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

Hybrid Dimensional Reduction Techniques

1.Jyostna devi Bodapati , 2.N. Veeranjaneyulu ,3. B.Suvarna
1,3. Department of CSE,2. Department of IT,
1,2,3. Vignan's University.
jyostna.bodapati82@gmail.com, veeru2006n@gmail.com, sv1720@gmail.com

Abstract— This paper introduces a novel hybrid dimensionality reduction technique. According to the literature, Linear projection techniques like PCA, LDA degrades the performance of the classifiers when applied on real time data. Recent literature claims that non-linear projection techniques like KPCA, KLDA though gives better classification accuracy than the linear projection of data but it is hard to find the suitable kernel function. A combination of linear and non-linear projection is used to project the data. We proposed methods that combine linear and non-linear projection methods in a meaningful way. Extensive studies were conducted to prove the performance of classifiers in the hybrid subspace.

Index Terms—Linear projection, Non-linear projection, Di-mensionality reduction, PCA, LDA, ICA, CCA, NMF, KPCA, KDA, Autoencoders,LLE

I.INTRODUCTION

The problem of high dimensionality has been gaining in-creased focus in the recent literature on pattern recognition and machine learning. This is due to the increase in the availability of the high volumes of data in various fields. Multiple sensors are being used to extract the data and each sensor gives multiple features. On the other side performance of a classifier depends on the type of features that are used to represent the data. But storing the high dimensional data is a challenge as they occupy more space and also demands more computational resources. Another challenge with high dimensional data is in case of parametric models the more the number of features used for data representation the more the number of parameters to be estimated. Many areas of Machine learning depend on the extensive data analysis and data visualization. To address these issues many dimensionality reduction methods have been proposed in the literature. The problem of dimensionality reduction came as an answer to analyze and visualize the huge amounts of multi-variate data. Besides decreasing the dimensionality these methods typically improve the performance of the classification accuracy.

Following are the advantages of dimensionality reduction:

Reduction in the computational and storage requirements. Uncorrelated features in the reduced subspace improves the efficiency of certain parametric models like GMM.

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

Improves generalization ability of the models

Dimensionality reduction techniques are basically categorized into linear and non-linear methods depending on how the input data is related to the projected data. PCA, LDA, ICA, CCA, NMF come under linear dimensionality reduction techniques as the projected data is linearly related to the data in the input space. Techniques like KPCA, KLDA, auto encoders come under non-linear dimensionality reduction techniques as the projected data is non-linearly related to the data in the input space. Another way of interpreting it is linear dimensional reduction techniques makes use of transformations that can be performed on matrix.

Principal component analysis (PCA) and Linear Discriminant Analysis (LDA) techniques are similar in the way data is linearly projected to a lower-dimensional space. The main difference between these two techniques is that PCA is an unsupervised approach as the class label information is not required where as LDA is a supervised approach as it uses class label information. In PCA the data is projected in the direction of the maximum variance. 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 are much smaller than the number of dimensions in general (typical number classes are in tens where as number of dimensions are in hundreds or even thousands). This method would be helpful only when the number of classes are sufficiently large.

Objective of PCA is to maximize variance of the projected data or in other words minimize the reconstruction error. CCA (canonical correlation analysis) [2] on the other hand maximizes cross correlation of the projected data. Assuming the features are correlated to each other, CCA finds the linear combinations of those features that have high correlation with each other.

Non-negative matrix factorization (NMF) [11] do not make use of class label information and comes under the category of unsupervised dimensionality reduction. It is similar to PCA except the coefficients in the linear combination must be non-negative. NMF has an advantage for applications involving large matrices as it can be solved iteratively.

PCA and LDA are non-iterative algorithms and they do not have the problem of reaching local optima. NMF computation involves solving hard non-convex objective that can be solved iteratively and there is chance of reaching a local optima.

Independent Component analysis (ICA) [Herault & Jutten, 86] finds the components that are as independent as possible and doesnt work if data is Gaussian. ICA finds the direc-tions such that data projected onto the directions that have maximum statistical independence by minimizing the mutual information.

Major challenge with linear models is that they allow linear projections alone which is not sufficient for the data whose intrinsic manifold is more complex in nature. Significant limitation of classical PCA is: it depends on mean, the first order moment and covariance, the second order moment of the data which is not the case in case of kernel PCA. Mul-tiple dimensions in the data brings more non-linearity to the data. The disadvantage of the linear dimensionality reduction techniques is that they are able to find a linear subspace alone which are poor representation of complex non-linear data.

To represent the data in non-linear manifold, many non-linear dimensionality reduction approaches are proposed in the literature. Kernel PCA (KPCA) and Kernel LDA (KLDA) are two non-linear methods that make use of kernel transforma-tions. In these methods first data is transformed to a nonlinear space and in that space the data is projected.

Auto encoder is another non-linear dimensionality reduction technique that makes use of the principles of neural network. In this method data is given as input to the network and the output expected from the network is the data with little or no loss. The network is trained 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 as bottle neck features.

Similar to PCA, Locally linear embedding (LLE) algorithm does not make use of class labels and comes under the category of unsupervised dimensionality reduction. It takes the high-dimensional data as input and computes low-dimensional data that preserves neighborhood embedings of the input data. As the objective of LLE is convex there is no chance of reaching in local minima as it is in the case of NMF. LLE is able to learn the global structure of nonlinear manifolds by exploiting the local symmetries of linear reconstructions [Sam T. Roweis1, Lawrence K. Saul2].

II. LINEAR DIMENSIONALITY REDUCTION TECHNIQUES

PCA, LDA, ICA, CCA and NMF techniques are described in this section

A. Principal component analysis (PCA) [1] :

PCA is an unsupervised dimensionality reduction technique that uses orthogonal projection. Data is mapped onto the directions of maximum variance in the data. The number of directions of projection(l) are called principal components. Usually the possible directions of projection are less than or 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. Hence it is guaranteed that the resulting features are uncorrelated.

Let the data D = $fx_n g^N_{n=1}$, each data point x_n is of d dimensional and assume that the data is to be reduced to ldimensions with the constraint that l < d. The data is to be transformed to a new feature space a.

Steps:

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

$$C = {}_N^{1\ PN}{}_{n=1}(x_n\ \ )(x_n\ \ )^T$$

ii By solving the following characteristic equation we get the eigen vectors,

$$Cv_i = {}_i v_i$$

where $_i$ is the eigenvalue corresponding to the eigen-vector v_i such that $v_i^T v_i = 1$; and $v_i^T v_j = 0$; $8i \neq j$

iii Compute the projection of x as follows

$$a_i = (x\ \ )^T v\sim_i \qquad i = 1; 2; :::; d$$

Limitations: When PCA is applied on data that has different scales (one variable in millions and another in hundreds) the projected data leads to issues. This issue can be addressed by normalizing the data before applying PCA. A d-dimensional data can be projected upto d-1 dimensions. Evaluating the covariance matrix in an accurate manner is a non-trivial task

[15]. Even the simplest invariance could not be captured by the PCA unless the training data explicitly provides this information [19]. Performance of classifier on projected data would be poor if the direction of separation between classes is not in the direction of maximum variance of the data.

Extensions: Algorithms for dimensionality reduction are computationally expensive and repeated computations due to accumulated data are computationally prohibitive. To address this issue, an out-of-sample extension scheme is proposed to extend to newly-arrived data points [21]. An approach called local linear approach [4] to dimension reduction provides accurate reduced representations of data. In the hierarchical PCA [17], a image is divided into various parts and then PCA is applied on each part separately and then the results are combined. 2DPCA [13] is based on two-dimensional matrices of images instead of the 1-dimensional vectors. This avoids the transformation of images into 1-dimensional vectors before extracting features.

B. Fisher Discriminant Analysis(FDA)

Fisher Discriminant Analysis(FDA) was proposed by Fisher in 1936. FDA [7] can be used for binary classification task in addition to dimension reduction. Unlike PCA, 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 = $fx_n; y_n g^N_{n=1}$, where each data point x_n 2 R^d and y_n 2 f+1; 1g. 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 a = $w^t x$. In solving FDA, measure of class separability, is considered as the objective function :

$$J(w) = \frac{(m_+ \quad m)^2}{(s^2 + s^2)} \qquad (1)$$

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

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

where S_W represents the total within-class scatter matrix and S_B represents the between-class scatter matrix of the data in the input space.

We can convert the problem of maximizing J in equation 2 into an objective function with associated conditions as follows:

$$L_P = \frac{1}{2} w^t S_B w + \frac{1}{2}(w^t S_W w \quad 1) \qquad (3)$$

Solving the above equation tells that the directions to be projected are in the direction that maximizes the separation of means of the classes.

Linear discriminant analysis (LDA) is a generalized version of FDA. The same can be extended to multi-class data and is called as Multiple discriminant analysis (MDA). LDA results at most M 1 directions for projections where data belongs to M different classes. LDA fails if the direction of separation between classes is not in the mean but in the direction of maximum variance of the data.

Extensions: Other variants of LDA proposed in the literature are: Non-parametric LDA [Fukunaga, Orthonormal LDA [Okada and Tomita], DiscLDA [20]. Linear discriminant analysis (LDA) and K-means clustering are together used to form a method called (LDA-KM [18]) to get the features that are best separabble in the reduced subspace.

C. Independent Component Analysis (ICA) :

ICA is a linear projection method that is proposed by Herault in 1991. ICA works on the assumption that data are linearly mixed by a group of independent sources and ungroup these sources based on their statistical independency measured by mutual information. Let v_1; v_2; v_3; :::v_d denote the projection directions of independent components. ICA finds these directions such that data projected onto these directions have maximum statistical independence by minimizing the mutual information or maximize the non-Gaussianity.

D. Canonical Correlation Analysis (CCA) :

Now consider two feature spaces that is sets of variables x and y, x is a vector of u variables y is a vector of v variables. CCA finds a projection direction u in the space of x, and a projection direction v in the space of y, so that projected data onto u and v has max correlation. CCA simultaneously finds dimension reduction for two feature spaces. The objective function of CCA is:

$$\max_{u2R^p;v2R^q} \frac{u^T X^T Y v}{\sqrt{(u^T X^T Xu)(v^T Y^T Y v)}}$$

E. Non-Negative Matrix factorization (NMF) :

NMF [11] also known as Positive Matrix Factorization, is a dimensionality reduction method that involves factoring the given data matrix into a low rank, sparse and non-negative factors. Let V be a non-negative matrix of dimension: n m, NMF algorithm decomposes the matrix V into two matrices W and H such that V can be approximated using W and H, i.e. V WH. W and H are low rank sparse and non-negative factors of V that have non-negative elements. W is of dimension n r, and is called the basis matrix because its row contains set of basis vectors. H is of dimension r m, and is called a weight matrix because its row contains coefficient sequences. The rank r of the factorization is chosen such that (n + m)r < nm. The columns of H are in one-to-one correspondence with the columns of V . Thus the result (WH) can be interpreted as weighted sum of each of the basis vectors in W , the weights been the corresponding columns of H. The additive properties resulting from the non-negative constraints of NMF results in basis vector that represents local components of the original data.

III. NON-LINEAR DIMENSIONALITY REDUCTION TECHNIQUES

Two non-linear dimensionality reduction techniques kPCA and Auto encoders are used to demonstrate the work in this paper. In kPCA first data is transformed to a nonlinear space and data is projected in that transformed space.

A. Kernel Principal Component Analysis [6]:

In 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 space. The transformation from input space to the kernel space ('(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 = $fx_n g^N_{n=1}$, where each data point x_n 2 R^d. Let '(x) is the representation of x in the kernel space and a 2 R^l be the final reduced dimension representation, l is expected to be less than d. This transformation can be represented as follows:

x ! 'fxg ! 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:

$$C'v_i = {}_iv_i \quad (4)$$

where C' is the covariance matrix of the data in the kernel space. The challenge here is computing C' is not possible when implicit kernel(Gaussian kernel) is used to transform data from input space to the kernel space. As we cannot directly compute C', we solve the following characteristic equation which is equivalent to (4).

$$K_i = {}_iN_i \quad (5)$$

characteristic equation in terms of the Equation (2) is the e

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:

$$a_i = \sum_{n=1}^{N} {}_{in}\tilde{K}(x_i; x_n) \quad i = 1; : : : ; l$$

The vector a gives the reduced dimension representation (a_i) and is non-linearly related to the input space(x).

Kernel PCA involves finding the eigenvectors of the kernel gram matrix, $\tilde{K}$, of size N N rather than finding the eigenvectors of the d d covariance matrix of conventional linear PCA.

In principle K^e have N significant eigenvalues and hence the possible number of directions(l) for projection is bounded by N (usually N > d) instead of d as it is in case of PCA. Therefore there is no guarantee that the reduced dimension is less than d as the number of examples (N) is usually much much greater than the number of dimensions (d).

B. Kernel Fisher Discriminant Analysis [kFDA]

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 '.

$$a = w^{t\prime}(x) = w$$

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

$$J(w) = \frac{(m_+ \quad m)^2 \qquad w^t S \; w}{(s \qquad + s \;)^2 \qquad {}_w t_S \qquad w} = \frac{B}{W} \tag{6}$$

where m_+, m , S_B, S_W, s_+, s 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,

$$\max J(w) = \frac{(m_+ \quad m)^2}{w \qquad {}_+ + \qquad + {}_2 \, kwk^2} \tag{7}$$

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

$$a = {}_n K(x; x_n) \tag{8}$$

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 are very few then there is significant loss in the information. If this is extended to multiple classes then it is known as generalized discriminant analysis (GDA) [8].

C. Locally linear embedding (LLE)

Similar to PCA, Locally linear embedding (LLE) [9] algorithm does not make use of class labels and comes under the category of unsupervised dimensionality reduction. It takes the high-dimensional data as input and computes low-dimensional data that preserves neighborhood embedings of the input data. As the objective of LLE is convex there is no chance of reaching in local minima as it is in the case of other approaches.

Let us assume the given data D = x_1; x_2:::x_n contains n examples. Each example x_i 2 R^d. Assume that D can span the input space sufficiently. That is each data point and its neighbors are close to each other and lie on the same stripe of the manifold. These stripes can be reconstructed using the data points on its neighborhood. Error between the original and reconstructed can be computed using the cost function:

$$(W) = \sum_i (x_i \qquad \sum_j W_{ij} x_j)^2$$

subject to $\sum_j W_{ij} = 1$ and $W_{ij} = 0$ if X_j does not belong to the set P of neighbors of X_i

Sum of the squared error gives the total error in reconstruction. The weight W_{ij} refers to the contribution of the j^{th} data point in reconstructing the i^{th} data point. By taking the derivative of the above cost function and equating that to zero gives the weight parameters.

D. Auto encoder based dimensionality reduction

Auto encoder [16] is a neural network based encoder used especially for dimensionality reduction. An MLFNN is trained with the same data given at the input and output layers that are linear. One of the hidden layer is a linear layer with lesser number of neurons than the input layer and we call this layer as bottle neck layer from which the features are extracted.

Figure 1 shows an example architecture with a 5 layer MLFNN with one input, one output and 3 hidden layers. The first and third hidden layers are non linear and the second hidden layer is a linear layer which is called as bottleneck layer from which features are being extracted. Figure 1 shows the details of the architecture.

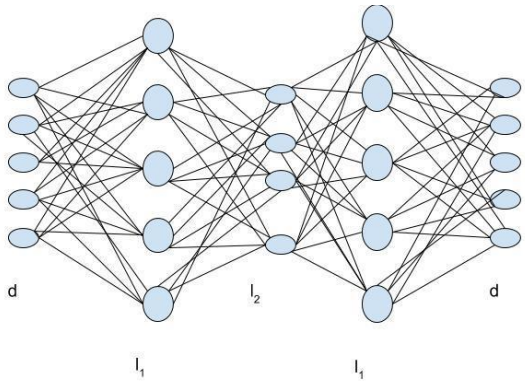

Fig. 1. A sample architecture used for auto encoder

Extensions to non-linear methods: Building a piecewise linear model [3] of the data provides compression that is superior to the globally linear model produced by PCA that is superior to the global nonlinear model constructed by a five-layer auto associative neural network. LLE and Isomap tech-niques are analyzed and enhanced their visualization power for data scattered among multiple clusters [10]. LPP [12] is proposed as an alternative to PCA. Neighborhood preserving projections [14] (NPP) is proposed as a novel linear dimension reduction method that has good preserving property than PCA. A reduced-dimension remapping [5] of pattern data is proposed in an unsupervised nonlinear manner, but with a constraint that the overall variance in a representation of the data be conserved.

IV. PROPOSED APPROACH

If we use KPCA alone, the number of reduced dimensions are bounded by number of examples, and in reduced space, features are uncorrelated. If we use GDA alone the number of reduced dimensions are bounded by number of classes, and after reduction the features are discriminative. We can combine these two feature reduction techniques to make the resultant features both discriminative and uncorrelated. this can be done in either of the following ways but the later approach is not that meaningful.

i KPCA followed by GDA: If we apply KPCA followed by GDA the resultant dimensions are bounded by num-ber of examples in the data set, and are uncorrelated. Hence applying GDA after KPCA is meaningful only
when the number of classes are large enough.

ii GDA followed by KPCA : This is not much meaningful as after applying GDA the resultant dimensions are bounded by number of classes(usually number of classes are very small compared to number of dimen-sions). Applying KPCA after GDA would not be much meaningful.

V. CONCLUSION

Based on this survey we can conclude that dimension reduction not only provides better visualization of the data but also helps in improving the performance of the machine learning algorithms. Linear dimensional reduction techniques are simple and elegant as long as the data is not in the non-linear manifold. PCA,LDA and CCA are trivial optimization problems where as NMF and ICA are non-convex optimization problems and are possible to reach in a local minima. Non-Linear dimensional reduction techniques are much complex and gives better representation of the data in a non-linear subspace. In case of KPCA and KLDA choosing proper kernel is an issue and in case of auto encoders choosing the structure of the network is a challenge. As today's real world data is much complex it is wise to use non-linear methods for dimensionality reduction and can also be combined as proposed in the previous section.

REFERENCES

[1] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2.11 (1901), pp. 559–572.

[2] Harold Hotelling. "Relations between two sets of variates". In: Biometrika 28.3/4 (1936), pp. 321–377.

[3] Nandakishore Kambhatla and Todd K Leen. "Fast nonlinear dimension reduction". In: Neural Networks, 1993., IEEE International Conference on. IEEE. 1993, pp. 1213–1218.

[4] Nandakishore Kambhatla and Todd K Leen. "Dimen-sion reduction by local principal component analysis". In: Neural Computation 9.7 (1997), pp. 1493–1516.

[5] Yoh Han Pao. Self-organization of pattern data with dimension reduction through learning of non-linear variance-constrained mapping. US Patent 5,734,796. 1998.

[6] Bernhard Scholkopf,¨ Alexander Smola, and Klaus-Robert Muller¨. "Nonlinear component analysis as a kernel eigenvalue problem". In: Neural computation 10.5 (1998), pp. 1299–1319.

[7] Bernhard Scholkopft and Klaus-Robert Mullert. "Fisher discriminant analysis with kernels". In: Neural networks for signal processing IX 1.1 (1999), p. 1.

[8] Gaston Baudat and Fatiha Anouar. "Generalized discriminant analysis using a kernel approach". In: Neural computation 12.10 (2000), pp. 2385–2404.

[9] Sam T Roweis and Lawrence K Saul. "Nonlinear dimensionality reduction by locally linear embedding". In: Science 290.5500 (2000), pp. 2323–2326.

[10] Michail Vlachos et al. "Non-linear dimensionality reduction techniques for classification and visualization". In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2002, pp. 645–651.

[11] Patrik O Hoyer. "Non-negative matrix factorization with sparseness constraints". In: The Journal of Machine Learning Research 5 (2004), pp. 1457–1469.

[12] X Niyogi. "Locality preserving projections". In: Neural information processing systems. Vol. 16. MIT. 2004, p. 153.

[13] Jian Yang et al. "Two-dimensional PCA: a new approach to appearance-based face representation and recognition". In: Pattern Analysis and Machine Intelligence, IEEE Transactions on 26.1 (2004), pp. 131–137.

[14] Yanwei Pang et al. "Neighborhood preserving projections (NPP): a novel linear dimension reduc-tion method". In: Advances in Intelligent Computing. Springer, 2005, pp. 117–125.

[15] P Jonathon Phillips et al. "Overview of the face recognition grand challenge". In: Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on. Vol. 1. IEEE. 2005, pp. 947–954.

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural net-works". In: Science 313.5786 (2006), pp. 504–507.

[17] Abhishek Agarwal et al. "Efficient hierarchical-PCA dimension reduction for hyperspectral imagery". In: Signal Processing and Information Technology, 2007 IEEE International Symposium on. IEEE. 2007, pp. 353–356.

[18] Chris Ding and Tao Li. "Adaptive dimension reduction using discriminant analysis and k-means clustering". In: Proceedings of the 24th international conference on Machine learning. ACM. 2007, pp. 521–528.

[19] Chunming Li et al. "A statistical PCA method for face recognition". In: Intelligent Information Technol-ogy Application, 2008. IITA'08. Second International Symposium on. Vol. 3. IEEE. 2008, pp. 376–380.

[20] Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. "DiscLDA: Discriminative learning for dimensionality reduction and classification". In: Advances in neural information processing systems. 2009, pp. 897–904.

[21] Yariv Aizenbud, Amit Bermanis, and Amir Averbuch. "PCA-Based Out-of-Sample Extension for Dimensionality Reduction". In: arXiv preprint arXiv:1511.00831 (2015).