ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

Principal Component Analysis - A Survey

Mr. S. Omprakash[1], S. Gokila[2]

Assistant Professor, Department of Computer Science,

Kovai Kalaimagal College of Arts & Science, Coimbatore[1]

M. Phil., Scholar, Kovai Kalaimagal College of Arts and Science, Coimbatore[2]

Abstract: Principal Component Analysis (PCA) is a multivariate procedure that investigates a data slab in which clarifications are designated by numerous inter-correlated measureable reliant variables. Its objective is to excerpt the imperative evidence from the table, to characterise it as a set of new orthogonal variables called principal components, and to show the outline of resemblance of the clarifications and of the variables as points in maps. The worth of the PCA model can be estimated using cross-validation procedures. Statistically, PCA be contingent upon the Eigen-decomposition of positive semi-definite matrices and upon the singular value decomposition of rectangular matrices. The number of principal components is less than or equivalent to the number of unique variables. It is a way of classifying patterns in data, and conveying the data in such a way as to highpoint their resemblances and modifications. This is the survey paper of the previous concept and procedures for PCA.

Keywords: Data mining, Principal Component Analysis (PCA)

I. INTRODUCTION

Principal Component Analysis (PCA) is probably the most popular multivariate statistical technique and it is used by almost all scientific disciplines. It is also likely to be the oldest multivariate technique. The number of principal components are less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components [3]. Principal components are guaranteed to be independent if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables.

Many papers of PCA are survey here and find that, no papers are combined with the techniques of PCA and the relevant example. That's why this type of paper is written where all these things are combined together which is easy to understand the beginner. PCA is a widely used mathematical tool for high dimension data analysis. Just within the fields of computer graphics and visualization alone, PCA has been used for face recognition [10], motion analysis and synthesis [7], clustering [4], dimension reduction [2], etc.

II. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. Principal Component Analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Principal components analysis is similar to another multivariate procedure called Factor Analysis. They are often confused and many scientists do not understand the difference between the two methods or what types of analyses they are each best suited.

A. Objectives of principal component analysis:

• PCA reduces attribute space from a larger number of variables to a smaller number of factors and as such is a "non-dependent" procedure (that is, it does not assume a dependent variable is specified).

• PCA is a dimensionality reduction or data compression method. The goal is dimension reduction and there is no guarantee that the dimensions are interpretable (a fact often not appreciated by (amateur) statisticians).

•To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

B. Properties of Principal Component

Technically, a principal component can be defined as a linear combination of optimally-weighted observed variables. The output of PCA are these principal components, the number of which is less than or equal to the number of original variables. Less, in case when we wish to discard or reduce the dimensions in our dataset. The PCs possess some useful properties which are listed below:

1. The PCs are essentially the linear combinations of the original variables, the weights vector in this combination is actually the eigenvector found which in turn satisfies the principle of least squares.
2. The PCs are orthogonal, as already discussed.
3. The variation present in the PCs decrease as we move from the 1st PC to the last one, hence the importance. The least important PCs are also sometimes useful in regression, outlier detection, etc.

C. Steps in PCA

Step 1: Normalize the data

First step is to normalize the data that we have so that PCA works properly. This is done by subtracting the respective means from the numbers in the respective column. So if we have two dimensions X and Y, all X become $\boldsymbol{x}$- and all Y become $\boldsymbol{y}$-. This produces a dataset whose mean is zero.

Step 2: Calculate the covariance matrix

Since the dataset we took is 2-dimensional, this will result in a 2x2 Covariance matrix.

Step 3: Calculate the eigenvalues and eigenvectors

Next step is to calculate the eigenvalues and eigenvectors for the covariance matrix. The same is possible because it is a square matrix. λ is an eigenvalue for a matrix A if it is a solution of the characteristic equation:

$$\det(\lambda I - A) = 0$$

Where, I is the identity matrix of the same dimension as A which is a required condition for the matrix subtraction as well in this case and 'det' is the determinant of the matrix. For each eigenvalue λ, a corresponding eigen-vector v, can be found by solving:

$$(\lambda I - A)v = 0$$

Step 4: Choosing components and forming a feature vector:

We order the eigen values from largest to smallest so that it gives us the components in order or significance.
Here comes the dimensionality reduction part. If we have a dataset with n variables, then we have the corresponding n eigen values and eigenvectors. It turns out that the eigenvector corresponding to the highest eigen value is the principal component of the dataset and it is our call as to how many eigenvalues we choose to proceed our analysis with. To reduce the dimensions, we choose the first p eigenvalues and ignore the rest. We do lose out some information in the process, but if the eigen values are small, we do not lose much.

Step 5: Forming Principal Components:

This is the final step where we actually form the principal components using all the math we did till here. For the same, we take the transpose of the feature vector and left-multiply it with the transpose of scaled version of original dataset.

$$\text{NewData} = \text{FeatureVector}^T \times \text{ScaledData}^T$$

Here,
NewData is the Matrix consisting of the principal components,
FeatureVector is the matrix we formed using the eigenvectors we chose to keep, and
ScaledData is the scaled version of original dataset
('T' in the superscript denotes transpose of a matrix which is formed by interchanging the rows to columns and vice versa. In particular, a 2x3 matrix has a transpose of size 3x2)

If we go back to the theory of Eigen values and eigenvectors, we see that, essentially, eigenvectors provide us with information about the patterns in the data. In particular, in the running example of 2-D set, if we plot the eigenvectors on the scatter plot of data, we find that the principal eigenvector (corresponding to the largest Eigen value) actually fits well with the data. it, does not carry much information and hence, we are at not much loss when deprecating it, hence reducing the dimension.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

III. LITERATURE REVIEW

In this Chapter we review the literature related to Principal Component Analysis (PCA) methods in brief. For better understanding we classify the literature into various categories. In this section, we present significant research carried out in PCA.

Principal components of CMB non-Gaussianity- The skew-spectrum statistic introduced by Munshi & Heavens has recently been used in studies of non-Gaussianity from diverse cosmological data sets including the detection of primary and secondary non-Gaussianity of cosmic microwave background (CMB) radiation. We describe how the bias induced in the estimation of primordial non-Gaussianity due to secondary non-Gaussianity may be evaluated for arbitrary primordial models using a PCA analysis. The PCA approach allows one to infer approximate (but generally accurate) constraints using CMB data sets on any reasonably smooth model by use of a look-up table and performing a simple computation. This principle is validated by computing constraints on the Dirac–Born–Infield spectrum using a PCA analysis of the standard templates.

Principle Component Analysis Based Cooperative Spectrum Sensing in Cognitive Radio – This paper introduces an improved cooperative spectrum sensing (CSS) algorithm for cognitive radio (CR) networks based on the machine learning technique. In this scheme, spectrum sensing consists of two steps. The offline training is performed by K-means clustering method and the threshold is defined by the classification result towards unsupervised data. As for the online classification stage, the similarity between the received signal and the cluster in training data is exploited for channel availability decision. The features of both the received signal and the training samples are extracted by principle component analysis (PCA). The fusion center will make final decision according to the local sensing result from differently sensors sequentially. Each sensor is responsible for the analysis towards certain period of the received signal. In this way, the initial sensing result (which will be updated due to the latest input signal) is available within short period of time.

Human Movements Separation Based on Principle Component Analysis – With more and more attention to terrorist attacks, rescue after disaster and medical treatments, the study on human motions has become a hot topic in recent years. Thanks to the unique mechanism of humans, the m-D signatures, which contain extensive information, of each segment are obviously distinct. It remains a great challenge to separate the movement of humans' each part. In this paper, a method for human movements separation based on a principle component analysis (PCA) is proposed. As one of the classical methods in the blind source separation problems, PCA decomposes the signal to a series of orthogonal basis functions to construct the Eigen subspace. The original signal can be represented by the linear combination of the orthogonal basis functions. In addition, information criterion is utilized to determine the minimal number of output for PCA.

Jointly Informative and Manifold Structure Representative Sampling Based Active Learning for Remote Sensing Image Classification – Active Learning (AL) methods that select unlabeled samples only querying by informative measures (i.e., uncertainty and/or diversity criteria) have been extensively investigated. However, these methods usually do not exploit the manifold structure of the unlabeled data from the geometrical point of view, a choice that might lead to a sample bias and consequently undesirable performances. To control and possibly overcome such drawbacks, this paper explores AL methods based on joint informative and manifold structure representative sampling (JI-MSRS). In JI-MSRS, a portion of the unlabeled samples that are added at each iteration is selected according to the informative measures, whereas another portion is selected according to their capability to represent the data cluster structure. Four popular manifold learning methods, namely, principle component analysis (PCA), linear discriminant analysis, kernel PCA, and neighbourhood preserving embedding, are used to model the data structure.

An Events Rearrangement Strategy-Based Robust Principle Component Analysis – Random noise in seismic data can affect the performance of reservoir characterization and interpretation, which makes denoising become an essential procedure. This letter focuses on suppressing random noise in post stack seismic data while preserving the edges of desired signals. Due to the lateral continuity of seismic data, polynomial fitting (PF) method can be a good alternative in attenuating random noise. However, discontinuities exist widely in post stack seismic data, which might be damaged by the PF filter. By contrast, principle component analysis (PCA)-based filters have better performance in edge preserving, but there appear artifacts in the demonised results using the PCA-based filters. Thus, we propose an edge-preserving polynomial PCA filter which combines advantages of the PF and PCA methods by optimizing a PCA problem with a weighted polynomial constraint.

A Probabilistic Approach to Outdoor Localization Using Clustering and Principal Component Transformations – A probabilistic approach f o r outdoor location estimation using GSM received signal strength (RSS) from base stations (BSs) is presented. The proposed approach first divides the region of interest into different clusters based on deviations

from the path loss mode lf or each RSS component. In each cluster, the proposed algorithm uses principal component analysis (PCA) to intelligently transform RSS into new uncorrelated dimensions. This retains accuracy by not losing the substantial RSS correlations in each cluster, b u t also accommodates the different RSS distributions in each cluster. Our experiments are conducted in a real GSM outdoor environment. The proposed approach is compared with a traditional probabilistic algorithm f o r three different area partitioning methods. The experimental results show that the positioning accuracy is significantly improved and our clustering scheme gives good support f o r location estimation.

IV. CONCLUSION AND FUTURE WORK

In this work, we reviewed the state-of-the-art of PCA literature, which include multiple PCA methods. Subsequently, we discussed how these PCA methods address the problems faced by classical PCA. In this work, we focus on study of PCA methods. In subsequent we propose a common framework for PCA methods and identify issues to be addressed, we bring out some novel PCA methods, which alleviate the problems faced by both the existing PCA methods and classical PCA methods, we establish general properties of PCA methods by performing a theoretical analysis and we extend our feature partitioning ideas to cluster analysis and subspace classification. Principal Component Analysis are useful as data reduction but not for understanding the structure of the data. In future work, we use to proposing a framework which brings the existing PCA methods under a common framework and identify the issues need to be addressed in this framework.

REFERENCES

[1]. Donough Regan and Dipak “Principal components of CMB non-Gaussianity”, Royal Astronomical Society, 2015.
[2]. Xin Chen, Fukang Hon, Hai Huang, Xiaojun Jing, “Principle Component Analysis Based Cooperative Spectrum Sensing in Cognitive Radio”, IEEE, 2016.
[3]. Xiaoran Shi, Feng Zhou, Mingliang Tao, Zijing Zhang,, “Human Movements Separation Based on Principle Component Analysis,” IEEE Sensors Journal, 2017.
[4]. Alim Samat, Paolo Gamba,Fellow, Sicong Liu,, “Jointly Informative and Manifold Structure Representative Sampling Based Active Learning for Remote Sensing Image Classification,” IEEE Journals, 2017.
[5]. Yuchen Wang, Wenkai Lu, and Benfeng Wang, “An Events Rearrangement Strategy-Based Robust Principle Component Analysis,” IEEE Journals, 2017.
[6]. Kejiong Li, John Bigham, Laurissa Tokarchuk and Eliane L. Bodanese, “A Probabilistic Approach to Outdoor Localization Using Clustering and Principal Component Transformations,” IEEE Journals, 2013.
[7]. [7] H. Hotelling, (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417–441, and 498–520.
[8]. S. Huang, M. O. Ward, E. A. Rundensteiner. Exploration of dimensionality reduction for text visualization. In CMV '05: Proceedings of the Coordinated and Multiple Views in Exploratory Visualization, pages 63-74, Washington, DC, USA, 2005. IEEE Computer Society.
[9]. I. T. Jollie, Principal Component Analysis. Springer, second edition, 2002.
[10]. Y. Koren and L. Carmel. Visualization of labeled data using linear transformations. Info Vis, 00:16, 2003.
[11]. Kyungnam Kim, Face Recognition using Principle Component Analysis.
[12]. K. Pearson, (1901). On Lines and Planes of Closest Fit to Systems of Points in Space (PDF). Philosophical Magazine 2 (11): 559–572. Doi: 10.1080/14786440109462720.
[13]. A. Safonova, J. K. Hodgins, and N. S. Pollard. Synthesizing physically realistic human motion in low dimensional, behavior-specic spaces. ACM Trans. Graph., 23(3):514-521, 2004.
[14]. Jon Shlens, A tutorial on Principal Component Analysis, 25 March, 2003.
[15]. Lindsay I Smith, A tutorial on Principal Component Analysis, February 26, 2002.
[16]. M.A. Turk and A.P. Pentland, Face Recognition Using Eigen faces, IEEE Conf. on Computer Vision and Pattern Recognition, pp. 586-591, 1991.