OVERVIEW ON PRINCIPAL COMPONENT ANALYSIS ALGORITHM IN MACHINE LEARNING

Article *in* international research journal of science and technology · October 2020

CITATIONS
5

READS
2,261

1 author:

Karunakar Pothuganti
Electrogenics
34 PUBLICATIONS **315** CITATIONS

SEE PROFILE

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:10/October -2020 Impact Factor- 5.354 www.irjmets.com

OVERVIEW ON PRINCIPAL COMPONENT ANALYSIS ALGORITHM IN MACHINE LEARNING

Swathi P *1, Dr. Karunakar Pothuganti*2

*1Lecturer, Dept. of Computer Science, Sree Chaitanya degree college, Karimnagar,Telangana.

*2Department of R &D, Electrogenics, Karimnagar, India.

ABSTRACT

In this paper, we have assessed a calculation utilizing Principal Component Analysis (PCA) for its application in information analysis. In the examination field, it is hard to comprehend the enormous measure of information and is very tedious as well. This way, to maintain a strategic distance from wastage of time and for the simplicity in understanding, we have investigated a PCA calculation that can diminish the immense element of the information into 2-dimensional. The technique for PCA is utilized to pack the most extreme measure of data into initial two segments of the changed network known as the principal components by dismissing the other vectors that convey the insignificant data or repetitive information. The primary target of the paper is to isolate two mixes state A and B having various focuses for every one of the four sensors also, distinguishes which sensors have the comparative or unique focus with the assistance of different plots that clarifies the connection between's the various factors.

Keywords: PCA, Data Analysis, Eigen values.

I. INTRODUCTION

In Principal Component Analysis (PCA) is one of the design acknowledgement techniques and one of its applications is to investigate the high dimensional information that is not anything but difficult to comprehend by only taking a gander at the enormous measure of information. For information analysis, I have to decrease the high element of the information into low measurement and afterwards making a plot and decipher the results. PCA is utilized to introduce the essential data into scarcely any straightforward plots in particular score plot and stacking plot. In the field of examination, it is tough to break down an enormous measure of information. PCA calculation is utilized to figure the connection between the tremendous corresponded informational index [1]. In linear algebra, PCA has its numerical calculation that clarifies the connection between's the information containing the factors as sections and perceptions or tests as lines. The objective of the PCA calculation is to diminish the enormous connected factors into a modest number of factors[2]. These connected factors are called principal components [3]. The primary thought process is to build up a network that contains the most significant measure of data in the initial two sections and at that point venture the information utilizing 2-dimensional plot in MATLAB programming.

II. ALGORITHM

We have considered an algorithm in which the different variables have diverse correlated variables and the principal objective is to isolate two different mixes state An and Bhave considered an algorithm in which the different variables have diverse correlated variables and the fundamental objective is to isolate two different mixes state An and B had various fixations for four sensors[3]. The progression by step PCA algorithm given in fig 1 is actualized in MATLAB programming.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:10/October -2020 Impact Factor- 5.354 www.irjmets.com

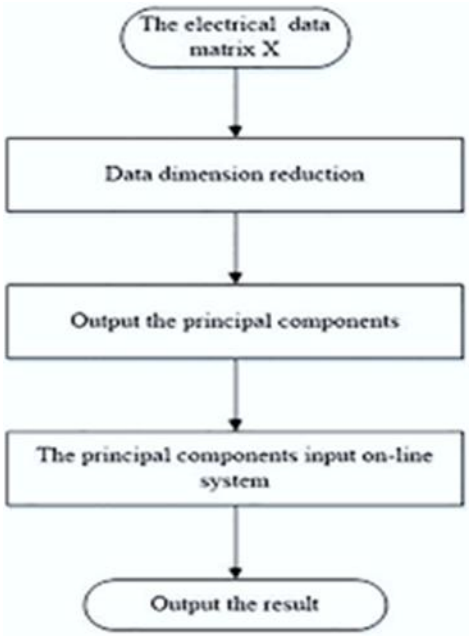

Fig-1: Flow chart for PCA Algorithm

The means of the PCA Algorithm are given underneath

We start with the data set 'A' which is a matrix of measurement m x n, where m columns speak to the variables while n segments speak to the examples for example perceptions. We will presently linearly change this matrix into another matrix 'B' of a similar measurement m x n, so that for some matrix Z given by condition (1).

$$B=Z*A \qquad (1)$$

Normalization is the significant aspect of the algorithm in which we have to figure the mean of the first data matrix and take away off the mean for discovering principal components as given in condition (2).

$$Mean(m) = 1/N \sum_{n=1}^{N} A[m,n] \qquad (2)$$

Compute Covariance matrix of A, which will be of dimension m x m in condition (3). Here, every component of the covariance matrix CA speaks to all conceivable pair of covariance[4]. Indeed all slanting components speak to fluctuation, and the non-slanting components of the matrix are covariance.

$$C_{A=}A * \frac{A^T}{(n-1)} \qquad (3)$$

We are needed to choose a few highlights that the changed matrix 'B' should show which identifies with the highlights of relating covariance matrix C_B. It thought to have the least covariance and most significant fluctuation[4]. Little change might be excess data. Accordingly, we have to augment the fluctuation and limit the covariance. The Eigenvalues are organized in the sliding request since the most significant value of Eigenvalue tells the family member significance of the principal comparing component as appeared in fig 2.

III. DATA ANALYSIS

The objective of our administrative work is to examine the data of dimension 11x4 as given in the matrix beneath in Table I. We have four unique sensors having various fixations state C1, C2, C3...and C11. There are two mixes state An and B, and we have taken four unique focuses for Compound An and seven unique fixations for Compound B. Our goal is to isolate two unique mixes relying upon their value of focuses.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:10/October -2020 Impact Factor- 5.354 www.irjmets.com

Table: I Original data used for data Analysis

Different Compounds Concentration		Four Different Sensors			
		Sensor1	Sensor2	Sensor3	Sensor4
Compound A	C1	503	58	23	42
	C2	675	59	39	65
	C3	429	35	33	49
	C4	163	18	17	5
Compound B	C5	639	17	47	21
	C6	105	1	23	9
	C7	106	7	17	5
	C8	118	11	17	3
	C9	110	2	18	4
	C10	636	65	19	9
	C11	313	26	21	10

The subsequent advance is to assess the standardized matrix of dimension 11x4 by taking away the mean from the first matrix of dimension 11x4 as given underneath in condition [5]. The data is standardized, so we can undoubtedly compute the difference.

$$\begin{bmatrix} 159.2727 & 29.0909 & -2.0909 & 21.8182 \\ 328.2727 & 32.0909 & 13.9091 & 43.8182 \\ 83.2727 & 7.0909 & 7.9091 & 28.8182 \\ -182.7273 & -7.9091 & -9.0909 & -15.1818 \\ 294.2727 & -10.9091 & 20.9091 & 0.8182 \\ -240.7273 & -24.9091 & -2.0909 & -10.1818 \\ -237.7273 & -19.9091 & -1.0909 & -15.1818 \\ -227.7273 & -15.9091 & -8.0909 & -17.1818 \\ -235.7273 & -24.9091 & -8.0909 & -15.1818 \\ 291.2727 & 37.0909 & -7.0909 & -11.1818 \\ -31.7273 & -0.9091 & 5.0909 & -11.1818 \end{bmatrix}$$

The third step gives the diminished covariance matrix of dimension 4x4 that can likewise be computed by just increasing the first data matrix with the render of the unique data matrix given underneath in condition.

$$\begin{bmatrix} 5.6779 & 0.4505 & 0.1531 & 0.3392 \\ 0.4505 & 0.0537 & 0.0043 & 0.0316 \\ 0.1531 & 0.0043 & 0.0099 & 0.0138 \\ 0.3392 & 0.0316 & 0.0138 & 0.0457 \end{bmatrix}$$

The Eigen vector-matrix computed as given underneath in condition that shows the connection between the uncorrelated variables. The slanting matrix is done and arranged in the diminishing request in condition [6]. The slanting values are the Eigenvalues that are removed from the matrix put in a segment, and every Eigenvalue conveys the amount of the difference is contained by the principal components and organize it in the diminishing request.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:10/October -2020 Impact Factor- 5.354 www.irjmets.com

$$\begin{bmatrix} 0.9947 & -0.0873 & -0.0316 & 0.0441 \\ 0.0792 & 0.3857 & 0.8030 & -0.4473 \\ 0.0268 & 0.0553 & -0.5078 & -0.8593 \\ 0.0598 & 0.9168 & -0.3103 & 0.2442 \end{bmatrix}$$

$$\begin{bmatrix} 5.7383 & 0 & 0 & 0 \\ 0 & 0.0275 & 0 & 0 \\ 0 & 0 & 0.0211 & 0 \\ 0 & 0 & 0 & 0.0004 \end{bmatrix}$$

The last advance of the algorithm gives the figuring of last score matrix in the condition that has the most extreme data contained in the initial two sections known as principal components PC1 and PC2 that are organized as indicated by their measure of fluctuation in the diminishing request.

$$\begin{bmatrix} 161.9812 & 17.2050 & 12.6208 & 1.1337 \\ 332.0678 & 24.6636 & -5.2609 & -1.1351 \\ 85.3280 & 22.3238 & -9.8944 & 0.7406 \\ -183.5372 & -1.5209 & 8.7494 & -0.4123 \\ 292.4600 & -27.9905 & -28.9309 & 0.0825 \\ -242.0897 & 1.9557 & -8.1754 & -0.1590 \\ -238.9818 & -0.9061 & -3.2115 & -4.3435 \\ -229.0252 & -2.4569 & 3.8598 & -0.1653 \\ -237.5759 & -3.3966 & -3.7354 & 3.9959 \\ 291.8089 & -21.7631 & 27.6515 & -0.3893 \\ -32.4361 & -8.1140 & 6.3269 & 0.6519 \end{bmatrix}$$

What is more, thus, the last two segments state PC3 and PC4 can be ignored that has a modest quantity of data that can be repetitive [8].

IV. RESULTS AND DISCUSSION

The data is investigated utilizing PCA algorithm in which we can decipher that out of four sensors, and one sensor is having less relationship when contrasted with other three sensors as appeared in fig 2.

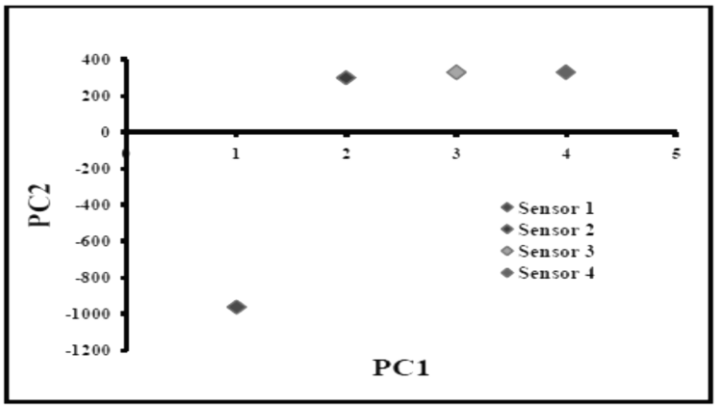

Fig-2 loading plot for four different sensors

As should be evident that sensor 2, sensor three and sensor 4 are near one another and have very high correlation when contrasted with sensor 1. The sensor 1 is on the negative side of the beginning have a negative correlation. Too, the sensors on a similar side of starting point are having comparable compound focuses[9]. We have distinguished the variables that are near one another having an exceptionally high correlation and additionally, the variables that are far separated from each speaking to the negative correlation. In fig 3, we have plotted the cylindrical graph for the Eigenvalues known as the Eigenvalue range that gives a connection between the Eigenvalues and Eigenvector number[10]. Eigenvector number is the all outnumber of Eigenvalues that are four in number for the given data matrix[11]. All Eigenvalues are more prominent than one.

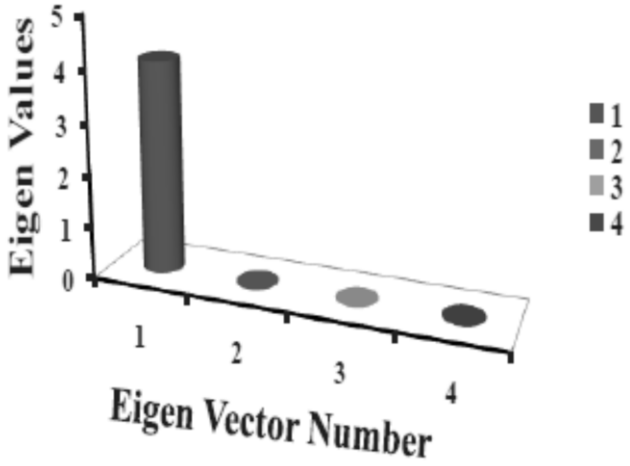

Fig-3: Eigenvalue Spectrum

We can see that the groupings of Compound A are having their fixations separated from one another for C1, C2, C3 and C4 clarifying that the variables have extraordinary fixations and they are disparate[12]. While, the meagre few centralizations of Compound B for C6, C7 and C9 are appearing comparative pattern as appeared in Table I. What is more. Likewise, C5 and C10 are giving a similar correlation.

V. CONCLUSION

We proposed an algorithm that is utilized to decrease the dimension of the first data completed for data analysis from 11-dimensions to the 2-dimensional data set. The objective of the algorithm is to restrict the most extreme data just in the first two sections called as principal components and disregard the rest of the sections conveying the immaterial measure of data. To decrease the dimensionality of the data, I have utilized PCA that gives better outcomes. Additionally, I can have an away from the correlation between the various variables at the point when they are spoken to in the 2D plot in MATLAB programming.

VI. REFERENCES

[1] Stojanovic, Branka, and Aleksandar Neskovic. "Impact of PCA based fingerprint compression on matching performance." In Telecommunications Forum (TELFOR), 2012 20th, pp. 693-696. IEEE, 2012.

[2] Vishal Dineshkumar Soni. (2018). ROLE OF AI IN INDUSTRY IN EMERGENCY SERVICES. International Engineering Journal For Research & Development, 3(2), 6. https://doi.org/10.17605/OSF.IO/C67BM

[3] Pothuganti Karunakar. et al,2020. “Analysis of Position Based Routing Vanet Protocols using Ns2 Simulator”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-9 Issue-5, March 2020, DOI: 10.35940/ijitee.E2717.039520.

[4] Ankit Narendrakumar Soni (2018). Application and Analysis of Transfer Learning-Survey. International Journal of Scientific Research and Engineering Development, 1(2), 272-278.

[5] Saporta G and Niang N. 2006. Correspondence analysis and classification. In: Greenacre M, Blasius J, eds. Multiple Correspondence Analysis and Related Methods. Boca Raton, FL: Chapman & Hall. 371-392.

[6] Ankit Narendrakumar Soni (2018). Image Segmentation Using Simultaneous Localization and Mapping Algorithm. International Journal of Scientific Research and Engineering Development, 1(2), 279-282.

[7] Dray S. 2008. On the number of principal components: a test of dimensionality based on measurements of similarity between matrices. Comput Stat Data Anal. 52: 2228-2237.

[8] Vishal Dineshkumar Soni. (2018). Prediction of Geniunity of News using advanced Machine Learning and Natural Language processing Algorithms. International Journal of Innovative Research in Science Engineering and Technology, 7(5), 6349-6354. doi:10.15680/IJIRSET.2018.0705232

[9] Bell, Anthony and Sejnowski, Terry. (1997) “The Independent Components of Natural Scenes are Edge Filters.” Vision Research 37(23), 3327-3338.

[10] Ankit Narendrakumar Soni (2018). Data Center Monitoring using an Improved Faster Regional Convolutional Neural Network. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 7(4), 1849-1853. doi:10.15662/IJAREEIE.2018.0704058

[11] PCA and LDA in DCT domain ,Weilong Chen, Meng Joo Er *, Shiqian Wu Pattern Recognition Letters 26 (2005) 2474–2482

[12] Vishal Dineshkumar Soni. (2018). Artificial Cognition for Human-robot Interaction. International Journal on Integrated Education, 1(1), 49-53. https://doi.org/10.31149/ijie.v1i1.482