INTERNATIONAL JOURNAL OF ADVANCED SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Efficient Dimensionality Reduction for Big Data Using Clustering Technique

Vallabh Dhoot[1], Shubham Gawande[2], Pooja Kanawade[3], Akanksha Lekhwani[4]
Students, Computer Science & Engineering, K.K.Wagh Institute of Engineering and Research, Nashik, India[1 2 3 4]

***Abstract*:-Clustering is unsupervised classification of patterns (observations, data items, or feature vectors) into teams (clusters). The drawback of clustering has been addressed in several contexts by researchers in several disciplines and so reflects its broad charm and quality in concert of the steps in exploratory data analysis. Clustering is useful in several exploratory pattern analysis, grouping, machine learning and making decisions as well as situations including data mining, document retrieval, image segmentation and pattern classification. We are living in a digital world. Every day, people generate massive amount of data and store it, for further analysis and management. The amount of knowledge in our world has been exploding. Big Data refers to extremely large datasets that may be analyzed computationally to reveal patterns, trends, and associations especially relating to human behavior and interactions. Due to the short growth of such information, solutions need to be studied so as to handle and extract price and information from these data sets. Therefore an analysis of the different classes of available clustering techniques with big datasets may provide significant and useful conclusions. The proposed system is to study and analyze some of the popular existing clustering techniques and impact of dimensionality reduction on Big Data.**

Keywords:- Data Mining, Clustering Techniques, Big Data, Dimensionality Reduction.

I INTRODUCTION

We are living in a digital world where we generate lots of data in gigabytes every single day. The global internet population grew up to 18.5% from 2013-2015. The data every day comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. The number is really huge with different applications coming each providing us one or another services the consumption of data has been tremendously increased. All of this has to be stored somewhere. For that purpose we have huge data warehouses, recently introduced cloud storages are been used to store this Big Data. Instagram, YouTube, Facebook, Twitter, Netflix are the big data generating services, having data been stored in different forms. This production of big data needs to be managed so that it can be used by various corporations which can be useful for people. But, it has many issues like storing this big data is a huge task to be performed as well as this size makes different operations like analytical operations, retrieval operations and various process operations on them very hectic, difficult and time consuming. Solution to these problems can be clustering of this big data in a compact format but that must be still an informative version of whole data. The aim of clustering of this huge amount of data is to provide a good quality of clusters. This can be a huge benefit for everyone from normal users to researches and corporate.

Big Data can be defined as datasets which are large or complex that the traditional data analyzing systems are inadequate. The factors on which Big Data can be categorized are: -

Volume:- It is an important characteristic to deal with when Big Data is concerned, as this requires substantial changes in architecture of storage systems as well as operations.

Velocity:- It leads to high demand for online processing of data, where processing speed is required to deal with data flows.

Variety:- It is third characteristics with different data types such as text, images, videos which are been produced by various sources like Smartphone's, laptops, sensors, etc.

The main aim of this paper is to provide readers with a proper analysis of the clustering algorithms on big data with the use of dimensionality reduction techniques. It provides experimental results from a variety of numerical datasets. Comparison of results between purely clustering algorithm outputs and our proposed system results is provided so that further researchers as well as students can refer to it while studying big data.

II LITERATURE SURVEY

The field of Data Mining has a different behavior towards Big Data. It can deal with data-sets having size gigabytes or even tera bytes. The main concern over here is that the algorithms which are used in data mining operations work on small data sets and do not give better results on large data sets. To work efficiently with large data sets, the algorithms must have high scalability. Clustering high dimensional data has always been a challenge for clustering techniques.

Clustering is unsupervised classification of patterns (observations, data items, or feature vectors) into teams (clusters). The drawbacks of clustering have been addressed in several contexts by researchers in several disciplines and so reflect its broad charm and quality in concert of the steps in exploratory data analysis. Clustering is useful in several exploratory pattern analysis, grouping, decision making and machine learning situations including data mining, document retrieval, image segmentation and pattern classification.

Adil Fahad, etal[1] performed a survey on clustering algorithms for Big Data. They have categorized 24 Clustering Algorithms as Partition-based, Hierarchical-based, Density-based, Grid-based and Model-based. Depending on the size of datasets, handling capacity of noisy data and types of datasets, Clusters are formed and the complexity of algorithms are

INTERNATIONAL JOURNAL OF ADVANCED SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

calculated. They concluded that no clustering algorithm performs well for all the evaluation criteria. All the clustering algorithm suffers from Stability problem. MacQueen [7] defined a technique for partitioning N-dimensional population into k-sets, which they named as K-means. They successfully concluded that k-means is computationally feasible and economical and has been a successful implementation for differentiating the data within a class.

S.Nazim presented a comparative review of dimensionality reduction techniques in regard with information visualization. The survey analyzed some DR methods supporting the concept of dimensionality reduction for getting the visualization of information with minimum loss of original information. As, we are dealing with Big Data. The issue of stability of clusters comes into picture. The theories [8] state that k-means does not break down even for arbitrarily large samples of data. The focus is on the behavior of stability of clusters formed by k-means algorithm.K-Means is closely related to principal component analysis [9]. The outcomes subject with regard to effectiveness of the solution obtained from k-means.

Unsupervised dimensionality reduction and unsupervised learning are associated closely [10]. The result provide new perception towards the observed quality of output obtained by PCA-based data reduction.

III PROPOSED SOLUTION

After a vast survey performed in literature [1] [2] and [3] we have studied the results and chosen two clustering techniques. But, still performing clustering on Big Data is an issue. To study and differentiate data is an issue as there are several dimensions and which dimensions are necessary to select creates problem. And with these all reasons we got motivated to study the clustering algorithms and dimensionality reduction process to achieve the following:

To propose a system which performs clustering on numerical data with the study and comparison of clustering algorithms performs dimensionality reduction on the data to reduce noise, dimensions for proper use of it.

Therefore, the proposed paper presents the study of clustering algorithm, their advantages, disadvantages and comparison with proper study and application of dimensionality reduction process its algorithm on big data. The experiment is performed on numerical data sets. Section II provides the review of clustering algorithm from the literature [1]. Section III has the dimensionality reduction method.

IV PROPOSED SYSTEM

As there are so many algorithms. This section shows the different advantages and comparison from literature [1] to be used in this system. Clustering means grouping of similar type of data. We are using Partition – Based algorithm in which data objects are divided into a number of partitions, where each partition represents a cluster and each object must belong to exactly one cluster. The partitioning algorithms are k-means, FCM, k-medoids, PAM, k-modes, CLARA and CLARANS.

K-Means

k-means classifies a given set of 'n' data objects in 'k' number of clusters. A centroid is defined for each cluster.

All the data objects are placed in a cluster having centroid nearest (or similar) to that data object. All data objects are bound to clusters based on the new centroids. In each iteration centroids change their location step by step. Centroids move in each iteration. This process is continued until none of the centroid moves. As a result, k clusters are formed having similar data objects.

Algorithm

Input- 'k' number of clusters, 'n' number of objects and their attributes (I have taken co-ordinates).

Output- 'k' clusters with similar data objects.

Steps-

1. Randomly select 'k' initial cluster centers.
2. Repeat
   i. Find the Euclidean distance of each of the centroids (the selected centers) with given data objects.
   ii. Update the centroid; i.e. associate the given data object to the cluster with minimum Euclidean distance.
3. GOTO step 2 if centroid changes, otherwise Stop.

FCM

Fuzzy C- means is a representative algorithm of fuzzy clustering which is based on K-means concepts to partition dataset into clusters. It is referred to soft clustering which means data objects can be in more than one cluster. These indicate the strength of the association between the data elements and a particular cluster. Fuzzy clustering is a process of assigning of membership levels and then using them to assign data to one or more clusters.

The algorithm attempts to partition a finite collection of 'n' elements X={X1, X2,… ,Xn} into a collection of 'C' fuzzy clusters with respect to some given criterion. The algorithm returns a list of 'C' cluster centers C={C1, C2, … , Cj} and a partition matrix W= wij ϵ [0,1], i=1,2, …, n and j=1,2,..,c.

Where each element wij tells the degree to which element Xi belongs to cluster Cj

Avg min $\sum_{i=1}\sum_{j=1} wij^m \| Xi - Cj\|^2$, $Wij=(1/\sum^c k=1(\|Xi-Cj\|/\|Xi-Ck\|)^{(2/m-1)})$

The fuzzy C-means algorithm minimizes intra- cluster variance as well. It iteratively searches the cluster centers and updates the memberships of objects. A FUZZY RULE states that the sum of the membership values of a data object to all clusters must be 1. The higher the membership value, the more likely a data object will belong to that cluster.

$J=\sum^n(i=1)\sum^c(k=1)\mu^m ik|pi-vk|^2$, Where J=objective function,

INTERNATIONAL JOURNAL OF ADVANCED SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

n=number of objects, c=number of defined clusters,

µik=likelihood values by assigning the object i to the cluster k,

m=fuzziness factor (a value<1),

|pi-vk|=Euclidean distance between i^{th} object pi and the k^{th} cluster center.
The centroid of the k^{th} cluster:-

$vk=((\sum^{n}(i=1)\mu^{m}(ik)\ pi)/(\sum^{n}(i=1)\mu^{m}(ik)))$

Algorithm

Input- 'c' number of clusters, the fuzzy parameter 'm', initialize the fuzzy partition matrix, and *stop=false*.

Output- A list of c cluster centers and a partition matrix are produced.

Steps-

1. Do:
2. Calculate the cluster centroids and the objective value
3. Compute the membership values stored in the matrix.
4. If the value of *J* between consecutive iterations is less than the stopping condition, then *stop=true*.
5. While(!stop)

TABLE I COMPARISON OF FCM AND K-MEANS

Parameter	FCM	K-Means
Type of Clustering	Soft Clustering	Hard Clustering
Type of Dataset	Numerical	Numerical
Size of Dataset	Large	Large
Input Parameter	3	3
Complexity	O(n)	O(nkd)

The main problem for using these algorithms arises when they are applied on Big Data. Many un-necessary attributes are even considered while the formation of clusters which are irrelevant to the application. So to minimize such irrelevance of clustering algorithms, the dataset can be reduced such that the reduced dataset is as useful and informative as the original one. Moreover when the reduced dataset is passed as input to the clustering algorithms, clusters will be more correctly classified. Dimensionality reduction techniques are one of the solutions of reducing the datasets.

V DIMENSIONALITY REDUCTION

Analyzing Big Data has been a great difficulty with n-number of dimensions (attributes) it adds much more difficulty to study the data. Dimensionality Reduction is the process of reducing random variables which are not that necessary to differentiate data. The whole process is divided into two parts:

Feature Selection
Also known as Variable Selection, Attribute Selection and Variable Subset Selection. It is a process of selecting important features (variables, attributes) for building models. Using feature selection technique we can remove the redundant or irrelevant features from the data without much loss of information.

Feature Extraction
The transformation from high- dimensional data to reduced feature data is called as Feature Extraction. The extracted feature are expected to be relevant information from the given input dataset. There are various algorithms for feature selection and extraction. The main algorithm which will be used to reduce the data will be:

Principal Component Analysis (PCA)
PCA is a statistical method which uses orthogonal transformation (it is linear transformation which doesn't change even after performing rotation ad reflection operation upon the data) to convert set of observation of possibly correlated attributes into a set of values of unrelated data variables. It identifies patterns and finds patterns to reduce the dimensions of the data with minimal loss of information. The attributes are been converted to Principal attributes which are important and necessary for defining the data.
Algorithm
Input: Data
Output: Data with only Principal attributes
Steps:

1. Perform the orthogonal transformation
2. Select the Eigen vectors (Those attributes which are not affected by the above operation) and Eigen values
3. Sort the Eigen vectors and Eigen values according to decreasing order.
4. Select some subset of Eigen vectors as per their values as Principal Attributes of the data.

Complexity: O (p^2n+p^3), where p is the features and n is data points.
The program flow for this project is as follows:
First the preprocessing first the preprocessing is done where three techniques are used to filter out the relevant data from the dataset.
Zero reduction: the attribute containing number of zeros more than a predefined threshold is eliminated.

Mean and variance: In this technique, the variance is calculated and the attribute having variance less than the predefined threshold is

INTERNATIONAL JOURNAL OF ADVANCED SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

eliminated.

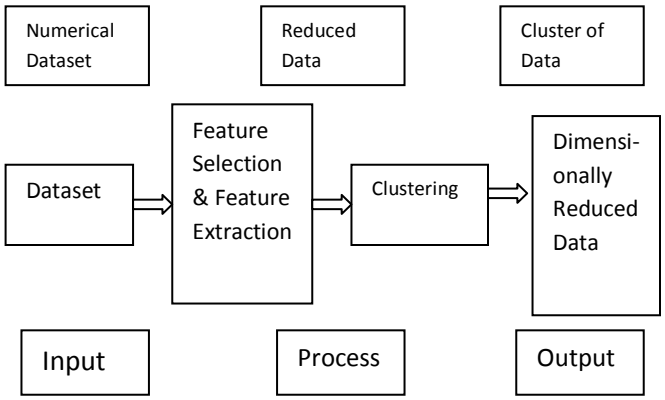


Figure 1 Flow Diagram

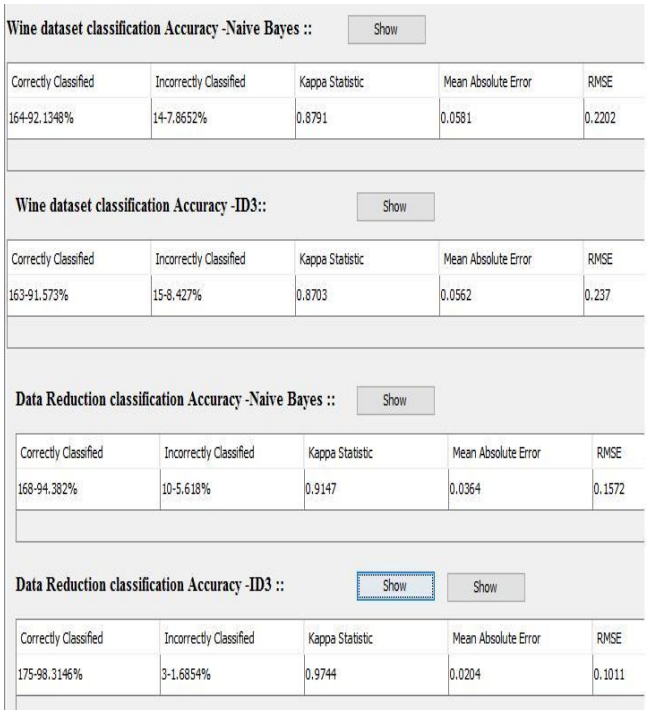
Pearson correlation: In this, similarity between two attributes is calculated using the Pearson correlation formula, and if the similarity is above a predefined threshold the second attribute is eliminated.

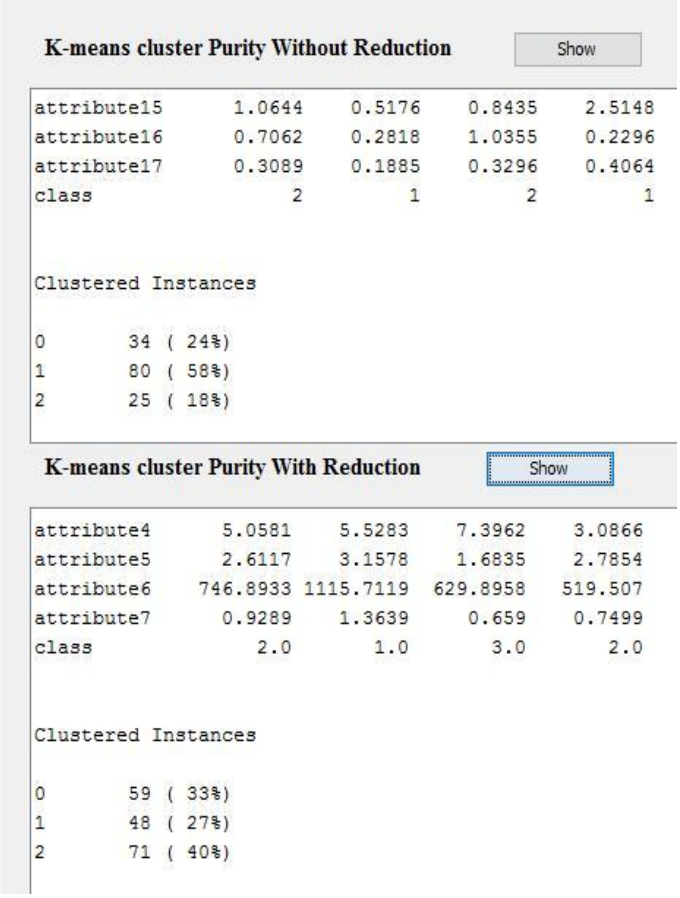
Next, the PCA technique of dimensionality reduction is carried out on the filtered dataset. After this, clustering techniques (K-means and FCM) are applied to the reduced dataset.

VI EXPERIMENTAL RESULTS

This project aims to study the impact of dimensionality reduction on big data. The results show the difference between the performance of pure clustering results and clustering of dimensionally reduced data. In addition to that, the results also show that the data which has been reduced is as much as useful as the original dataset. It can be said that despite of small or reduced size, the new dataset which is reduced using PCA and other preprocessing techniques is as important and useful as the large dataset. To validate these results and the code, different testing strategies can be used. To compare the results, an analysis software is used which gives the results with the help of two classification algorithms: Naïve Bayes and ID3 algorithms. These two algorithms measure the performances of the datasets before and after reduction i.e. analyzing whether after reduction the dataset is informative or not. The results show that after reduction data set can be more correctly classified by clustering algorithms. In addition to this there is reduced error rate and applying dimensionality reduction before clustering improvises the purity of clusters.

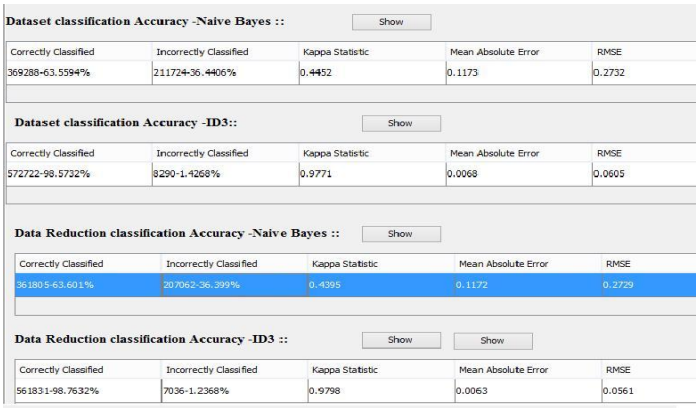
The results show that after reduction of dataset the percentage to which the dataset can be correctly clustered and classified is increased whereas the error rate is reduced and with the use of dimensionality reduction techniques in clustering of big data more pure clusters are formed. A standard wine dataset is used which is used as training set for this project. Below are results obtained





INTERNATIONAL JOURNAL OF ADVANCED SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Similar results are obtained for another stock market dataset.



Dataset classification Accuracy -Naive Bayes :: Show

Correctly Classified	Incorrectly Classified	Kappa Statistic	Mean Absolute Error	RMSE
369288-63.5594%	211724-36.4406%	0.4452	0.1173	0.2732

Dataset classification Accuracy -ID3:: Show

Correctly Classified	Incorrectly Classified	Kappa Statistic	Mean Absolute Error	RMSE
572722-98.5732%	8290-1.4268%	0.9771	0.0068	0.0605

Data Reduction classification Accuracy -Naive Bayes :: Show

Correctly Classified	Incorrectly Classified	Kappa Statistic	Mean Absolute Error	RMSE
361805-63.601%	207062-36.399%	0.4395	0.1172	0.2729

Data Reduction classification Accuracy -ID3 :: Show Show

Correctly Classified	Incorrectly Classified	Kappa Statistic	Mean Absolute Error	RMSE
561831-98.7632%	7036-1.2368%	0.9798	0.0063	0.0561

K-means cluster Purity Without Reduction Show

```
attribute52          0.0268      0.0252      0.0288      0.0252
attribute53          0.0238       0.019      0.0259      0.0245
attribute54          0.0151      0.0151      0.0167      0.0127
class                     2           1           2           2


Clustered Instances

0       145249  ( 25%)
1       260730  ( 45%)
2       175033  ( 30%)
```

K-means cluster Purity With Reduction Show

```
attribute51          0.0005      0.0003           0       0.001
attribute52          0.0264      0.0228           0      0.0453
attribute53          0.0233      0.0233           0      0.0364
attribute54          0.0148      0.0087           0      0.0303
class                     2           2           2           2


Clustered Instances

0       248511  ( 44%)
1       114005  ( 20%)
2       206351  ( 36%)
```

VII CONCLUSION

This paper provides a comprehensive study of the partition - based clustering algorithms proposed in the literature [1]. The process been performed on numerical dataset in order to reduce the dimensionality of Big Data. In order to future directions for performing the process with different datasets and other algorithms and to guide the process of dimensionality reduction process for big data, we proposed a framework to perform the operations on a number of clustering algorithms. The framework is developed from a theoretical study that would automatically perform the operations which can be helpful for researchers, students and corporate. Therefore, even other than these algorithms other techniques can be added to the framework according to the input and properties. In general from the literature and the operations performed we can conclude that:

- Big Data is not easy to handle. It is difficult to handle the noise and dimensions.
- Extracting important features is tricky as per the user's need. Features can vary according to the need. So, dimensions must be reduced with proper study.
- Numerical dataset gets clustered and reduced as per the required way.

VIII FUTURE WORK

As for future work, we can try for the following questions:

- How to perform these techniques for character datasets?
- Can we make a framework for this system?
- Are the "principal" attributes helpful for differentiating a data from another?
- Applying parallelism and analyzing the computational performance of the system

REFERENCES

[1] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis ", *on Emerging Topics on Computing*, IEEE, 11 June 2014.

[2] A. A. Abbasi and M. Younis, ``A survey on clustering algorithms for wireless sensor networks," *Comput. Commun.*, vol. 30, nos. 14_15, pp. 2826_2841, Oct. 2007.

[3] C. C. Aggarwal and C. Zhai, ``A survey of text clustering algorithms," in *Mining Text Data*. New York, NY, USA: Springer-Verlag, 2012, pp. 77_128.

[4] R. Xu and D. Wunsch, ``Survey of clustering algorithms," *IEEE Trans.Neural Netw.*, vol. 16, no. 3, pp. 645_678, May 2005.

[5] S. Guha, R. Rastogi, and K. Shim, ``Cure: An ef_cient clustering algorithm for large databases," in *Proc. ACMSIGMOD Rec.*, Jun. 1998, vol. 27, no. 2, pp. 73_84.

[6] Improving Fusion of Dimensionality Reduction Methods for Nearest Neighbor Classification Deegalla, S.; Bostrom, H. Machine Learning and Applications, 2009. ICMLA '09. International Conference on Year: 2009

[7] J. MacQueen, ''Some methods for classification and analysis of multivariate observations,'' in Proc. 5th Berkeley Symp . Math. Statist. Probab., Berkeley, CA, USA, 1967, pp. 281–297

[8] O Shamir,"Model Selection and Stability in k-means Clustering"

[9] C Ding,"Principal Component Analysis and Effective K-means Clustering"

[10] Chris Ding and Xiaofeng He, "K-Means Clustering via Principal Component Analysis", In proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004