRCDA	**0.5482**	**0.7466**	**0.8294**	**0.8767**	0.9044	**SRCDA**	**0.5435**	**0.737**	**0.7954**	**0.8341**	0.863
PIE Face Dataset, Task: PIE29						**COIL Dataset, Task: COIL_1**					
Algorithms	**1**	**2**	**3**	**4**	**5**	**Algorithms**	**1**	**2**	**3**	**4**	**5**
RCDA	0.1117	0.7205	0.8136	0.8246	0.8263	RCDA	0.0797	0.2591	0.3443	0.4365	0.4772
LDA	N/A	0.6056	**0.8148**	**0.8683**	**0.9033**	LDA	N/A	0.3823	0.4560	0.5355	0.6492
PCA	0.2058	0.3337	0.4357	0.5257	0.5918	PCA	**0.6903**	**0.7858**	**0.8414**	**0.8802**	0.9019
NN	0.2117	0.3358	0.4434	0.5312	0.6019	NN	0.2986	0.3631	0.3946	0.4215	0.8999
SRCDA	**0.5260**	**0.7229**	0.8121	0.8659	0.8946	**SRCDA**	0.6155	0.7211	0.8205	0.8753	**0.9085**

5 Conclusion

In this paper we have proposed a novel approach to Semi-supervised Dimensionality Reduction and have evaluated the proposed algorithm against several other state-of-the-art algorithms with various datasets. We have used label propagation with the help of a graph to better predict the labels of unlabelled data. Additionally we have added a regularization term that helps incorporate the manifold structure in the objective function. In future we aim to implement the method on medical data[16,17].

References

1. Bishop, C.M.: Pattern recognition and machine learning. springer (2006) 2
2. Zhu, X.J.: Semi-supervised learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2005) 2
3. Cunningham, J.P., Ghahramani, Z.: Linear dimensionality reduction: Survey, insights, and generalizations. The Journal of Machine Learning Research **16**(1), 2859–2900 (2015) 2
4. F.R.S., K.P.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901) 2, 5
5. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis. In: Gerstner, W., Germond, A., Hasler, M., Nicoud, J.D. (eds.) Artificial Neural Networks — ICANN'97. pp. 583–588. Springer Berlin Heidelberg, Berlin, Heidelberg (1997) 2
6. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.R.: Fisher discriminant analysis with kernels. In: Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468). pp. 41–48. Ieee (1999) 3, 5
7. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding **290**(5500), 2323–2326 (2000). https://doi.org/10.1126/science.290.5500.2323 3
8. He, X., Niyogi, P.: Locality preserving projections. In: Advances in neural information processing systems. pp. 153–160 (2004) 3
9. Huang, S., Elgammal, A., Huangfu, L., Yang, D., Zhang, X.: Globality-locality preserving projections for biometric data dimensionality reduction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2014) 3
10. Huang, K.K., Dai, D.Q., Ren, C.X.: Regularized coplanar discriminant analysis for dimensionality reduction. Pattern Recognition **62**, 87–98 (02 2017). https://doi.org/10.1016/j.patcog.2016.08.024 3, 4, 5
11. Cai, D., He, X., Han, J.: Semi-supervised discriminant analysis 3
12. Samaria, F.S., Harter, A.C.: Parameterisation of a stochastic model for human face identification. In: Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. pp. 138–142. IEEE (1994) 5, 6
13. Yale face database b, `http://cvc.yale.edu/projects/yalefaces/yalefaces.html` 5, 6
14. Sim, T., Baker, S., Bsat, M.: The cmu pose, illumination, and expression (pie) database. In: Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition. pp. 53–58. IEEE (2002) 6
15. Nene, S.A., Nayar, S.K., Murase, H., et al.: Columbia object image library (coil-20) (1996) 6
16. Barlow, H., S. Mao, and M. Khushi, Predicting High-Risk Prostate Cancer Using Machine Learning Methods. Data, 2019. 4(3): p. 129. 8
17. Khushi, M., Dean IM. Teber ET., et. al. Automated classification and characterization of the mitotic spindle following knockdown of a mitosis-related protein. BMC Bioinformatics, 2017. 18(16): p. 566. 8