A REVIEW ON LINEAR AND NON-LINEAR DIMENSIONALITY REDUCTION TECHNIQUES

[1]Arunasakthi. K, [2]KamatchiPriya. L

[1] *Assistant Professor*
Department of Computer Science and Engineering
Ultra College of Engineering and Technology for Women,India.
[2]*Assistant Professor*
Department of Computer Science and Engineering
Vickram College of Engineering, Enathi, Tamil Nadu, India.

ABSTRACT

Analysis on the high dimensional data is the main problem in several applications like content based retrieval, speech signals, fMRI scans, electrocardiogram signal analysis, multimedia retrieval, market based applications etc., to improve the performance of the system, the dimensions should be reduced into lower dimension. There are many techniques for both linear and non linear dimensionality reduction. Some of the techniques are suitable linear sample data and not suitable for non linear data and sample size is another criteria in dimensionality reduction. Each technique has its own features and limitations. This paper presents the various techniques used to reduce the dimensions of the data.

KEYWORDS

High dimensional data, Dimensionality reduction, sample size, linear and non-linear techniques

1. INTRODUCTION

Dimensionality reduction is a process of extracting the essential information from the dataset. The high-dimensional data can be represented in a more condensed form with much lower dimensionality to improve the classification accuracy and to reduce computational complexity. Dimensionality reduction becomes a viable process to provide robust data representation in relatively low-dimensional space in such many applications like electrocardiogram signal analysis and content base retrieval [1].

The mathematical representation of the problem is defined as follows: Let us consider the high dimensional dataset X with D-dimensional data. Feature extraction involves to find the low dimensional dataset Y with d-dimensional data which are meaningful low dimensional data, where d□D. Initially, the high dimensional data D is mapped on to the low dimensional subspace

d. In this paper, the points x_i and x_j represents the i^{th}, j^{th} record of the high dimensional data D and y_i, and y_j represents the i^{th}, and j^{th} record in the low dimensional data d.

2. DIMENSIONALITY REDUCTION TECHNIQUES

Dimensionality reduction reduces the number of variables which improves the performance of the classification. Processing of the high dimensional data leads the increase of complexity both in execution time and memory usage. There are number of techniques available to reduce the dimensions of the dataset. Each and every technique reduces the dimensions of the data based on particular criteria. In recent years, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are regarded as the most fundamental and powerful tools of dimensionality reduction for extracting effective features of high-dimensional vectors in input data. Depending on the data, the reduction techniques are classified as linear techniques and non linear techniques. In following sections we deal with these different techniques. Generally, there are two types of data like linear data and non linear data.

3. LINEAR DIMENSIONALITY REDUCTION TECHNIQUES

Data which has linear relationship is called as linear data and others are called as non linear data. There are number of techniques available to handle this type of linear data. This section deals with Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA).

3.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a supervised dimensionality reduction technique. The feature selection in traditional LDA is obtained by maximizing the difference between classes and minimizing the distance within classes. For better separation, the high dimensional space is reduced into low dimensional subspace [1]. Let $X= [x_{1. . .}x_n] \in \mathbf{R}^{m\times n}$ be the training samples, where m and b the member of data samples and C_i denotes the class of data i. Generally, the conventional LDA finds the Projection matrix $\mathbf{W} = [\mathbf{w}1. . . \mathbf{wn}] \in \mathbf{R}^{m\times n}(n < m)$ whose columns $\{\mathbf{w}_k\}$ (k = 1, . . . , n) constitute the bases of the n-dimension linear subspace. The data samples x_j is projected onto W which gives the projected n-dimension vector y_j that denotes the features of x_j. The optimal Projection matrix is obtained by using equation (1),

$$J(W) = \frac{tr(S_b)}{tr(S_w)} \quad (1)$$

Where tr(.) denotes the trace of matrix, S_b and S_w represents the between class and within class scatter matrix in the feature space and J is the Fisher scalar used for measuring the class seperability. Between class and within class matrixes are calculated as (2) and (3).

$$S_b = \frac{1}{n}\sum_{i=1}^{C} n_i(\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T \quad (2)$$

$$S_w = \frac{1}{n}\sum_{i=1}^{C}\sum_{j\in C_i}(\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T \quad (3)$$

Where, $\bar{y}$, $\bar{y_i}$ denotes the global mean and mean vector of the i^{th} class and superscript T indicates the transposition operation in the feature space. The between class and within class scatter matrix for the data X can be identified by substituting (4) on equations (2) and (3).

$$y_j = W^T x_j \quad (4)$$

$$S_b = \frac{1}{n}\sum_{i=1}^{C} n_i W^T(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T W \quad (5)$$

$$S_w = \frac{1}{n}\sum_{i=1}^{C}\sum_{j\in Ci} W^T(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T W \quad (6)$$

Where $\bar{x}$ and $\bar{x}_i$ represents the global mean and mean vector of i^{th} class vector. Using (5) and (6) the equation (1) can be written as,

$$J(W) = \frac{tr(W^T G r_b W)}{tr(W^T G r_w W)} \quad (7)$$

The optimal objective function is obtained using (7).

The LDA [2] usually uses the global structure information of the total training samples to determine the linear discriminant vectors and these vectors are all global. For a test sample, the use of global linear discriminant vectors to extract features from the samples may lead to erroneous classification, whereas the use of local linear discriminant vectors might produce correct classification. When the global data structure is not completely consistent with the local data structure, Local Linear Discriminant Analysis (LLDA) is more powerful than the traditional LDA algorithms and LLDA can effectively capture the local structure of samples.

To preserve the local intrinsic structure, Joint Global and Local structure Discriminant Analysis (JGLDA) is a novel approach for linear dimensionality reduction [3]. Rotational LDA algorithm is an iterative algorithm, rotates the original feature vectors with respect to the centered of their own class separately, until overlapping error is minimized [4]. The generalization performance is improved in Fisher Linear Discriminant Analysis (FLD) [5].

The major issues of LDA are Small Sample Size (SSS) problem and Common Mean (CM) problem. Small Sample Size problem occurs when the dimensions of the data exceeds the number of samples. There are number of techniques to overcome this SSS problem. Null space based LDA(NLDA)[6], Shrunken centroids regularized discriminant analysis [7], LDA with generalized singular value decomposition [8], null space LDA [9], discriminative common vector (DCV) [10], kernel discriminative common vector (kDCV) [11], orthogonal Centroid Method (OCM) [12], weighted piecewise LDA [13], and LDA over PCA [14] is used to solve the Small Sample Size problem.

Since the objective function of the conventional LDA is based on the distance criteria using L2-norm,it is sensitive to outliers. It comes to know that, the robustness of the LDA with L1-norm is better than the L2-norm [15].

3.2 Principal Component Analysis

Principal Component Analysis (PCA) is one most popular unsupervised technique to handle the curse of dimensionality and it plays important role in pattern recognition and machine learning. The projection on PCA is done by maximizing the correlation between the data. The projection provides the low dimensional subspace which can represent all the data without losing the any information [16]. The main idea of PCA is to transform the high dimensional input space onto the feature space where the maximal variance is displayed. The mathematical formulation for PCA is given below.

Let $X=[x_1 \ldots x_m]^T$ be the input vector that denotes X as the m-dimensional data input data. The sample mean of the given input is calculated as in (8),

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

Where, $\bar{X}$ denotes the Mean of the sample data X and n denotes the number of samples. The covariance of the matrix is identified by using equations (4) and (5).

$$C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x}) \tag{9}$$

$$= \frac{1}{n}\sum_{1}^{n} x_i x_i^T - \bar{x}\bar{x}^T \tag{10}$$

PCA is performed by finding the Eigen values and Eigen vectors of the covariance matrix and rearranged in descending order according to the corresponding Eigen values using a transformation matrix T, which can produce the new form of input vector X.

$$PC = T(x - \bar{x}) \tag{11}$$

In equation (11), T represents the transformation function and PC is the new form of vector which is minimally correlated. To reduce the dimensionality we can select top k number of components where (k<m). This is the general process involves on the Principal Component Analysis (PCA).

Though the conventional PCA is performed in many applications, it has some problems. The main problem of PCA is that the Mean Square Error (MSE) is dominated with the large number of errors. PCA based on L2-norm becomes sensitive to outliers. To overcome this problem PCA based on L1-norm is proposed to improves the robustness [17][18]. In [15],[20],[21], the projection follows Laplacian distribution and L1-PCA is formulated by applying Maximum likelihood estimation to the original given data. The problem of L1-PCA is solved by using weighted median method [20] and convex programming method [21] and it becomes computationally expensive.

In [23], Rotational PCA (R1-PCA) is proposed to combine the advantages of both L1-PCA and L2-PCA. It is rotational invariant and successfully reduces the effect of outliers. PCA based on Maximum Currentropy Criterion (MCC)[24], Robust Two dimensional PCA (RTDPCA) solves the problems of outliers and robustness[19].

Conventional PCA is not probabilistic. Moghaddam[25] extended the conventional PCA into a Probabilistic framework and Probabilistic PCA (PPCA) is derived from the linear latent variable model which can be used to handle the One dimensional(1 -D) data vector. Probabilistic second order PCA (PSOPCA) is a model to follow the classical latent variable model and used different learning [26]. 2D-PCA is used to extend the PCA to handle the 2-Dimensional data (such as image) vectors [27]. Parameter estimation in PPCA requires latent variables which lead to get slower convergence [28]-[30]. To overcome these problems Bilinear Probabilistic Principal Component Analysis (BPPCA) was proposed in the curse of dimensionality on the two dimensional data is solved by using [31].

4. NON LINEAR TECHNIQUES

In real world, most of the data are in the form of non linear. Handling these types of data for further analysis is difficult. There are many techniques, which can handle this type of non linear data.

4.1 Support Vector Machine

Support Vector Machine is a supervised technique for classification which classifies the data into different classes based on the hyper plane and it considers only the support vectors for the problem of classification. Each set of input record or instance has own class labels. The major work of this SVM is to find the relationship among the input dataset and its output labels. Generally, SVM is used for two class classification and its class may be 0 or 1 otherwise -1 or 1. Let us consider $X=(x_1 \ldots x_D)$ be a high dimensional data and each x_i has its own class labels Y= [-1, 1].

In the case of linear data, SVM tries to find the hyper plane with minimum distance from the data points from the boundary. If the data is non-linearly distributed, the data is transformed by using non-linear transformation functions. The training set and the corresponding output is defined as, $T= \{(x_1,y_1), (x_2,y_2) ,\ldots,(x_n,y_n)\}$ $x_i \epsilon R^n$ Where, $y_i \epsilon \{-1, +1\}$ denotes the corresponding output. The optimal hyper plane is identified by Eq. (1),

$$Y = w^T + b = 0 \tag{11}$$

Here, $w \epsilon R^n$ and $b \epsilon R$. The empirical risk is measured with the soft margin loss function by introducing the regularization terms and the slack variables $\Psi = (\Psi_1 \ldots \Psi_n)$. The soft margin function is expressed in Eq. (2).

$$\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b)) \tag{12}$$

The Support Vector Machine Problem is defined using the regularization term is expressed in (3) and (4) and (5) represents the supporting hyper planes which are parallel to the decision plane.

$$\min \ \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \Psi_i$$
$$y_i(w^T x_i) + b \geq 1 - \Psi_i, \ \Psi_i \geq 0, \ \ i = 1,..,n \tag{13}$$

$$w^T + b = 1 \quad (14)$$

$$w^T + b = \mathbf{1} \quad (15)$$

Where, C>0 is the constant parameter. Minimization of the regularization term $\frac{1}{2}||w||^2$ maximizes the margin between the parallel hyper planes.

The general framework of LDA is based on the most intuitive LDA with zero within class variance. It is found that SVM and LDA are almost similar. Commonly, the projection on these techniques is done by maximizing the variance between classes and minimizing the variance within classes [32]. The problems of LDA do not affect the performance of SVM and it reduces the empirical error by maximizing distance between the margin and the hyper plane. Maximizing the width of the margin leads to trade-off between the empirical error and complexity and it is robust under noisy environment. In SVM, Small Sample Size Problem (SSS), and Common Mean (CM) problem does not affect the performance of the classification as in LDA. Compared with other techniques, SVM reduces the Structural risk and empirical error [33]. Recursive SVM (RSVM), Large Scale maximum Margin Discriminant Analysis (large scale (MMDA) and Maximum Margin Projection(MMP) are some other techniques used to reduce the dimensions of the data[34]-[37].

Complexity of SVM will be increased when the number of Support Vectors are increased and it makes difficult to find the hyper plane among the wide range of data [38][39]. To minimize the complexity issues of SVM, Separable Case Approximation (SCA) reduce the number of support vectors by minimizing the distance between the original weight vectors [40]. [41] Eliminates the support vectors which are linearly dependent on other support vectors. [42] Keerthi et.al. Proposed greedy method SVM which is directly related with the training cost.

SVM takes long time for training and large amount of memory while using large datasets. The execution speed of SVM is improved by Sequential Minimal Optimization (SMO), decomposed Support Vector Machine (DSVM) and Core Vector Machine (CVM) [43]-[50]. In [51], SVM handles the online classification problems.

4.2 Independent Component Analysis

Independent Component Analysis (ICA) is an essential unsupervised method to extract the independent features from the high dimensional data. The main goal of ICA is to recover the independent data given from the observation that are linearly dependent on another data. ICA finds the correlation among the data, and decorrelates the data by maximizing or minimizing the contrast information. It has been applied in many applications in Image Processing, Signal Processing, etc,. ICA is mostly related to the Blind Source Separation (BSS) method. In this method, the source represents the independent components; blind represents the little assumption on the mixed high dimensional data. ICA focuses on to find the independency among the data.

Let consider the $X=[x_1 \ldots x_m]$ be the m-dimensional input vector with Non-Gaussian distribution. ICA mainly focus onto maximize the non linearity by using non-linear transformation function.

$$(x_1,\ldots,x_m)^T = f\,(s_{1\ldots,}s_k)^T \qquad (16)$$

Where, f is a real valued m-dimensional vector function. ICA involves to minimize the linearity among the data points and finds the covariance with the new independent data vectors. So that maximum independency is achieved.

In [53], A. Hyvarinen uses tanh non-linearity function to maximize the non-linearity among the data points. It involves finding the orthogonal matrix of the original data matrix and used Batch Covariance Algorithm to find the variance among the data points. Minimizing the dependency among the data provides the stronger optimization than the uncorrelated methods like PCA [33]. The main problem of ICA is to find the independent components that lead to high computational complexity. ICA features are identified by using several approaches like Infomax [52], Negentropy maximization [53], etc. [52] Maximizes the entropy of the data of the ICA framework, which are non linearly transformed. Negentropy maximization [53] extracts feature which has minimum dependence. Fast ICA is proposed in [54] to speed up the process of finding the independent components.

Generally, Independent Component Analysis (ICA) is extended into supervised technique for feature extraction by Conditional Independent Component Analysis (CICA)[55]. The Kullback-Leibler divergence between the joint and the product of marginal conditional distribution of the output is minimized and the CICA uses dual objective function which shows the information bottleneck method [56], which extracts the redundant data that preserve the class information maximally. Independent discriminant Component Analysis (IDCA)[57] and bilinear Discriminant Component Analysis(BDCA) [58] are some other approaches related to CICA. Discriminative ICA (dICA)[59] is a semi supervised approach proposed to improve the performance with other techniques such as PCA,LDA, ICA.

4.3 Multi Dimensional Scaling (MDS)

Multi Dimensional Scaling (MDS) is the collection of non-linear techniques to transform the high dimensional data into low dimensional data. The error between the pair wise distance between the low dimensional data and high dimensional data is expressed in stress function [60]. The examples of stress functions are raw stress function and Sommon cost function.

Let (x_i, x_j) and (yi, yj) be the high dimensional data points and low dimensional data points respectively. The raw stress function is defined by,

$$R(Y) = \sum_{i}^{j} (\left|\left|x_i - x_j\right|\right| - ||y_i - y_j||)^2$$

And the Sommon Cost function is defined by,

$$S(Y) = \frac{1}{\sum_{i,j} ||x_i - x_j||} \sum_{i \neq j} \frac{(\left|\left|x_i \quad x_j\right|\right| \quad ||y_i \quad y_j||)^2}{||x_i - x_j||}$$

Where, $||x_i-xj||$ denotes the Euclidean distance between the high dimensional data points and $||yi-y_j||$ denotes the Euclidean distance between the low dimensional data points.

Minimizing the stress function reduces the error which leads to improve the performance of the system. Eigen decomposition of a pair wise distance similarity matrix, Conjugate gradient method, Pseudo- Newton method are some of the methods to reduce the stress function[61].MDS is used in many applications like fMRI Analysis[62], molecular modeling[63],etc.,

5. EXPERIMENTAL RESULTS

Generally, most of the real world data are in the form of non linear. For this Analysis, we choose three high dimensional data like Insurance Benchmark, Spam and cancer datasets from UCI repository. Here Insurance dataset contains 85 attributes and 750 records, Spam contains 57 attributes and 4600 records and cancer dataset contains 57 attributes and 26 instances for the analysis. Analyzing and computing all these high dimensional data is very difficult and not all these variables affect the result of the classification. So, we tried to reduce the dimensionality by removing the irrelevant data for this analysis.

Though removing of attributes from the data takes less execution time, there may be a loss of data which may affect the accuracy of the classification. In this project, the dimensionality of the data has been reduced and the performance is measured in terms of both Accuracy and Elapsed time and the code was implemented by using Matlab.

5.1 Result of SVM Classification

For this paper we use these high dimensional data on SVM Classification and the classification Accuracy and Elapsed time are measured and the results are shown in the table 5.1

DATASET	ACCURACY	ELAPSED TIME (in sec)
Insurance Benchmark	39.4667	1.6659
Spam	35.5217	0.2643
Cancer	76.9231	1.2634

Table 5.1 Result of SVM Classification

To show the effectiveness of the dimensionality reduction, the high dimensional data is processed by the Linear Discriminant Analysis. From this analysis 85, 58, and 57 variables are transformed into 31, 30 and 38 on insurance, spam and cancer datasets respectively. Next these high dimensional data is process with the Principal Component Analysis that produce the better Results than the LDA method by giving 24, 11 and 7 respectively. ICA is performed on the high dimensional data which gives the results as 15, 2, and 7 for Insurance, spam and cancer datasets respectively which is better compared to other techniques and the dimensions of low dimensional data are shown in table 5.2

DATASET	No. of variables in HDD	Linear components	Principal Components	Independent Components
Insurance Benchmark	85	31	15	15
Spam	58	30	2	2
Cancer	57	38	15	7

Table 5.2 Dimensions of low dimensional data

The low dimensional data from the above techniques are again processed on the SVM classification and the Performance of the SVM classification calculated which proves the performance of SVM with low dimensional data is better than that of the SVM with high dimensional data and the result is shown in table 5.3.

DATASETS	LDA			PCA			ICA		
	Dimension	Accuracy	Time	Dimension	Accuracy	Time	Dimension	Accuracy	Time
Insurance (85)	24	52	1.54s	15	49.86	1.21s	15	58.93	1.65
Spam (57)	11	72.82	5.05s	2	70.08	1.14s	2	74.43	1.21
Cancer (58)	7	38.46	0.01s	15	61.53	0.09s	7	69.23	0.009

Table 5.3 Performance Analysis

CONCLUSION

In this paper, we present the various techniques to reduce the dimensions of the original data. From the survey, it comes to know that, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are the powerful techniques to handle the linear types of data and Independent Component Analysis (ICA), Support Vector Machine (SVM) and Multi Dimensional Scaling (MDS) are effectively worked on non linear data. But most of the real world data are in non linear form. So, the survey is concluded with the non linear techniques are efficient compared with the linear techniques.

REFERENCES

[1] R. Fisher, “The statistical utilization of multiple measurements,” Ann.Eugenics, vol. 8, no. 4, pp. 376–386, Aug. 1938.

[2] Zizhu Fan; Yong Xu; Zhang, D. ,"Local Linear Discriminant Analysis Framework Using Sample Neighbors", IEEE Transactions on Neural Networks, , On page(s): 1119 - 1132 Volume: 22, July 2011

[3] Quanxue Gao; Jingjing Liu; Hailin Zhang; Xinbo Gao; Kui Li, "Joint Global and Local Structure Discriminant Analysis",IEEE Transactions on Information Forensics and Security, page(s): 626 - 635 Volume: 8 April 2013

[4] Sharma, A. ; Paliwal, Kuldip K. ," Rotational Linear Discriminant Analysis Technique for Dimensionality Reduction", IEEE Transactions on Knowledge and Data Engineering, Year: 2008 , Page(s): 1336 - 1347
[5] Bin Zou; Luoqing Li; Zongben Xu; Tao Luo; Yuan Yan Tang , " Generalization Performance of Fisher Linear Discriminant Based on Markov Sampling " , IEEE Transactions on Neural Networks and Learning Systems, page(s): 288 - 300 Volume: 24, Issue: 2, February : 2013
[6] Yuxi Hou; Iickho Song; Hwang-Ki Min; Cheol Hoon Park , " Complexity-Reduced Scheme for Feature Extraction With Linear Discriminant Analysis " , IEEE Transactions on Neural Networks and Learning Systems, page(s): 1003 - 1009 Volume: 23, Issue: 6, June 2012
[7] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," Biostatistics, vol. 8, no. 1, pp. 86–100, 2007.
[8] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 8, pp. 982–994, Aug. 2004.
[9] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," Pattern Recognit., vol. 33, no. 10, pp. 1713–1726, 2000.
[10] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 1, pp. 4–13, Jan. 2005.
[11] H. Cevikalp, M. Neamtu, and M. Wilkes, " Discriminative common vector method with kernels," IEEE Trans. Neural Netw., vol. 17, no. 6, pp. 1550–1565, Nov. 2006.
[12] H. Park, M. Jeon, and J. B. Rosen, " Lower dimensional representation of text data based on centroids and least squares," BIT Numerical Math.,vol. 43, no. 2, pp. 427–448, 2003.
[13] M. Kyperountas, A. Tefas, and I. Pitas, " Weighted piecewise LDA for solving the small sample size problem in face verification," IEEE Trans.Neural Netw., vol. 18, no. 2, pp. 506–519, Mar. 2007.
[14] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, " Eigenfaces versus Fisherfaces: Recognition using class specific linear projection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711–720, Jul. 1997.
[15] Q. Ke and T. Kanade, "Robust subspace computation using L1 norm," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-03- 172, Aug. 2003.
[16] "Principal Components Analysis", 36-490, Spring 2010
[17] A. M. Martinez and A. C. Kak, "PCA versus LDA," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 2, pp. 228–233, Feb. 2001.
[18] N. Kwak , "Principal component analysis based on L1-norm maximization", IEEE Transaction on Pattern Analysis and Machine Intelligence, volume. 30, no. 9, pages.1672 -1680 Year : 2008
[19] Xu Chunming ; Jiang Haibo ; Yu Jianjiang ,"Robust two-dimensional principle component analysis ", IEEE transaction on signals and control Publication Year: 2010 , Page(s): 452 – 455
[20] A. Baccini, P. Besse, and A. D. Falguerolles, "A L1-norm PCA and a heuristic approach, ordinal and symbolic data analysis," in Ordinal and Symbolic Data Analysis, E. Diday, Y. Lechevalier, and P. Opitz, Eds. New York, NY, USA: Springer-Verlag, 1996, pp. 359–368.
[21] Q. Ke and T. Kanade, "Robust L1 Norm Factorization in the Presence of Outliers and Missing Data by Alternative Convex Programming," Proc. IEEE Conf. Computer Vision and Pattern Recognition, June 2005.
[22] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," IEEE Trans. Biomed. Eng., vol. 59, no. 3, pp. 653–662, Mar. 2012.
[23] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: Rotational Invariant L1- Norm Principal Component Analysis for Robust Subspace Factorization," Proc. 23rd Int'l Conf. Machine Learning, June 2006.
[24] Ran He; Bao-Gang Hu; Xiang-Wei Kong ,"Robust Principal Component Analysis Based on Maximum Correntropy Criterion", IEEE Transactions on Image Processing,On page(s): 1485 - 1494 Volume: 20, Issue: 6, June 2011
[25] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 696–710, Jul. 1997.

[26] S. Yu, J. Bi, and J. Ye, “Probabilistic interpretations and extensions for a family of 2D PCA-style algorithms,” in Proc. KDD Workshop Data Min. Using Matri. Tensors, Las Vegas, NV, Aug. 2008, pp. 1–7.

[27] J. Yang, D. Zhang, A. F. Frangi, and J. Yang, “2-D PCA: A new approach to appearance-based face representation and recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 1, pp. 131–137, Jan. 2004.

[28] J. Ye, “Generalized low rank approximations of matrices,” Mach. Learn.,vol. 61, nos. 1–3, pp. 167–191, 2005.

[29] J. Yang, D. Zhang, A. F. Frangi, and J. Yang, “2-D PCA: A new approachto appearance-based face representation and recognition,” IEEE Trans.Pattern Anal. Mach. Intell., vol. 26, no. 1, pp. 131–137, Jan. 2004.

[30] S. Yu, J. Bi, and J. Ye, “Probabilistic interpretations and extensions for a family of 2D PCA-style algorithms,” in Proc. KDD Workshop Data Min. Using Matri. Tensors, Las Vegas, NV, Aug. 2008, pp. 1–7.

[31] J. Zhao , P. Yu and J. Kwok "Bilinear probabilistic principal component analysis", IEEE Transaction on Neural Networks and Learning system., volume. 23, no. 3, pages.492 -503, publication year: 2012

[32] Q. Tao, G. Wu, and J. Wang, “The theoretical analysis of FDA and applications,” Pattern Recognition., vol. 39, no. 6, pp. 1199–1204, 2006.

[33] S. Moon and H. Qi ,"Hybrid dimensionality reduction method based on support vector machine and independent component analysis", IEEE Transaction on Neural Networks, volume. 23, pages.749 - 761 year 2012

[34] Q. Tao, D. Chu, and J. Wang, “Recursive support vector machines for dimensionality reduction,” IEEE Trans. Neural Netw., vol. 19, no. 1, pp. 189–193, Jan. 2008.

[35] Zoidi, O.; Tefas, A.; Pitas, I., "Multiplicative Update Rules for Concurrent Nonnegative Matrix Factorization and Maximum Margin classification", IEEE Transactions on Neural Networks and Learning Systems,, On page(s): 422 - 434 Volume: 24, Issue: 3, March 2013

[36] I. W.-H. Tsang, A. Kocsor, and J. T.-Y. Kwok, “Large-scale maximum margin discriminant analysis using core vector machines,” IEEE Trans. Neural Netw., vol. 19, no. 4, pp. 610–624, Apr. 2008.

[37] Fei Wang, Bin Zhao, and Changshui Zhang,” Unsupervised Large Margin Discriminative Projection”, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 22, NO. 9, SEPTEMBER 2011

[38] C. J. C. Burges, “Simplified support vector decision rules,” in Proc. 13thInt. Conf. Mach. Learn., 1996, pp. 71–77.

[39] E. Osuna and F. Girosi, “Reducing the run-time complexity of supportvector machines,” in Proc. Int. Conf. Pattern Recognit., Brisbane,Australia, Aug. 1998, pp. 1–10.

[40] D. Geebelen , J. A. K. Suykens and J. Vandewalle ,"Reducing the number of support vectors of SVM classifiers using the smoothed separable case approximation", IEEE Transaction on Neural Networks and Learning System, volume. 23, pages.682 -688, Publication year: 2012

[41] T. Downs, K. E. Gates, and A. Masters, “ Exact simplification of support vector solutions,” J. Mach. Learn. Res., vol. 2, pp. 293–297, Dec. 2001.

[42] S. S. Keerthi, O. Chapelle, and D. DeCoste, “ Building support vector machines with reduced classifier complexity,” J. Mach. Learn. Res., vol. 7, pp. 1493–1515, Jul. 2006.

[43] J. Lopez and J. R. Dorronsoro, “Simple proof of convergence of the SMO algorithm for different SVM variants,” IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 7, pp. 1142–1147, Jul. 2012.

[44] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to the SMO algorithm for SVM regression,” IEEE Trans. Neural Netw., vol. 11, no. 5, pp. 1188–1193, Sep. 2000.

[45] P. Chen, R. Fan, and C. Lin, “A study on SMO-type decomposition methods for support vector machines,” IEEE Trans. Neural Netw., vol. 17, no. 4, pp. 893–908, Jul. 2006.

[46] F. Cai and V. Cherkassky, “Generalized SMO algorithm for SVM-based multitask learning,” IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 6, pp. 997–1003, Jun. 2012.

[47] I. W. Tsang, J. T. Kwok, and P. M. Cheung, “Core vector machines:Fast SVM training on very large data sets,” J. Mach. Learn. Res., vol. 6,pp. 363–392, Apr. 2005.

[48] C. C. Chang , C. W. Hsu, and C. J. Lin, “The analysis of decomposition methods for support vector machines,” IEEE Trans. Neural Netw., vol. 11, no. 4, pp. 1003–1008, Jul. 2000.
[49] C. Lin, “On the convergence of the decomposition method for support vector machines,” IEEE Trans. Neural Netw., vol. 12, no. 6, pp. 1288–1298, Nov. 2001.
[50] J. Dong, A. Krzyzak, and C. Y. Suen, “Fast SVM training algorithm with decomposition on very large data sets,” IEEE Trans. Pattern Anal.Mach. Intell., vol. 27, no. 4, pp. 603–618, Apr. 2005.
[51] Di Wang; Hong Qiao; Bo Zhang; Min Wang "Online Support Vector Machine Based on Convex Hull Vertices Selection", Neural Networks and Learning Systems, IEEE Transactions on, On page(s): 593 - 609 April 2013
[52] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” Neural Comput., vol. 7, no. 6, pp. 1129–1159, Nov. 1995.
[53] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis. New York: Wiley, 2001.
[54] A. Hyvarinen and E. Oja, “Independent component analysis: Algorithms and applications,” Neural Netw., vol. 13, nos. 4–5, pp. 411–430, 2000
[55] S. Akaho, “Conditionally independent component analysis for supervised feature extraction,” Neurocomputing, vol. 49, nos. 1–4, pp. 139– 155, Dec. 2002
[56] N. Slonim and N. Tishby, “Agglomerative information bottleneck,” in Proc. Adv. Neural Process. Syst. 12, 2000, pp. 617–623.
[57] U. Amato, A. Antoniadis, and G. Gregoire, “Independent component discriminant analysis,” Int. J. Math., vol. 3, no. 7, pp. 735–753, 2003.
[58] M. Dyrholm, C. Christoforou, and L. C. Parra, “Bilinear discriminant component analysis,” J. Mach. Learn. Res., vol. 8, pp. 1097–1111, May 2007.
[59] S. Dhir and S.-Y. Lee, “ Discriminant independent component analysis “ , IEEE Transaction on Neural Networks, volume. 22, pages : 845 -857 2011
[60] T. Cox and M. Cox. Multidimensional scaling. Chapman & Hall, London, UK, 1994.
[61] T. Cox and M. Cox. Multidimensional scaling. Chapman & Hall, London, UK, 1994.
[62] J.B. Tenenbaum. Mapping a manifold of perceptual observations. In Advances in Neural Information Processing Systems, volume 10, pages 682–688, Cambridge, MA, USA, 1998. The MIT Press.
[63] J. Verbeek. Learning nonlinear image manifolds by global alignment of local linear models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(8):1236–1250, 2006.

AUTHORS:

Arunasakthi. K
KamatchiPriya. L