International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 1468 – 1474

Dimension Reduction in Big Data Environment-A Survey

Saheli Ghosh
Computer Science & Engineering Department
Techno India University, West Bengal
Saltlake, Kolkata-91, India
sahelighosh92@gmail.com

Mainak Sen
Computer Science & Engineering Department
Techno India University, West Bengal
Saltlake, Kolkata-91, India
mainaksen.1988@gmail.com

Abstract—Relational database management system is able to tackle data set which is structured in some way and by means of querying to the system user gets certain answer. But if the data set itself does not lie under any sort of structure, it is generally very tedious job for user to get answer to certain query. This is the new challenge coming out for the last decade to the scientists, researchers, industrialists and this new form of data is termed as big data. Parallel computation not only from the concept of hardware, but different application dependent software is now being developed to tackle this new data set for solving the challenges generally attached with large data set such as data curation, search, querying, storage etc. Information sensing devices, RFID readers, cloud storage now days are making data set to grow in an increasing manner. The goal of big data analytics is to help industry and organizations to take intelligent decisions by analyzing huge number of transactions that remain untouched till today by conventional business intelligent systems. As the size of dataset grows large also with redundancy, software and people need to analyze only useful information for particular application and this newly reduced dataset are useful compare to noisy and large data.

Keywords-*big data; dimension reduction; feature extraction; feature selection; PCA; ICA.*

*****

I. INTRODUCTION

Big data offers huge improvement over traditional data storage or warehousing technologies. However big data also comes with few challenges that must be solved before its benefits can be fully leveraged since big data stores unstructured data. This system provides mechanism for data storage or retrieval, SQL querying and a "framework" for data mining and analysis. It must be noted that a data scientist must write an analysis program using the provided framework for data mining purpose. Personal information of the owner can be accessed by unethical IT specialists without getting the authorization from them. When system is design to receive bulk data, security measures should be built in to the system which checks the source of the data for authenticity unfortunately most system lack this basic security measures. As we know Big Data stores a huge amount of data on a daily basis, so we need to reduce the dimension of the datasets as per as the requirements of the application that can be done by feature selection and feature extraction.

The main motive of the data mining is to find some useful patterns or structures from the unstructured data and extract the patterns and explanation of the data those are random and noisy in nature.

In a data matrix columns are correlated and rows are not correlated to each other where n is the number of columns represents total number of observation and p is the number. of rows represent the total number of variables in the samples. Ideally n should be much greater than p (n>>p) but when p (number of variable) is very much larger than the n (number of observations) then this problem needs a solution and the solution is dimension reduction.

The six tasks that have been done by data mining are

1. Anomaly detection: Also known as outlier detection. It will be needed in terms of data analysis. We can detect the outlier by box plot.
2. Association rule learning: It defines some relationships that is unique among the variables. For example, when a customer buys video equipment, he or she will also buys another electronic gadget and relate the presence of a set of items with another range of values for another set of variables.
3. Clustering: A set of events or items that can be divided or segmented into similar sets of items or events or divide the observation into same homogeneous group for better analysis purpose. For example, we can group a huge amount of treatment data on the basis of side effects those are similar to each other. Also, we can divide a group of people those are interested to buy a new product.
4. Classification: Data mining can parcel the information so that diverse classes or classifications can be distinguished in light of blends of parameters. For instance, clients in a market can be ordered into discount seeking shoppers
   - shoppers in a rush
   - loyal regular shoppers
   - shoppers attached to name brands
   - Infrequent shoppers.

   This classification may be used in different analyses of customer buying transactions as a post mining activity.
5. Regression: motive is to find out how each and every variable is related to each other. It is a statistical process and two factors are present in this process- one is response and another one is predictor. For example, rainfall and production are two factors. If rainfall is considered as response then a prediction can be done based on the response factor that how much production will be done.

6. ***Summarization:*** main motive is to find a section of data or subset that holds the most info of the whole data set.

II. CHARACTERISTICS OF BIG DATA

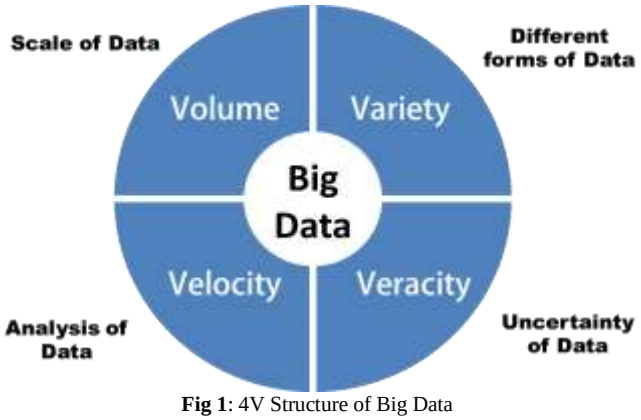

Fig 1: 4V Structure of Big Data

John Mashey made the term popular since its first appearance from 1990. In 2001, Gartner defined Volume, Velocity and Variety the three main characteristics of big data and in 2012, a new V known as Veracity was added to the existing one and later came to known as 4 V's [1]. Two new characteristics that have been introduced in big data are variability and value also known as 6V's structure and now a day a new characteristic that define 7V's structure is visualization.

A. Volume

Big data is all about a huge amount of data. Now a day in the online network a big scale of the data is available. A tremendous amount of users are present in that online network like a huge amount of users in Facebook, active Twitter users and bags of emails. An enormous amount of data is generated in every second from all this users' action.

B. Velocity

Velocity is enormous amount of Incoming data that demands to be processed. The number of SMS messages, face book comments, updates, and debit or credit card details, twitter activities, call details that have been received by a specific telecom company every second is humongous. Amazon Web Services Kinesis is a streaming application and also an illustration of the application that manages the velocity of data.

C. Variety

Now a day data generated from different sources are mostly unstructured. Variety is different types of data that is getting generated. Various tools and well defined techniques have been introduced to load the raw data. Different types of tools and analysis techniques are being used for various type of data.

D. Veracity

Veracity refers to the efficiency of the data degree to which we can use it for a decision making. Looking at the wrong data may drive incorrect decision and unnoticed errors may snowball over a period of time. So organizations demand to check the correctness of the incoming data as well as the analysis that have been performed over the data.

E. Variability

Variability refers the data with multiple meaning where meaning of the data is changing repeatedly. For this reason process of data managing and handling is damaged. Past meaning of the data will be replaced by the present meaning over the time. For example a coffee shop sells different types of coffee and someone takes the same flavor of coffee every day and the taste of the coffee is different each and every day then this is known as variability.

F. Value

The most important characteristic of the big data is value. Having access in big data is worthless until we make it into useful meaning. Now a day companies are started to generate the useful value from the big data. It is all about how business value is derived from the data. For example if a shopkeeper can find out a meaningful relationship among two products then he/she can convince the customer to buy the product or put them next to each other.

G. Visualization

Visualization is one of the most important characteristics in big data. A huge amount of data can be visualized by graphs, histogram, charts etc. rather than the spreadsheets or two or three dimensional visualizations, formulas for a better human understanding because of two attributes variety and velocity.

III. DIMENSION REDUCTION

Dimension reduction basically is a process of selecting a subset of features to be used in the model. It belongs to the preprocessing of the data sets in the data science project. It can be useful for both supervised and unsupervised learning problems. It is the process of reducing a number of variables features in review. Through this dimension reduction procedure, variables those are random in nature can be reduced under consideration by getting a set of principal variables. Discovering the axis of the data is the goal of the dimension reduction procedure. We need dimension reduction to find out hidden correlations, remove the noisy and redundant features, visualization and interpretation of the data, store and process the data easily. Dimension reduction can be divided into two subcategories- feature selection and feature extraction. Feature selection and Feature extraction both are the pre-processing steps for dimension reduction to improve the performance. Feature selection includes wrapper, filters and hybrid method, on other hand feature extraction includes Principal component analysis, Kernel principal component Analysis and Independent component analysis.

For example,

$$a+b+c+d = e$$

$$ab = a + b . \quad (1)$$

In equation (1) making representation of two variables into one represents feature extraction, is used to reduce the number of variables.

$$ab + c + d = e. \quad (2)$$

Now, if c = 0 or arbitrarily a small number it wouldn't be relevant therefore it could be taken out of the equation by doing so we would be using feature selection.

$$ab + d = e. \quad (3)$$

Through this we are selecting only the relevant variables and leaving out the irrelevant one.The main motivation of the dimension reduction is reducing the complexity of the classifier, reduce the extraction cost and survey class distinctness.

 ISSN: 2321-8169
1468 – 1474

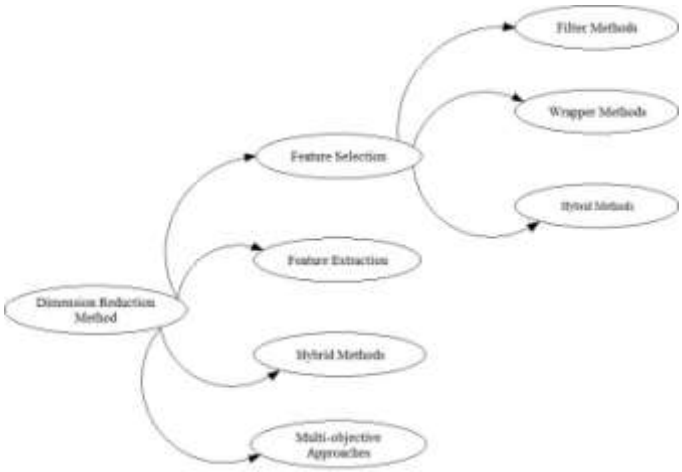

Fig2: Dimension reduction method

The dimension reduction problem in Data Mining is basically the big data reduction as the enormous accumulation of big data streams presents the 'scourge of dimensionality' with a large number of elements (factors and measurements) that builds the capacity and computational unpredictability of big data frameworks. There are many wide ranges of dimension reduction approaches [2]. Popular are DQC (Dynamic Quantum Clustering), BIGQuic, OFS (Online Feature Selection), Corsets, LHNFCSF (Linguistic hedges fuzzy classifier), Tensors, FH (Feature Hashing), Incremental partial least squares (IPLSs).

A. DQC

It empowers capable perception of high-dimensional big data. It plots subsets of the information on the premise of thickness among the majority of the connected factors in high-dimensional component space. The DQC is versatile to expansive frameworks due to its support for exceptionally disseminated information in parallel situations. The DQC depends on quantum mechanics methods from material science. It works by building a potential intermediary capacity to evaluate the thickness of information focuses.

B. BIGQuic

Customary dimensionality decrease calculations that utilize Gaussian most extreme probability estimator proved unable handle the datasets with more than 20,000 factors. The BIGQuic addresses the issue by applying a parallel gap and overcome technique that can be connected up to 1 million variables in the element space for dimensionality decrease [3].The outcomes demonstrate that the proposed calculation is profoundly adaptable and speedier than the current calculations, for example, Glasso and ALM [4, 5].

C. OFS

It is an approach where the online learners just work on little and settled length highlight sets. Be that as it may, the choice of dynamic components for exact expectation is a key issue in the methodologies presented in [6].

D. Corsets

Applying undergarments to lessen enormous information and high importance when utilized for information.

E. LHNFCSF

It is an Etymological supports fluffy classifier. Data reduction is the strength of the procedure and weakness is lack of efficiency.

F. Classifier

Classifier preparing with in significant highlight spaces. Strength is diagrams basic component measurements and satisfactory testing size and weakness is assumptions or suspicions should be more exact to diagram basic include measurement highlight space.

G. Tensors

It can be described as Tensor deterioration and guess. Computational complexity is very much high in this process.

H. FH

The component hashing (FH) technique diminishes include dimensionality by haphazardly doling out each element in the real space to another measurement in a lower-dimensional space [7]. This is finished by essentially hashing the ID of the unique elements. Normally, all dimensional diminishment techniques corrupt the information quality. Nonetheless, the vast majority of them save the geometric characteristics of the information.

I. Incremental partial least squares (IPLSs)

Incremental fractional minimum squares (IPLSs) are a variation of the fractional minimum squares strategy that adequately lessens measurements from extensive scale spilling information and enhances the order exactness. The expert postured calculation works in two-arrange include extraction prepare. To begin with, the IPLS embraces the objective capacity to refresh the verifiable means and to extricate the main projection heading. Second, the IPLS figures whatever is left of projection bearings that depend on the proportionality between the Krylov grouping and the halfway minimum square vectors [8].

IV. FEATURE SELECTION

The measure of high-dimensional information that exists and is accessible to all on the web has incredibly expanded in the recent years. In this manner, machine learning techniques experience issues in managing the huge number of information highlights, which is representing a fascinating test for analysts. Keeping in mind the end goal to utilize machine learning strategies adequately, preprocessing of the information is basic. Include choice is a standout amongst the most regular and imperative procedures in information preprocessing, furthermore, has turned into a vital segment of the machine learning process [9]. It is otherwise called variable choice, property choice, or variable subset choice in machine learning and measurements. It is the way toward recognizing important elements and expelling unessential, excess, or boisterous information. This procedure accelerates information mining calculations, enhances prescient precision, and expands understandability. Superfluous components are those that give no valuable data, and repetitive components give no more data than the present chosen highlights. It is a NP- hard problem.
On demand of some specific criteria, features those are not relevant or excess in the whole set are dismissed by selecting

 ISSN: 2321-8169
1468 – 1474

some features those are the subset of the original one through the feature selection procedure. In feature selection feature redundancy can be defined by a relationship between two features those values are correlated to each other. Pertinence of the features can be achieved by improved prescient precision of the classifier and righteous of cluster through segregating capacity of the feature.

Four steps need to be executed for feature selection

- Subset generation
- Subset evolution
- Stopping criteria
- Result validation

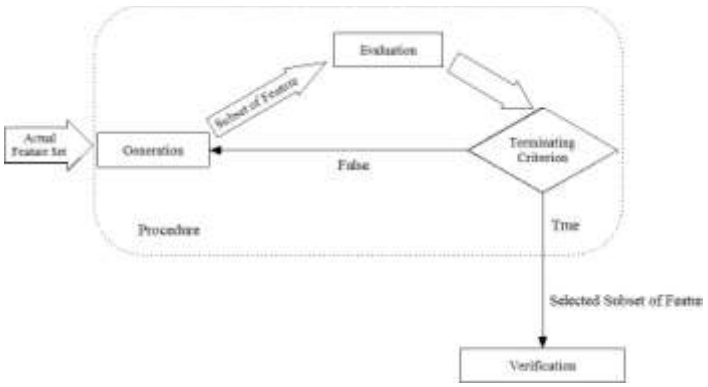

Fig 3: Feature Selection Process

A. Subset Generation

Subset generation is nothing but a pursuit technique that used to produce subset of the original features under some specific search approach and forward them to the subset evolution. Different types of search strategies are present in this step like heuristic, complete, sequential, and random. Subset generation and evolution are repeating process till a guaranteed stopping criterion is fulfilled.

In heuristic search in light of some past knowledge upcoming subset will be found example is genetic algorithm computational complexity is high in this search approach. In complete search under some specific criteria an optimal result is found that can be used through backtracking example is Branch and Bound. Some variations like sequential forward selection (SFS), sequential backward elimination (SBE), bi-directional selection are present in the sequential search process. It is a greedy hill climbing approach. Main motive of randomly search is to select a subset randomly example is Las Vegas algorithm.

B. Subset Evolution

In the first step subset generation a subset is generated and sent to the subset evolution stage in this stage newly generated subset will be evaluated under some evolution condition.

Evolution conditions can be divided into three types

- Filter model
- Wrapper model
- Hybrid model

B. AFILTER MODEL

In filter methods without involving a data mining algorithm a feature subset will be selected by intrinsic characteristics of the data set. Some conditions those are used in this method are distances, information, consistency and dependency.

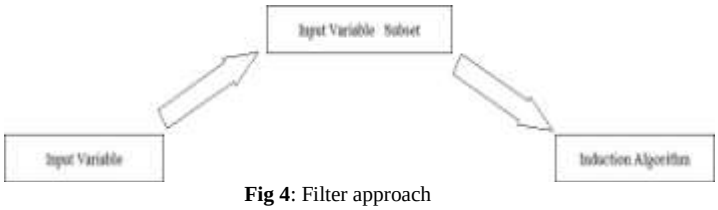

Fig 4: Filter approach

B.B WRAPPER MODEL

It finds out the best feature set by using some predetermined algorithm. Though this method gives the guarantee of a best result but it's not a good idea to use this method for the huge dataset because it is very much expensive computationally. Predictive accuracy is used in this method.

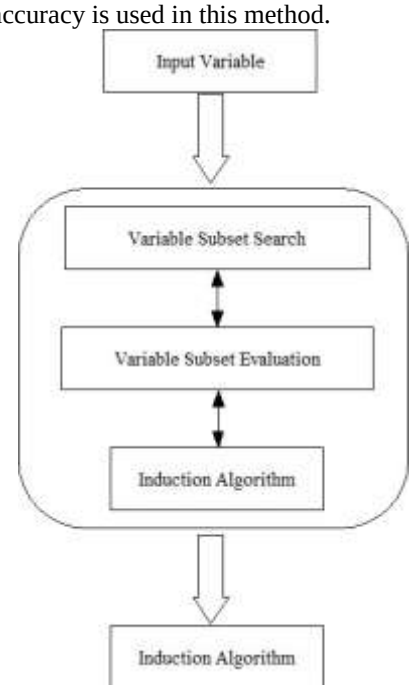

Fig 5: Wrapper method

B.C HYBRID MODEL

These methods are only applicable on the high dimensional datasets with some mining algorithm those are predefined and intrinsic characteristic of data set. It takes all the advantages of the filter and wrapper method by combining both methods.

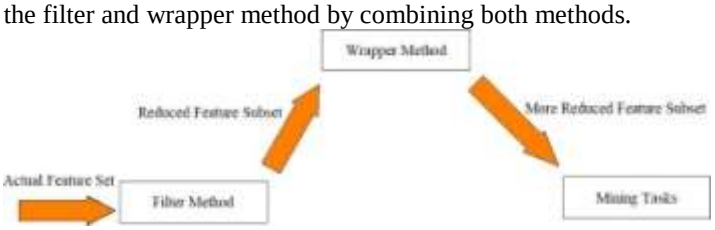

Fig 6: Hybrid method

C. STOPPING CRITERION

A stopping criterion method helps to decide when and where the feature selection process ought to stop. Some conditions those are used in this method frequently are

- Search is complete or not
- A better subset is not delivered due to Consequent expansion or cancellation of any component or feature.

- If classification error rate of a subset is less than the authorized error rate then the subset will be selected as a quality subset.

D. Result Validation

Result validation can be done by determining the result directly with the help of some previous knowledge over the data also need some knowledge present in the feature set those are redundant and not relevant at all. Another way for result validation is compare the previously generated result.

V. FEATURE EXTRACTION

Feature extraction is another type of dimension reduction method that works differently; in this method an entirely new set of variables will be created to represent the data. It's a linear combination of original attributes. Every new variable or feature is a function of original of the original data sets. The main two tasks done by the feature extraction are dimensionality reduction and transform the input vector into the feature vector. This method reduces the features. High dimensional information typically requires parcel of memory and power utilization. In FE method the extensive quantities of attributes are consolidated in view of the calculations utilized and they are changed over into bring down dimensional space. ICA and PCA both are the most used methods in feature extraction. Through the parameter extraction and classification extraction can be led autonomously or together.

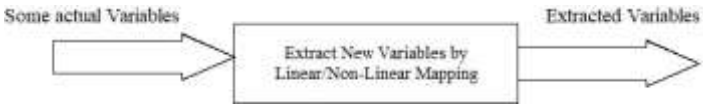

Fig 7: Feature Extraction Method

A. Principal component analysis(PCA)

PCA or principal component analysis is a dimension reduction technique and a descriptive method which emphasizes variation and eliminates multi co-linearity. Suppose we have 1000 of variables and by doing principal component analysis. We end with 1000 of PCAs then by calculating Eigen vector and Eigen value we can decide how many components we can keep and by keeping a fewer number of components we may be able to explain most of the variation present in the datasets. It mainly used for unsupervised learning problems. In this strategy, given an informational collection of perceptions on corresponded factors or variables, an orthogonal change is performed to change over it into an arrangement of uncorrelated factors or variables called the foremost parts or principal component. The quantity of original variables is not exactly or equivalent to the quantity of principal components. One dependable guideline is to consider those components or segments whose fluctuations are more prominent than one in the lessened space. Essential segments or components are ensured to be autonomous just if the factors are mutually regularly dispersed. PCA is not linear regression.

In face recognition, Principal Component Analysis [10] follows some steps to recognize a face, at first an image will be taken from the database (ORL database specifically used for face storage) then calculate the Eigen vector and Eigen value, and a set of the weights will be computed by projecting input image onto each of the Eigen value in face space in the next step comparing the weight of the input image to face space. From a determination it can be stated that input image is actually an image of a face, if the input image classifies as the face in previous step then in the next step we determine if the input image matches to the image that exist to the database. In the final step if it is an unknown face, add to the database and recomputed the Eigen face similarly we need to do this do this for the next image. Images vary in items of (1) lighting (2) expression (3) details. Time Complexity PCA Algorithm is O (p^2n+p^3) where n is defined as number of data points and the number of dimensions is p.

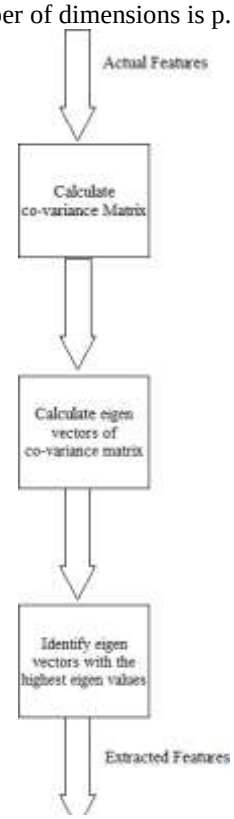

Fig 8: Principal Component Analysis

In this method new variables are created by linear combination of original data sets. Linear transformation function can be written as Z = XA,

Where X is the n x p matrix of n observation on p variables, Z is the n x p matrix of n values for each of p components; A is the p x p matrix of coefficients defining the line transformation.

Key steps for principal component analysis are,

- Start with the data with n observation on p variables.
- From a matrix of n × p with deviations from mean for each of the variables.
- Calculate the covariance matrix (p× p).
- Calculate the Eigen value and Eigen vector for the covariance matrix.
- Eigenvector with higher Eigen value is the principal component of data set.
- Choose the principal component and create a new feature vector.
- In the last step new data set will be derived.

B. Independent component analysis

A standout amongst the latest intense factual procedures for breaking down substantial data sets those are large in size and multivariate in nature is known as Independent component analysis (ICA). Linear transformation of original data is done for computational and theoretical straightforwardness. Some mainstream techniques those are applicable for linear transformation is known as principal component analysis, factor analysis, projection pursuit etc. Be that as it may, ICA [14] is unique in relation to different strategies, since in the portrayal it determines the components those are both factually autonomous and non-Gaussian in nature. Data those are autonomous in nature in a watched set of information blends are isolated by the ICA. Independent component analysis also used to take out some information and hidden patterns those are important from the multivariate data sets. A generative model has been introduced by independent component analysis as a database of items that is large in size and can generate detectable data values.

PCA is the older version and ICA is the new method for dimension reduction. In PCA premise vectors are spatially less restricted but in ICA premise vectors are spatially less restricted and statically autonomous. Premise vector are more affordable to figure in PCA not in ICA it is expensive in ICA. In PCA Premise vectors are orthogonal and positioned in arrange it is not done in ICA vectors are not in order. Projection error is minimized in PCA but in ICA statistical dependency of two premise vectors is minimized.

C. Linear discriminant analysis

Linear discriminant analysis (LDA) [11] is a method for dimension reduction. It is a simple classification method which is robust mathematically and sometimes produces some method whose accuracy is very good. LDA is dependent on the concept of searching for the variables those are linearly combined and separate two target classes. The main difference of LDA from PCA is that PCA does not consider the class information but LDA did and also LDA utilizes an inside diffuse framework or scatter matrix of all c classes $S_W = \sum_i^c = 1 \sum_j^{Ni} = 1 [(x_j - \mu_i) T (x_j - \mu_i)]$, and a scatter matrix: $S_B =$ Pc i=1 $[(\mu_i - \mu) T (\mu_i - \mu)]$, where N_i is the number of objects within class i, μ_i is the mean of class i (mean those are common), and the whole mean of all classes is μ. LDA preserves the discriminatory information as much as possible during the dimension reduction procedure. For the Eigen vector those are not zero then d = c − 1, which is related to the Eigen values (largest) of the matrix $S_w^{-1}S_B$, define the D X d feature extraction matrix M. This transformation maximizes the inside diffuse framework when within diffuse frame work is minimized and denotation of the determinant of a square matrix is done.

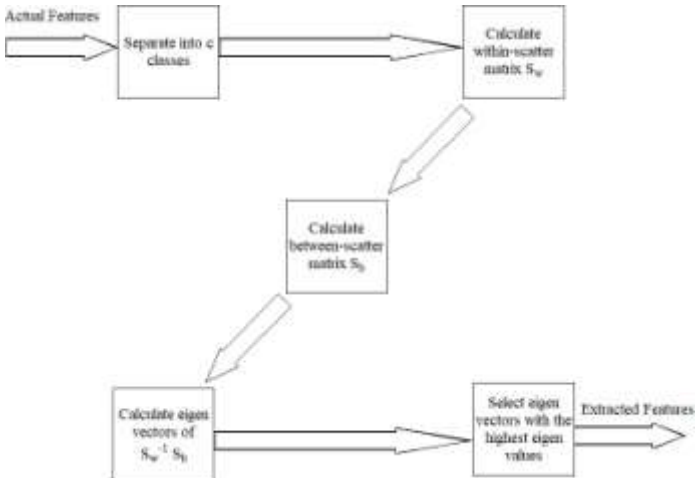

Fig 9: Linear Discriminant method

VI. CONCLUSION

As the use of different networking and smart devices are getting useful in our daily life, data gathered by those devices are also sensitive and requires much more storage space also with the provision that every data should be accessed at more or less same time from different locations. Big data concepts thus now days solves the problem but also each of the attributes of an entity is not needful for each application. For this purpose identification of correlated data is very much important and depending on that we need to select some features which are needful in our application to reduce time and space complexity. Not only from analyses point of view, the security in big data environment is a much concern for researcher.

REFERENCES

[1] Sabia and Sheetal Kalra, "Applications of big Data: Current Status and Future Scope", International Journal on Advanced Computer Theory and Engineering (IJACTE) 2319-2526, Volume -3, Issue -5, 2014.

[2] Muhammad Habib urRehman, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah and Samee U. Khan, "Big Data Reduction Methods: A Survey" Data Sci. Eng. (2016) 1:265–284 DOI 10.1007/s41019-016-0022-0.

[3] Hsieh C-J, (2013) "BIG &QUIC: sparse inverse covariance estimation for a million variables." Advances in neural information processing systems.

[4] Bi C, (2013) "Proper orthogonal decomposition based parallel compression for visualizing big data on the K computer". (2013)IEEE symposium on large-scale data analysis and visualization (LDAV).

[5] Bhagwat D, Eshghi K, MehraP (2007) "Content-based document routing and index partitioning for scalable similarity-based searches in a large corpus". In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining.

[6] Zhao D (2015) "COUPON: a cooperative framework for building sensing maps in mobile opportunistic networks". IEEE Trans Parallel DistribSyst 26(2):392–402.

[7] Dalessandro B (2013) "Bring the noise: embracing randomness is the key to scaling up machine learning algorithms". Big Data1(2):110–112.

[8] Ruhe A (1984) "Rational Krylov sequence methods for eigenvalue computation". Linear Algebra Appl 58:391–405.

[9] K.Sutha and Dr.J. JebamalarTamilselvi "A Review of Feature Selection Algorithms for Data Mining Techniques"International Journal on Computer Science and Engineering (IJCSE)ISSN : 0975-3397 Vol. 7 No.6 Jun 2015.

[10] S. Thakur , J. K. Sing, D. K. Basu, M. Nasipuri and M. Kundu, "Face Recognition using Principal Component Analysis and RBF Neural Networks", IJSSST, Vol. 10, No. 5 ISSN: 1473 - 804x online, 1473-8031.
[11] Dr. S.Vijayarani and S. Maria Sylviaa, "Comparative Analysis of Dimensionality Reduction Techniques" International Journal of Innovative Research in Computer and Communication EngineeringISSN(Online): 2320-9801 ISSN (Print): 2320-9798.
[12] Vipin Kumar and Sonajharia Minz, "Feature Selection: A literature Review" Smart Computing Review, vol. 4, no. 3, June 2014.
[13] I. T. Joliffe. "Principal Component Analysis". Springer-Verlag, New York, 1986.
[14] AapoHyvarinen, JuhaKarhunen, and ErkkiOja, A Wiley-Interscience Publication JOHN WILEY & SONS, INC.7 March 2001.
[15] Chattopadhyay A.K. at al. JISAS 2014 68(2) 39-54 Filzmoser et al. CSDA 2008 52 1694-1711.