A Novel Dimension Reduction Technique based on Correlation Coefficient

Vinay Soni, Ritesh Joshi

Abstract - In this paper, a novel simple dimension reduction technique for classification is proposed based on correlation coefficient. Existing dimension reduction techniques like LDA is known for capturing the most discriminant features of the data in the projected space while PCA is known for preserving the most descriptive ones after projection. Our novel technique integrates correlation coefficient and method of elimination for feature selection to reduce the dimensionality of input space. Our approaches are novel because our method finds alternatives to LDA and PCA in a 2D parameter space. Extensive experiments are conducted on various datasets.

Index Terms—Classification, Correlation Coefficient, Dimension Reduction.

I. INTRODUCTION

Curse of dimensionality is an impediment for most computer vision applications as the search space grows exponentially with the data dimension. Principal component analysis (PCA) [1, 2] and linear discriminant analysis (LDA) [3, 4, 5] are both well known techniques for feature dimension reduction. LDA constructs the *most discriminant* features while PCA constructs the *most descriptive* features in the sense of packing most "energy". LDA plays a key role in many research areas in science and engineering such as face recognition [6, 7], image retrieval [8, 9], and bioinformatics [10]. LDA is a simple algorithm that is used for both dimension reduction and classification. In either case, LDA aims at maximizing class separability in a low dimensional space by selecting the feature vectors *w* which maximize

$$\frac{\left|w^T S_B w\right|}{\left|w S_w w^T\right|},$$

where B_S measures the variance between the class means, and W_S measures the variance of the samples in the same class. PCA is a useful statistical technique that has found various applications in many fields [1, 2]. It is considered as one of the simplest and best-known *Data Analysis* techniques.

Vinay Soni 1
Jawaharlal Institute of Technology, Borawan
soni.vinay1@gmail.com

Ritesh Joshi 2
Sr. Asst. Professor
Medicaps Institute of Technology and Management, Indore
riteshjoshi.indore@gmail.com

Its goal is to replace the original (numerical) variables with new numerical variables called "Principal Components" that have the following properties: (1) They can be ranked by decreasing order of "importance" (this term can be given a precise meaning). The first few most "important" Principal Components account for most of the information in the data. In other words, one may then discard the original data set, and replace it with a new data set with the same observations, but fewer variables, without throwing away too much information. (2) These new variables are uncorrelated.
In computer vision community, when comparing LDA with PCA, there is a tendency to prefer LDA to PCA, because, as intuition would suggest, the former deals directly with discrimination between classes, whereas the latter deals without paying particular attention to the underlying class structure. For example, when the data of each class can be represented by a single Gaussian distribution and share a common covariance matrix, LDA will outperform PCA. However, LDA as well as other discriminant analysis techniques are not guaranteed to work where the assumptions of the method do not hold. An interesting result is reported by Martinez and Kaka [11] that this is not always true in their study on face recognition. PCA might outperform LDA when the number of samples per class is small or when the training data non-uniformly sample the underlying distribution [12, 13].

Correlation Coefficient

The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables [14]. More generally, note that (Xi − X)(Yi − Y) is positive if and only if Xi and Yi lie on the same side of their respective means. Thus the correlation coefficient is positive if Xi and Yi tend to be simultaneously greater than, or simultaneously less than, their respective means. The correlation coefficient is negative if Xi and Yi tend to lie on opposite sides of their respective means. Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

The above formula defines the population correlation coefficient, commonly represented by the Greek letter ρ (rho). Substituting estimates of the covariances and variances based on a sample gives the sample correlation coefficient [5], commonly denoted r :

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

II. THE NOVEL DIMENSION REDUCTION TECHNIQUE

A. Algorithm

We have introduced new algorithm which is used as dimension reduction technique. New dimension reduction technique is based upon finding correlation among various attribute and on the basis of highly correlated attributes (redundant attributes) are eliminated from input space. Here it is assumed that the datasets used for deriving results contain linear data. The proposed algorithm is as follows –

Algorithm NewDimReduce (D)
Input: *Database D, database of input values;*
Output: *IndptAttribute;*
Begin
Let NoofAttributes: m;
Let R (m, m)
Let row number: n;
Let copy of database D in Matrix n×m of L;
For all row k of matrix R do
For columns j = k +1 to m of Matrix R do
{
R (k, j) = ∞
Sum1 = $\sum_{i=0}^{n}(L(i,k)-\mu_k)*(L(i,j)-\mu_j)$
Sum2= $\sum_{i=1}^{n}(L(i,k)-\mu_k)^2$
Sum3= $\sum_{i=1}^{n}(L(i,j)-\mu_j)^2$
R (k_ij)= $\frac{sum1}{\sqrt{sum2 * sum3}}$
}

```
Let indptAttribute (m);
indptAttribute (1) =1
For j =2 to m do
{
        If R (1, j) < Threshold then do
                IndptAttribute (j) =j
        Else
                IndptAttribute (j) = ∞
}
For each k = 2 to m
        For j = k + 1 to m
                If R (k, j) > Threshold Then
                If indpntAttrbt (k) = k and indpntAttrbt (j) = j Then
                        indpntAttrbt (j) = ∞
                End If
                End If
        Next
Next
Return (List of elements having not value ∞ of IndptAttribute);
```

Figure 1: New Proposed Algorithm for Dimension Reduction

B. Working Explained

Functioning of Proposed New Dimension reduction algorithm based on correlation coefficient

Calculation of correlation coefficient							
R	1	2	3	4	...	m	
1	-	R12	R13	R14	...	R1m	
2	-	-	R23	R24	...	R2m	
3	-	-	-	R34	...	R3m	
4	-	-	-	-	...	R4m	
...	...	...	...	...	...	...	
m	-	-	-	-	-	-	

Calculation of correlation coefficient for all possible combination of attributes

Figure 2: Working of Newly proposed dimension reduction algorithm

Selection of Independent Attributes based on Condition							
R	1	2	3	4	...	m	
1	-	R12	R13	R14	...	R1m	
2	-	-	R23	R24	...	R2m	
3	-	-	-	R34	...	R3m	
4	-	-	-	-	...	R4m	
...	...	...	...	...	...	...	
m	-	-	-	-	-	-	

Selection of independent attributes as having **R value < threshold** in the first row of the matrix

Figure 3: Working of Newly proposed dimension reduction algorithm

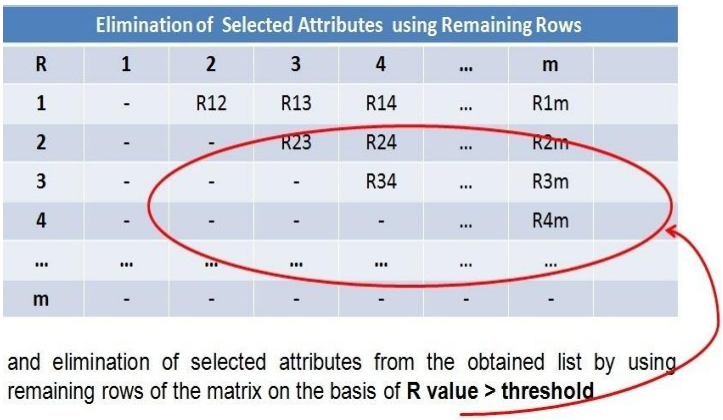

Figure 4: Working of Newly proposed dimension reduction algorithm

C. Time Complexity of Algorithm

This algorithm having time complexity of O (nm^2) and space complexity is 2((n+1) m+2) where n is number of instances and m is number of dimension.

III. EXPERIMENTAL RESULTS

The newly proposed algorithm is applied on numerical datasets like wine , glass identification and Breast Cancer Wisconsin used to reduce their dimensionality. There are 13 dimensions in Wine Dataset. After application of proposed algorithm remaining dimensions are as given figure 5.

Dimension Reduction				
Attribute No.	1	8	11	13

Figure 5: Dimension remaining after applying algorithm on Wine Dataset

In Glass Identification dataset number of dimension are 09, but after applying new dimension reduction algorithm, remaining dimension are as given in figure 6.

Dimension Reduction				
Attribute No.	1	6	8	9

Figure 6: Dimension remaining after applying algorithm on Glass Identification Dataset

In Breast Cancer Wisconsin dataset numbers of dimensions are 09, but after applying new dimension reduction algorithm, remaining dimension are as given in figure 7.

Dimension Reduction		
Attribute No.	1	9

Figure 7: Dimension remaining after applying algorithm on BCW Dataset

IV. CONCLUSION

REFERENCES

[1] I. T. Jolliffe, *Principal Component Analysis*. 2nd edition, New- York: Springer-Verlag, 2002.

[2] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks*, New York: Wiley, 1996.

[3] R. A. Fisher, "The use of multiple measurement in taxonomic problems," *Annals of Eugenics*, vol. 7, pp.179-188, 1936.

[4] R.A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, vol. 8, pp.376-386, 1938.

[5] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc., 2001.

[6] P. N. Belhummeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7): 711-720, 1997.

[7] K. Etemad and R. Chellapa, "Discriminant analysis for recognition of human face images," *Journal of Optical Society of American,* 14(8): 1724-1733, 1997.

[8] D. Swets and J. Weng, "Hierarchical discriminant analysis for image retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(5), 396-401, 1999.

[9] Y. Wu, Q. Tian, and T. S. Huang, "Discriminant EM algorithm with application to image retrieval," *Proc. of IEEE Conf. Computer Vision and Pattern Recognition,* South Carolina, June 13-15, 2000.

[10] W J. Ewans and G. R. Grant, *Statistical methods in bioinformatics*, Springer-Verlag, 2001.

[11] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 228-233, February 2001.

[12] R. Beveridge, K. She, B. Draper, and G. Givens, "A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition", *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Kauai, Hawaii, vol. 1, pp. 535-542, 2001.

[13] M. Zhu, A. M. Martinez, and H. Tan, "Template-based recognition of sitting postures," *Proc. IEEE Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, Madison, WI, 2003.

[14] Ervin Kregzig, "Advanced Engineering Mathematics", John Wiley, Edition 9, Illustrated.