International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

Projection of Data in a Non-Linear Subspace

N. Veeranjaneyulu[1] Jyostna Devi Bodapati[2]

[1]Professor, Department of Information Technology, Vignan's University, Valdlamudi, India.
[2]Assist Professor, Department of CSE, Vignan's University, Valdlamudi, India.
**** Corresponding author's Email: veeru2006n@gmail.com***

Abstract— ***The problem of high dimensionality has become a great challenge in machine learning community. This problem is gaining much focus due to the increase in the volume of data availability. The challenges related to high-dimensional data is more severe in case of generative models like Gaussian Mixture models (GMMs), as there is high correlation between the number of parameters to be estimated and the number of features used to represent the data. In parametric models like GMM the higher the number of dimensions the higher the number of parameters to be estimated. To attain reasonable accuracy large number of examples are required which is usually ten times the number of parameters but availability of large training data may not be possible in most of the real-time applications. So it is important to represent the high-dimensional data in a reduced dimensional space to design a more robust classifier. Dimensionality reduction of data helps for discriminative models like support vector machines, as the computational complexity required is less compared to the high dimensional data. There are two different types of projections that are popular in literature to reduce the data dimensionality: Linear projection and Non-linear projection. In both these methods data is projected onto the lower dimensions. In this work we carried out multiple studies on linear and non-linear projection of data before feeding the data to a model for classification. Our experimental studies show non-linearly projected data is better classified than linearly projected data. For our experimental studies and illustrations we use both synthetic and real-time, Brain Computer Interface data.***

Index Terms— Principal Component Analysis (PCA), Linear Projection of data, PCA in Kernel space (KPCA), Non-linear projection, Multi-modal GMMs (Gaussian Mixture Models).

I. INTRODUCTION

In the area of machine learning and pattern recognition, dimensionality reduction is a technique of reducing the number of features with which the data is being represented. In simple words it is the process of projecting the data from it's input space (usually higher dimensional) to a lower dimensional space. Usually less number of features are used to represent the data in the reduced space as compared to the dimensions in the original space. The type of projection can be either linear or non-linear.

Two prominent linear projection techniques for feature reduction in the literature are: **Principal Component Analysis (PCA),** and Linear Dimensionality reduction (LDA). In Both these approaches the data is linearly projected to a lower-dimensional space. Major difference between PCA and LDA is: in PCA label information is not required where as in LDA label information is used, while computing the directions for projection. Therefore PCA is an unsupervised approach and LDA is a supervised approach. In PCA the data is projected in the direction of the maximum variance of the data. The dimensionality of the resulting subspace is bounded by the number of dimensions. In LDA the data is projected in the direction that maximizes the separability between classes. In LDA the number of directions the data can be projected is limited to the number of classes. This is the major limitation of this method as the number of classes(C) in general is very few compared to the number of dimensions (d). LDA method would be helpful only when the number of classes is sufficiently large.

There are many non-linear dimensionality reduction approaches proposed in the literature: PCA in the kernel space (KPCA) and LDA in Kernel space (KLDA). These two are non-linear methods that make use of kernel transformations. In these methods first data is transformed to a nonlinear space and in that space the data is projected. Another non-linear dimensionality reduction technique is neural network based auto encoders. In this method data is given as input to the network and the output expected from the network is the data. We train the network such that it gives minimum loss and extract the features from the linear hidden layer. As the features are from a hidden layer these features are known are bottle neck features. Following are the advantages of dimensionality reduction:

- Reduction in the computational and storage requirements.
- Uncorrelated features in the reduced subspace improve the efficiency of certain machine learning models like GMM.

Dimensionality reduction makes the data visualization easy once it is reduced to two or three dimensions.

II. LINEAR DIMENSIONALITY REDUCTION

Two prominent linear projection techniques for feature reduction in the literature are: **Principal Component Analysis (PCA),** and Linear Discriminant Analysis (LDA).

A. Principal Component Analysis:

It is an unsupervised dimensionality reduction

technique that uses orthogonal projection. In this approach directions of maximum variance of the data are computed and data is projected onto those directions. The number of directions of projection (l) is called principal components. Usually the possible number of directions for projection is at most equal to the number of dimensions of the data in the original space(d). The first direction of the projection has the maximum variance and the second projection is in the direction of second maximum variance and so on. Each of these directions of projections are orthogonal to each other as they are the eigen vectors of the covariance matrix which is a symmetric positive semi-definite matrix. So it is guaranteed that the resulting features are uncorrelated.

Let the data D = $\{\bar{x}_n\}_{n=1}^{N}$ each data point $\bar{x}_n$ is of d dimensional and assume that the data is to be reduced to l dimensions with the constraint that l < d. The data is to be transformed to a new feature space a.

Steps:

1. Find the Co-variance matrix(C) of the data D using the following equation

$$C = \frac{1}{N}\sum_{n=1}^{N}(\bar{x}_n - \bar{\mu})^T$$

2. By solving the following characteristic equation we get the eigen vectors,

$$C\bar{v}_i = \lambda v_i$$

Where λ_i is the eigen value associated with the eigenvector v_i such that

$$\lambda_1 \geq \lambda_2 \geq .....\lambda_{d1};$$

Where the eigen vectors are orthogonal to each other and uncorrelated to each other.

3. Compute the projection of x as fallows

$$a_i = (\bar{x} - \bar{\mu})^T \vec{v}_i \qquad i=1,2,\ldots..,d$$

B. Fisher Linear Discriminant Analysis:

FLDA is a linear projection technique that makes use of class labels and comes under the category of supervised learning. In FDA the data is projected in the directions maximum separability. PCA on the other hand does not take into account of discriminant information. FDA requires label information for dimensionality reduction. In FDA the data is projected to the direction w that maximizes the measure of separation.

Let the given data D = $\{\bar{x}_n, y_n\}_{n=1}^{N}$, where each data point $\bar{x}_n \in R^d$ and $y_n \in \{+1,-1\}$. Assume that the data is to be projected to l dimensions with the constraint that l < d i.e. the data is to be transformed to a new feature space $\bar{a} = \bar{w}^t \bar{x}$.

In solving FDA, measure of class separability, is considered as the objective function:

$$J(\bar{w}) = \frac{(m_+ - m_-)^2}{(s_+^2 + s_-^2)} \qquad (1)$$

where m_+, m are the mean of the projected data of positive and negative classes respectively. s_+ and s are the scatter matrices projected data of the positive and negative classes respectively. Equation (1) can be expressed as follows:

$$J(\bar{w}) = \frac{\bar{w}^t S_B \bar{w}}{\bar{w}^t S_W \bar{w}} \qquad \text{---------- (2)}$$

In the above equation S_w denotes the total within-class scatter matrix and S_B denotes the between class scatter matrix of the data in the input space.

The above maximizing problem can be posed as a constrained optimization problem and the lagrangian of the optimization problem is:

$$L_p = \frac{1}{2}\bar{w}^t S_B \bar{w} + \frac{1}{2}(\bar{w}^t S_w \bar{w} - 1) \qquad \text{---------- (3)}$$

III. NON LINEAR DIMENSIONALITY REDUCTION

A. PCA in the kernel space:

In linear projection (PCA), the projection is in the input space. In kernel PCA the data is transformed to a kernel space and data is projected in that transformed space. The transformation from input space to the kernel space $(\bar{\varphi}(\bar{x})$ space)is non-linear hence it is a non-linear projection in the kernel space.

Let there are N data points in the given data, D = $\{\bar{x}\}_{n=1}^{N}$, where each data point $\bar{x}_n \in R^d$ X R^d. Let $\bar{\varphi}(\bar{x})$ is the data point x represented in the kernel space and $\bar{a} \in R^l$ the data point x represented in the kernel space, usually l is expected to be less than d.

$$\bar{x} \rightarrow \bar{\varphi}\{\bar{x}\} \rightarrow \bar{a}$$

Following is the characteristic equation in '(x) space that is to be solved to get the principal components for projection in the kernel space:

$$C^{\varphi}\bar{\vartheta}_i = \lambda_{i\bar{\vartheta}_i} \qquad \text{------- 4}$$

In the above characteristic equation, C^{φ} is the covariance matrix of the data in kernel space. The challenge here is computing C^{φ}. It is always not possible to compute C^{φ} especially when an implicit kernel (Gaussian kernel) is used to transform data from input space to the kernel space. As we cannot directly compute C^{φ}, we solve the following characteristic equation which is equivalent to (4).

$$\bar{K}\bar{\alpha}_i = \lambda_i N\bar{\alpha}_i \qquad \text{---------- (5)}$$

Equation (5) is the characteristic equation in terms of the Kernel Gram Matrix after mean subtraction (K^e). By solving this equation we get the values. The l, 's corresponding to the most significant eigen values are considered for projection.

The directions for projections can be computed using the following equation:

$$a_i = \sum_{n=1}^{N} \alpha_{in} \overline{K}(x_i, x_n) \qquad \text{i=1,...l}$$

The vector a gives the reduced dimension representation (a_i) of the data point (x_i). This projection is non-linear as the kernel used to transform data is a non-linear kernel. Once the data is transformed to a non-linear space, it is projected in that kernel space. The input data is non-linearly related to the projected data, So the projection is said to be non-linear. Kernel PCA involves finding the eigenvectors of the kernel gram matrix, K, of size NXN rather than finding the eigenvectors of the d covariance matrix of conventional linear PCA.

In principle $\overline{K}$ have N significant eigen values and hence the possible number of directions (l) for projection is bounded by N (usually N > d). Therefore there is no guarantee that the reduced dimension is less than d.

B. LDA in kernel space:

As KPCA is the non-linear extension to PCA, KFDA is the non-linear version of the LDA. To extend FDA to non-linear mapping, the data can be mapped to a new feature space(a) via some function φ.

$$\overline{a} = \overline{w}^t \varphi(\overline{x}) = \phi_{\overline{w}}$$

The objective function of KFDA in terms of data in the kernel space is represented as the following maximization function:

$$J(\overline{w}) = \frac{(m_+^{\phi} - m_-^{\phi})^2}{s_+^{\phi 2} + s_-^{\phi 2}} = \frac{\overline{w}^t S_B^{\phi} \overline{w}}{\overline{w} S_w^{\phi} \overline{w}} \qquad (6)$$

where $m_+^{\phi}, S_-^{\phi}, S_B^{\phi}, s_+^{\phi}, s_-^{\phi}$ corresponds to m_+, m , S_B, S_W, s_+, s respectively in the kernel feature space.

The Fisher discriminant maximizes the ratio between the quantities as seen in equation(6). The motivation for this choice is that the direction chosen maximizes the separation of the means scaled according to the variances in that direction.

The regularized Fisher discriminant chooses w to solve the following optimization problem,

$$MaxJ(\overline{w}) = \frac{(m_+^{\phi} - m_-^{\phi})^2}{\sigma_+^{\phi 2} + \sigma_-^{\phi 2} + \frac{\lambda}{2} \| \overline{w} \|^2} \qquad (7)$$

Clearly, the direction of the derivative is in the direction of ($_+$). The directions of projection in the kernel space is given by:

$$\overline{a} = \sum_{n=1}^{N} \alpha_n K(\overline{x}, \overline{x}_n) \qquad (8)$$

Here K^e have M solutions and hence the possible number of directions (l) for projection is bounded by the number of classes(M). If the number of classes is very few then there is significant loss in the information.

IV. Experimental Results

A In this section we provide experimental results to show that non-linear dimensionality reduction outperforms linear dimensionality reduction. We carried out experiments using one artificial dataset and one real world image dataset. We use PCA and KPCA to demonstrate all our experiments.

A. *Artificial data*

In this section, we show that non-linear projection helps to transform the complex decision boundary of the data in the original space to a linear decision boundary in the reduced subspace. As we can have the visualization of the data in lower dimensional space like 2D and 3D we first describe the experiments with a 2D artificial dataset.

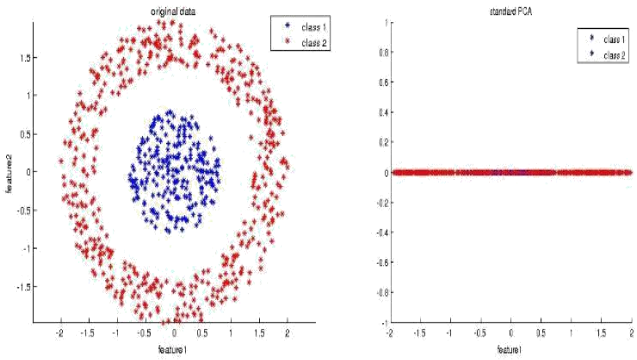

Fig 1: Data in the input space and Data projected to 1D

On seeing the data in Figure 1(a), one can observe that the data is non-linearly separable and the decision boundary is a complex non-linear boundary and is expected to be in the centre of the two classes. A linear classifier cannot be used to classify such data.

First we projected the data using conventional PCA that performs a linear projection in the input space and the same is shown in Figure 1(b). One can see that there is no clue to classify the data.

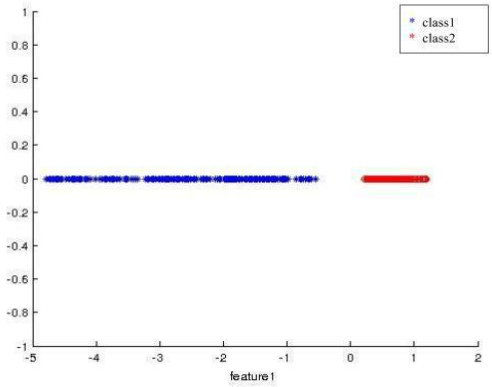

Figure 2: KPCA based non-linear projection in reduced subspace

In such cases linear dimension reduction techniques fail. In Figure 2 we show the projected data in 1D space after applying KPCA based non-linear dimensionality reduction. Now we can see that the data is linearly separable in the reduced subspace. To classify such data we can use any simple classifier such as naive Bayes, as the classes are linearly separable.

Kernel transformation: We tried different types of kernels to transform the data and our experiments show that Gaussian kernel is sufficient to transform the data such that the classes are linearly separable. Gaussian kernel, also known as RBF kernel, on two samples x_m and x_n, represented as feature vectors in some input space, is defined as:

$$k(\bar{x}_m, \bar{x}_n) = \exp^{\frac{-II\,\bar{x}_m - \bar{x}_n II}{2\alpha^2}}$$

For our experiments we set =0.28.

Projecting onto higher dimensional space: As mentioned in the previous section the possible number of directions to be projected is number of examples (N). Hence we can project to more than the number of dimensions in the input space. Figure 3 shows the projection of the 2D data to 2D and 3D space.

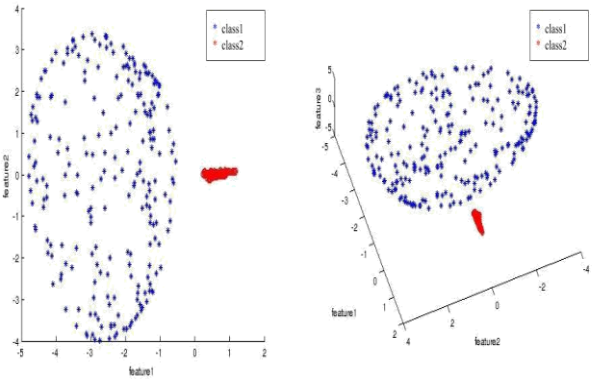

Figure 3: Projection to high-dimensional space

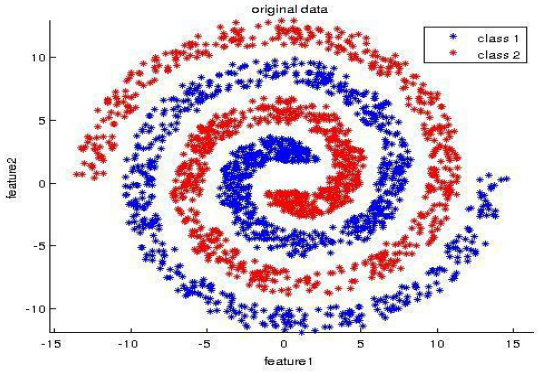

Figure 4: Non-linearly separable data in original space

Projection onto higher dimensions helps when the data is not separable in the lower dimensions. Consider the data given in Figure4 which is not separable even after projecting to 2D and 3D. It might become separable in the higher dimensions.

4.2 Real data

In this section, we use a real world image dataset to show that non-linear projection helps to improve classification accuracy. To illustrate the experimental studies, we use Support Vector Machine (SVM) based classifier.

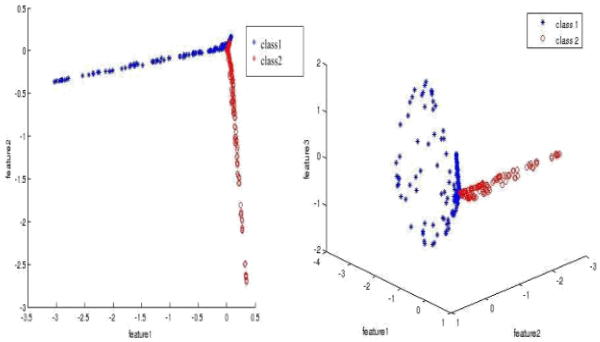

Figure 5: Non-linearly projection of spiral data onto 2D and 3D

Dataset	Classes	Features	Samples
BCI	2	117	400

Table 1: Summary of the dataset used for experimental results

Description of the dataset: This data set created from research toward the development of a brain computer interface (BCI). The data is collected from a single person. That person underwent 400 different trials. In each trial, he imagined movements of the hand writing (both left and right). His imaginations are captured using 39 different electrodes and each electrode gives 3 parameters, which leads to a total of 117 parameters for each trial.

Classifiers: To illustrate the experimental studies, Gaussian Mixture Model (GMM) based classifier is used. The reason for selecting GMM based classifier is the size of dimensionality has a direct impact on the accuracy of classifier. In case of GMM based classification, if the number of dimensions of the data increases the number of parameters to be estimated is large and if the number of dimensions of the data is small the number of parameters to be estimated are also small. In GMM based classification the number of parameters also depends on the type of covariance matrix used. If the covariance matrix is diagonal then the number of parameters to be estimated is linearly proportional to the dimensionality of the data. If the covariance matrix is non-diagonal or full then the number of parameters to be estimated is in quadratic relation to the dimensionality of the data.

GMM	PCA-GMM	KPCA-GMM
53.75 (117-d)	47.50 (10-d)	53.75 (10-d)
53.75 (117-d)	47.50 (3-d)	50.00 (3-d)
53.75 (117-d)	45.00 (1-d)	48.75 (1-d)

Table 2 shows the performance of linear and non-linear projections on BCI dataset.

The results are compared to the GMM based classifier without any dimensionality reduction. Three mixture components used for these experiments. Along with the accuracy the number inside () represent the size of the dimensionality of the data.

GMM	KPCA-GMM
53.75 (117-d)	53.75 (10-d)
53.75 (117-d)	56.25 (20-d)
53.75 (117-d)	57.50 (30-d)
53.75 (117-d)	57.50 (40-d)
53.75 (117-d)	56.25 (50-d)
53.75 (117-d)	56.25 (60-d)
53.75 (117-d)	55.00 (70-d)
53.75 (117-d)	55.00 (80-d)
53.75 (117-d)	52.50 (90-d)
53.75 (117-d)	53.75 (100-d)

Table 3: Classification Accuracy before and after dimensionality reduction

Following experiment shows the classification accuracies of the data after reducing the data onto different number of directions.

Gaussian components	GMM	KPCA-GMM
1	55.00 (117-d)	56.25 (10-d)
2	55.00 (117-d)	51.25 (10-d)
3	53.75 (117-d)	53.75 (10-d)
4	40.25 (117-d)	43.75 (10-d)
5	38.75 (117-d)	47.50 (10-d)

Table 4: Classification Accuracies before and after dimensionality reduction

Based on Table 4 we can see that as the number of components increase accuracy using GMM in the original space decrease as the number of parameters to be estimated are large. But in case of reduced space GMM performance increases as more number of components can better represent the data and number of parameters to be estimated is not too large as d is fixed to 10.

Based on the above experiments we can conclude that non-linear projection is better than linear projection in terms of classification accuracy. In generative models like Gaussian Mixture models (GMMs) high dimensional data brings severe problems as the number of parameters to be estimated is proportional to the dimensionality of the data. By reducing the number of dimensions using non-linear projection, the number of parameters to be estimated is low and in turn non-linear projection may better discriminate the data in the reduced subspace.

V. CONCLUSION

Based on the above experiments we can conclude that non-linear projection is better than linear projection in terms of classification accuracy. In generative models like Gaussian Mixture models (GMMs) high dimensional data brings severe problems as the number of parameters to be estimated is proportional to the dimensionality of the data. By reducing the number of dimensions using non-linear projection, the number of parameters to be estimated is low and in turn non-linear projection may better discriminate the data in the reduced subspace.

References

[1] Wang." Kernel principal component analysis and its applications in face recognition and active shape models" CoRR, abs/1207.3538, 2012.

[2] Scholkopf, A. Smola, and K.-R. Muller, "Kernel principal component analysis", pages 583–588, 1997.

[3] Scholkopf, A. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem" Neural computation, 10(5):1299–1319, 1998.

[4] S. T. Roweis and L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding" Science, 290(5500):2323–2326, 2000.

[5] Chapelle, B. Schölkopf, and A. Zien, "Semi-Supervised Learning ", Eds. London, U.K.: MIT Press, 2006, pp. 508, ISBN: 978-0-262-03358-9.

.