January - February 2020
ISSN: 0193 - 4120 Page No. 14546 - 14552

Dimensionality Reduction in Machine Learning Technique using Principal Componenet Analysis

S.Umadevi[1], S. NirmalaSugirtha Rajini[2],A. Punitha[3],Viji Vinod[4]
1. Research Scholar, Department of Computer Applications,Mail-Id : umakalyani70@gmail.com
2. Professor, Department of Computer Applications, Mail-Id: sugirtharaja77@yahoo.com
4. Professor, Department of Computer Applications, Mail-Id: vijivino@gmail.com
Dr. M.G.R Educational and Research Institute, Chennai,
3. Assistant Professor, Department of Computer Applications, Mail-Id: apunithaganesh@yahoo.com
Queen Mary's college, Chennai,

Article Info
Volume 82
Page Number: 14546 - 14552
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 28 February 2020*

Abstract:
Machine learning plays a vital role in today's world. In the internet data searching will be abundant. These data has to be trained to the system in order to perform the work of machine learning. In machine learning the main parameter that is the size of the data plays a vital role in the execution time. In other words we can say that the execution time is directly proportional to the volume of data. The data that are going to use for our research is been collected from a private production organization. When the analysis of the organization is done the datasets that is been collected in enormous in used. To simply the work and to get accurate production results dimensional reduction method is followed. For data reduction the Principal Component Analysis (PCA) technique is used. This technique can be used to reduce the size of the data that we will execute when finding out the production of an organization. The dimensionality reduction helps in finding out the exact volume of data that is been present in the datasets and reduce its size and helps in giving the exact accuracy results. In this paper we have analyze the concept of Principal Component Analysis along with clustering algorithm.

Keywords: *Principal Component Analysis, Dimensional Reduction, Clustering, Localization, Boundary Scan.*

I INTRODUCTION

Machine learning is generally referred as obtaining particular data that need from huge collection of data and training the system to work according to our proposed framework. If the execution time is very high the accuracy in getting the results will be affected. So this execution time has to be maintained low in order to get the correct accuracy in end results. Doing a detailed research in finding the relationship between the data is a major challenging in data science field. The structure of each and every data is been analyzed in all possible ways. Adata table consisting of all parameter data is mandatory for reference. Maintain the entire structure of all data it will be very easy to predict the size of the data. By implementing this it is very easy to perform dimension reduction of data before starting the execution[1]. Each and every data is been counted and clustered according to our need which will reflect in the output dimension of data. In order to perform dimension reduction of data the concept of Principal Component Analysis (PCA) is used. PCA is generally referred as one of the most important method in the field of machine learning because of its nature of reducing the size of the data. PCA will help in reducingthe dimensionality of the dataset by two ways [2]. The first way it collectsall the dimensions of all dataand encodes the important data alone according to our usage. Secondly it removes the least priority data from the data sheet.

January - February 2020
ISSN: 0193 - 4120 Page No. 14546 - 14552

By following these two steps it will be very easy to do the dimension reduction of data which will results in the less execution time. Followingthis method the data utilizes less space, which results in the classification of larger data in a small amount of time.PCA concentrates only on the dimension which will be more helpful in fetching out the needed data in less execution time.One of the worst problem in machine learning is that the dimension of data is not been taken into account anywhere which will results in more execution time. In reality when the dimension reached more than the expected size the execution time will be increased which results in the less accuracy in mining of data [3].

II RELATED WORKS

Chen-Fu et. al., proposed about the dimensionality reduction in the form of graphs. TheLancos algorithm is usedto reduce the dimension of data. Compute graphs are been used to find out the dimension reduction of each and every data. The concept of multiparty and Bi-party is been involved which indicates the most and least data and their dimensions. The past Lanczos algorithm is expanded to provide the accurate set of data. This model is not much efficient and time consuming[1].

CédricTraizetet. al., proposed the dimensionality reduction is not only involved in the data extraction, it also gives the extent towards the image system.. At each set of part the dimensionality reduction takes place but the efficient of the image is not obtained as per our expected needs. The concept of k means clustering is used which is time consuming and data loss will be very high [2].

Wenbo Yu et. al., proposedthe dimensionality reduction in the Hyperspectalmethod which requires a large storage space. The dimensionality reduction is high because of the exact localization of the data from the sets. So the space requirement and the accurate extraction will be very high resulting in low execution time[3].

Wei Wang et. al., proposed the less density of data is low impact in considering or to extract. To avoid this dimensionality is made to high to make to data to be intense. As the high dimensionality is implemented the space in the server gets consumed large area. Noise parity check method is executed which will not give best results n data reduction[4].

Honglei Zhang et. al., proposed the two methods PCA and the LDA are the main in the dimensionality reduction. This combination can reduce the dimensions of the vector in the matrix set. The edge predicting is not provides the better result. The PCA and the LDA can face several problems. The PCA is not getting expanded in the sets label. The LDA can be used in the label sets but the data extraction form the label set is not much effective. This method is time consuming and gives less accuracy in dimensional reduction[5].

III RESEARCH METHODOLOGY

A. Clustering

Cluster analysis is the foremost important analysis in reducing the size of the data. Clustering is the process in which the input data are been grouped. This grouping will be carried on between the important and non important data. So when we do this grouping the quality of data will be improved.

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

For every i, set

$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

For each j, set

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

}

-- (1)

To eliminate the unwanted databy this clustering method and only the needed data are been taken into account. So when doing this as first step the output data is given as the input for the PCA. Thus resulting in a good mining work where only the needed datagot as input.

January - February 2020
ISSN: 0193 - 4120 Page No. 14546 - 14552

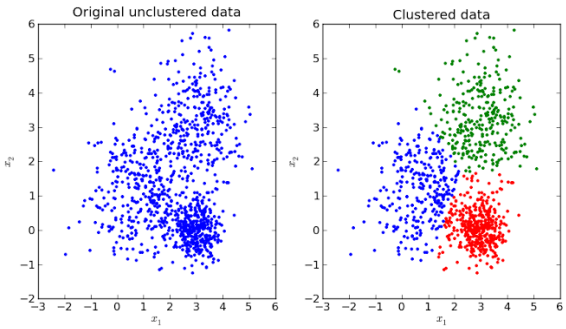

Fig.1. Clustering of Data

B. Principal Component Analysis

Principal component analysis (PCA) is a mathematical work in which the number of possibly correlated variables are been spitted into a smaller number of uncorrelated variables which is generally termed as principal components in data science. The principal component checks the possibility and availability of all data whose dimension has to be reduced. For each and every application the work of principal component varies. Usually it will be termed as cosine transform n the mathematical formulation.

1. Compute the sample covariance matrix

Σ =bn −1 Pni=1(Xi − Xn)(Xi − Xn) T.

2. Compute the eigenvalues λ1 ≥ λ2 ≥ · · · and eigenvectors e1, e2, . . . , of Σ.

3. Choose a dimension k.

4. Define the dimension reduced data

Zi = Tk(Xi) = X + Pk j=1 βijej where βij = hXi − X, eji

Now in case of data analysis this principal component is mainly used as a tool in reduction of size of data. Here the concept of Eigen values and matrix multiplication is carried out.The results that is been obtained is termed as component score in data analysis. PCA is a true and good Eigen vector calculation that is been most commonly used in data analysis. This PCA will help in localization of data that we need to be mined and fixes the exact location in each and every computation. So, to follow these stepsit's very easy to locate the data and process the data which will result in its dimension reduction. The following diagram represents the principal component analysis which shows matrix multiplication of data.

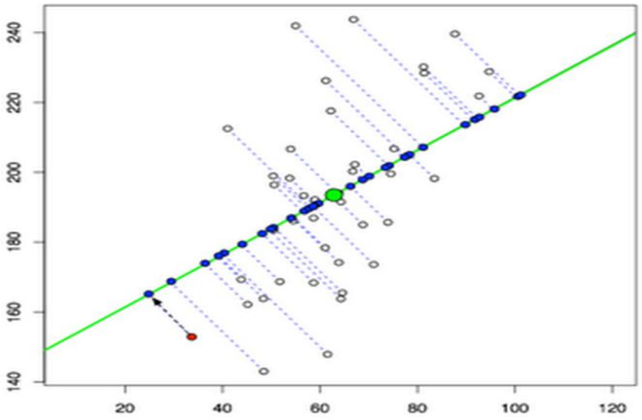

Fig.2. Principal Component Analysis

C. Principal Components (PC)

Principal component is generally defined as combination of optimally weighted variables and non weighted variables present in the data sheet from where we need to make our dimension reduction. Generally there will be two types of components that is been extracted. The first component extracted in a principal component analysis will give the results of the wanted data that we are going to mine and the second component extracted will be the data that we do not require at any cost. Actually both the components which we have derived will be irrelevant to each other either in size or number of data.This PCA will give two results of the covariance correlation of both needed and unneeded data. Eigen values of both the results are been calculated and matrix multiplication is carried out [14-16].

Mean is represented by following equation

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

(2)

D. Dataset Description

Here the production company data is used. All the production company data is tabulated in the attributes such as fresh milk, detergents, frozen and so on. The data are tabulated as 81 Rows with 3592 Columns. All samples are included same number of attributes.

January - February 2020
ISSN: 0193 - 4120 Page No. 14546 - 14552

IV PROPOSED DESIGN

In this proposed system we are reducing the dimension of datawith the help of PCA. The unwanted attributes are been reduced in order reduce the dimension of data. The following steps are been carried out in PCA.

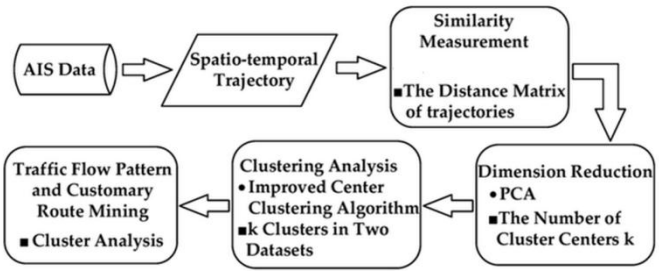

Fig.3: Block Diagram of Dimensionality Reduction

The following steps are the step by step process representing dimensional reduction of data using PCA.

• Data acquisition- Here the entire data are been collected from the Bench mark data sets. This is the first step involved.

• Data pre-processing- Needed attributes and un-needed attributes are been classified.

• Each attribute is been standardized and variance and standard deviation formula is been calculated.

Variance is defined as

$$\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n-\mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}\right\}^2=\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 \tag{3}$$

Where S is data covariance defined as

$$\mathbf{S}=\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\overline{\mathbf{x}})(\mathbf{x}_n-\overline{\mathbf{x}})^{\mathrm{T}}. \tag{4}$$

Standard Deviation is calculated from below

The vector

$$\vec{\bar{x}}=\frac{1}{n}\left(\vec{x}_1+\vec{x}_2\ldots+\vec{x}_n\right)=\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_p\end{bmatrix}$$

is called the sample mean vector

note

$$\bar{x}_i=\frac{1}{n}\left(x_{i1}+x_{i2}+\ldots+x_{in}\right)=\frac{1}{n}\left(\sum_{j=1}^{n}x_{ij}\right) \tag{5}$$

• Correlation of each attribute is calculated.

$$r=\frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) \tag{6}$$

Where sx is mean of attribute trained data 1

sy is mean of attribute trained data 2

•Eigen values of each attributes is been calculated in order to produce principal components (PC).

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}x_{ij}e_j-\mu\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}x_{ij}e_j\right)^2 && \leftarrow \mu=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}x_{ij}e_j\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}x_{ij}e_j\right)\left(\sum_{a=1}^{d}x_{ia}e_a\right) && -\sum_{j=1}^{d}\left(\frac{1}{n}\sum_{i=1}^{n}x_{ij}\right)e_j\\
&=\sum_{a=1}^{d}\sum_{j=1}^{d}\left(\frac{1}{n}\sum_{i=1}^{n}x_{ia}x_{ij}\right)e_je_a\\
&=\sum_{a=1}^{d}\left(\sum_{j=1}^{d}\mathrm{cov}(a,j)e_j\right)e_a && \leftarrow \mathrm{cov}(a,j)=\frac{1}{n}\sum_{i=1}^{n}x_{ia}x_{ij}\\
&=\sum_{a=1}^{d}(\lambda e_a)e_a && \leftarrow \sum_{j=1}^{d}\mathrm{cov}(a,j)e_j=\lambda e_a \quad \text{e is an eigenvector of the covariance matrix}\\
&=\lambda\|e\|^2=\lambda
\end{aligned} \tag{7}$$

• Find principal component coefficients are been formulated in order to find out the variance of each attributes.

$$\pi_{\mathbf{u}}(\mathbf{x})\equiv\langle\mathbf{u},\mathbf{x}\rangle\Rightarrow\mathrm{var}[\pi_{\mathbf{u}}]=\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u} \tag{8}$$

Where $\sum u$ is the covariance matrix

• Calculate eigenvectors is been done in each and every steps.

• Building classification model of PCA is been executed at the final stage.

V RESULTS AND DISCUSSIONS

Figure 4 shows the results obtained using classical PCA on an isometric nonlinear manifold, in the shape of an S-Curve, with 1000 data points. The

January - February 2020
ISSN: 0193 - 4120 Page No. 14546 - 14552

MDS embedding appears not to be as good as the other embeddings because MDS is not able to take nonlinearities into account, and the Euclidean distance may not be as good a distance measure if the data lies on a nonlinear manifold. However, if the number of data points, N, is set very small, then the results for MDS appears less distorted compared to the nonlinear dimensionality reduction algorithms. Figure 4 for the results of using the dimensionality reduction algorithms on the S-Curve manifold, this time with only 200 data points. We can conclude from this that MDS may be better than the other algorithms for sparse data sets, but is not able to accurately capture the nonlinearities if the original data lies on a nonlinear manifold. Although PCA, Isomap and HPCAare all able to reduce the dimension of the nonlinear isometric manifold, they require the data points to be relatively dense.

	0	1	2	3	4	5	6	7	8	9	10
0	1000025	5	1	1	1	2	1	3	1	1	2
1	1002945	5	4	4	5	7	10	3	2	1	2
2	1015425	3	1	1	1	2	2	3	1	1	2
3	1016277	6	8	8	1	3	4	3	7	1	2
4	1017023	4	1	1	3	2	1	3	1	1	2
5	1017122	8	10	10	8	7	10	9	7	1	4
6	1018099	1	1	1	1	2	10	3	1	1	2
7	1018561	2	1	2	1	2	1	3	1	1	2
8	1033078	2	1	1	1	2	1	1	1	5	2
9	1033078	4	2	1	1	2	1	2	1	1	2
10	1035283	1	1	1	1	1	1	3	1	1	2
11	1036172	2	1	1	1	2	1	2	1	1	2
12	1041801	5	3	3	3	2	3	4	4	1	4
13	1043999	1	1	1	1	2	3	3	1	1	2
14	1044572	8	7	5	10	7	9	5	5	4	4

Fig. 4. Reduced Production Data Sets

The figure 2 shows the results of PCA algorithm. The results of sum of squared error distance and Execution Time of corresponding clusters are been clearly put in the figure 5. The Figure 6 shows the relation between SSE and Number of clusters. It finally shows that when number of clusters increases the sum of squared error distance values decreases and vice versa. TheFigure 7shows the ExecutionTime and No of Clusters. It shows that if the number of clusters increases the Execution time increases also increases and vice versa.

PCA			
Dataset	*No of Clusters*	*SSE*	*Execution Time(in ms)*
Reduced Dataset	1	12604	0.513
	2	9513	0.631
	3	8075	0.689
	4	7641	0.777
	5	1959	0.862

Fig.5. Results of Number of Clusters, SSE and Execution Time

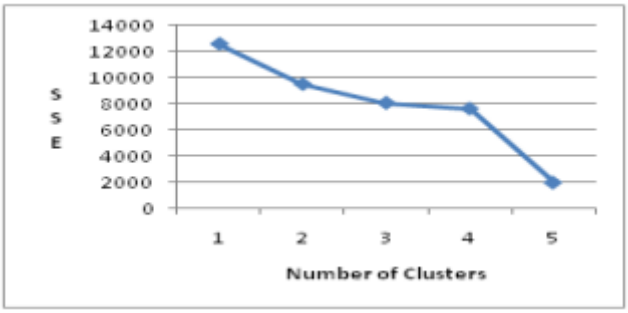

Fig. 6. Shows SSE and Number of Clusters

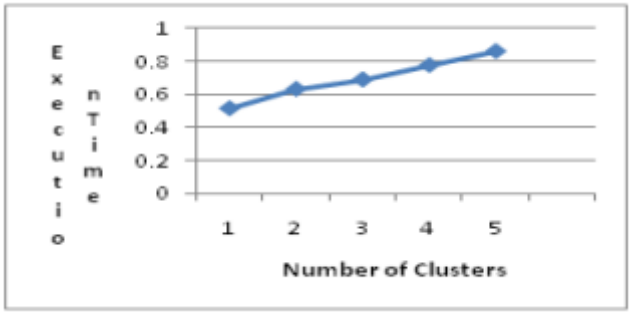

Fig.7. Shows Execution Time and Number of Clusters

VI CONCLUSION

The Multi-dimensionality-Systems decrease systems for information investigation has been assessed in this work. A rundown of some current direct and nonlinear dimensionality decrease strategies has been introduced. PCA is helpful in

recognizing huge directions and straight relationships in high dimensional information. PCA just as old style MDS are inadmissible if the informational collection contains nonlinear connections among the factors. General MDS procedures are fitting when the information is exceptionally nonmetric or inadequate. On the off chance that the first high-dimensional informational collection contains nonlinear connections, at that point nonlinear dimensionality decrease systems are increasingly fitting.

For fundamental kinds of nonlinear manifolds, an assessment was performed on different dimensionality decrease procedures. Since the nonlinear dimensionality decrease procedures rely upon a decent neighborhood, the outcomes when utilizing various neighborhoods are altogether unique. A decent neighborhood will rely upon the specific attributes of the informational index. On the off chance that the area is excessively little, the worldwide structure may not be caught successfully. In the event that the area is too enormous, the nonlinearities of the informational index may not be mapped suitably. Since a "decent" neighborhood is dependent on the informational collection, it is hard to sum up a worthy worth that will work under all conditions.

VII REFERENCES

[1]. Viktoria Koscinski&Chen-Fu Chiang(2018), “Dimensionality Reduction of the Complete Bipartite Graph with K Edges Removed for Quantum Walks”, IEEE 2018.

[2]. Jordi Inglanda ; Cédric Traizet,“Temporal Dimensionality Reduction for Land Cover Map Production Using High Resolution Image Time Series”.

[3]. Wenbo Yu , Miao Zhang &Yi Shen(2018),”Learning a Stable Local Manifold Representation for HyperspectralLinear DimensionalityReduction”,IEEE.

[4]. Wei Wang , Wei-guo Shen , Ya-xin Sun , Bin Chen & Rong Zhu(2018),“Dimensionality reduction via adjusting data distribution density”, IEEE.

[5]. Honglei Zhang &Mancef Gabbouj(2018),”Feature Dimensionality Reduction with Graph Embedding and Generalized Hamming Distance”, IEEE..

[6]. Lu Liang , Yi Xia , Lina Xun , Qing Yan & Dexiang Zhang(2018), “Class-Probability Based Semi-Supervised Dimensionality Reduction for Hyperspectral Images”,IEEE.

[7]. BastianSeifert , KatharinaKorn ,SteffenHartmann & Christian Uhl(2018) , “Dynamical Component Analysis (DYCA): Dimensionality Reduction for High-Dimensional DeterministicTime-Series”,IEEE.

[8]. Yujing Zhang , Ying Qiao , Zongxiang Lu & Wei Wang(2018),” New Dimensionality Reduction Method of Wind Power Curve Based on Deep Learning”,IEEE.

[9]. Yumeng Liu & Aina Sui(2018),” Research on Feature Dimensionality Reduction in Content Based Public Cultural Video Retrieval”,IEEE.

[10]. Xiangrong Zhang , Yaru Han , Ning Huyan , Chen Li ; Jie Feng , Li Gao &Xiaoxiao Ma(2018) ,“Partial-Spectral Graph-Based Nonlinear Embedding Dimensionality Reduction for Hyperspectral Image Classificaiton”,IEEE

[11]. RojaSahu, Rajesh Kumar Sahoo, Shakti KetanPrusty, Pratap Kumar Sahu. "Urinary Tract Infection and its Management." Systematic Reviews in Pharmacy 10.1 (2019), 42-48. Print. doi:10.5530/srp.2019.1.7

[12]. Majeed, A.S. Eco-friendly design of flow injection system for the determination of bismarck brown R dye (2018) International Journal of Pharmaceutical Research, 10 (3), pp. 399-408.

[13]. Shabna e., shriharisubhashri r., sridevi r., monisha p., kavimani s. (2018) role of certain drugs in cardiotoxicity induction for the experimental purpose. Journal of Critical Reviews, 5 (4), 1-5. doi:10.22159/jcr.2018v5i4.26326

January - February 2020
ISSN: 0193 - 4120 Page No. 14546 - 14552

[14]. Ritter, V.C., Nordli, H., Fekete, O.R. and Bonsaksen, T., 2017. User satisfaction and its associated factors among members of a Norwegian clubhouse for persons with mental illness. *International Journal of Psychosocial Rehabilitation. Vol 22 (1) 5, 14.*

[15]. Ferrazzi, P., 2018. From the Discipline of Law, a Frontier for Psychiatric Rehabilitation. *International Journal of Psychosocial Rehabilitation, Vol 22*(1) 16, 28.

[16]. Bornmann, B.A. and Jagatic, G., 2018. Transforming Group Treatment in Acute Psychiatry: The CPA Model. *International Journal of Psychosocial Rehabilitation, Vol 22*(1) 29, 45.