High Dimensional Data Clustering Using Partition Based Reduction Algorithm

D, Ashok Kumar[1], P.Velayudham[2]
Department of Computer Science
1. Supervisor Manonmaniam Sundaranar University, Tirunelveli - 627 012, Tamil Nadu, India
2. Research Scholar Manonmaniam Sundaranar University, Tirunelveli - 627 012, Tamil Nadu, India

Abstract - Clustering is an unsupervised classification of patterns into small clusters. The deviations of clustering techniques are quality of cluster and data analysis techniques. Clustering is useful in several exploratory pattern analysis, grouping, machine learning and decision making as well as situations including data mining, document retrieval, image segmentation and pattern classification. Every day, people generate massive amount of data and store it into the database, for further analysis and management, it increases the data volume as terabyte or petabyte for the nature of big data. 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. The short growth of information, solutions need to be studied to extract the information from big datasets. In this study the PCAFC-Means clustering algorithm is proposed. The proposed PCAFC-Means algorithm with its fuzzy computing techniques provides the solution for transformation problems. The proposed PCAFC-Means algorithm is tested with UCI HealthNews, Diabetes Datasets. The experimental results evaluate the performance of existing PCAK-Means algorithm and proposed PCAFC-Means algorithm. The result shows that the proposed PCAFC-Means algorithm outperformed the existing PCAK-Means algorithm.

Index- Optimal Partition Clustering, High Dimensional Data, and Dimensionality Reduction.

I. INTRODUCTION

In this digital world, people generate massive amount of data every day. The global internet population grew up for last decades and comes from everywhere. The sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPRS signals. The application and services provided to the users are the consumption of data that has been extremely increased as a big data, that is stored into data warehouse, and cloud storage. YouTube, Facebook, Vine, Twitter, Netflix are the big data generating services to generate the different forms of big data. This production of big data needs to be managed by various corporations which is useful for people. Big data process has many issues like storage of data, size of big data makes difficult and time consuming for different operations like analytical operations, retrieval operations. The solutions for big data problems, to clustering the big data into a compact format. The proposed system is to study and analysis the popular existing supervised and unsupervised reduction techniques and used to get useful knowledge from big data. The impact of big data is the curse of dimensionality. In this study, the Principal Component Analysis with Fuzzy C- Means(PCA-FC-Means)clustering algorithm is proposed. The proposed PCA-FC-Means algorithm is tested with Healthcare (HealthNews, CancerRNA sequence, Diabetes) datasets.

Big Data[1] can be defined as datasets which are large or complex that the traditional data analyzing systems are inadequate to classified. The factors on which big data can be categorized are V3(Volume, Velocity, Veracity)

• 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 smartphones, laptops, sensors, etc.

NATURES OF BIG DATA

- The analysis of big data can be approached in several ways, but the underlying problem remains a statistical problem.
- Learning methods for big-data are analytical methods that generalize and adapt the classical statistical methods to new large data sets.
- The underlying problems, in many respects, remain the same, although with different emphasis.
- Big-data analysis requires the ability to assess the effectiveness of the model with a non-classical inferential approach that takes advantage of the large amount of data available.

II. LITERATURE REVIEW

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 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[2] has always been a challenge for clustering techniques to produce a high quality of clusters. 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 is calculated. They concluded that no clustering algorithm performs well for all the evaluation criteria. All algorithms suffer from Stability problem.

Fan J, Han F, Liu H [4] 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.

Many author presented a comparative review[5][6][7][8] of dimensionality reduction techniques in regard with information visualization. The survey analyzed some dimension reduction 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 Hsieh C-J et al [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 provides new perception towards the observed quality of output obtained by PCA-based on K-means concepts to partition dataset into clusters.

This paper is organized as follows, Section III describe the Existing PCA with K-Means algorithm. Section IV describes the Proposed PCAFC-Means Algorithm. Section V describes the Experimental Results. Section VI gives the Conclusion.

III. Existing Principal Component Analysis with K-Means.

Clustering big data is a challenging issue because of its huge volume of data. The reduction falls with the

unexpected loss of information. The information loss occurs due to learning algorithms unknown to identify which attributes have been converted to Principal attributes. PCA is the best and simple non-linear dimensionality reduction technique. Since the data volume increases the PCA performance goes down. The partition algorithms help to improve the strength PCA.

The K-Means is one of the most important partition based clustering algorithm defining the data 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. Since the Reduction performs little more efficient only for numerical data sets. Principal Component Analysis(PCA)[9] is a statistical method which uses orthogonal transformation (it is linear transformation which does not change even after performing rotation and reflection operations. Principal Component Analysis(PCA) [9] is a statistical method which uses orthogonal transformation (it is linear transformation which does not change even after performing rotation and 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 have been converted in to Principal attributes which are important and necessary for defining the data.

The implementation process of Principal Component Analysis (PCA) algorithm is represented in Algorithm.1

Algorithm.1 - Principal Component Analysis

Input : $D = d_1,d_2,d_3,...,d_i,...,d_n$ // Set of data points(n x d Matrix).
$d_i = x_1,x_2,x_3,...,x_i,...,x_m$ // Set of attributes of one data point.
k // Number of Principal Attributes.
Output : Reduced Dimensional Matrix R^p

1. Input dataset $D = d_1,d_2,d_3,...,d_i,...,d_n$ where D is a n × d Matrix. $d_i = x_1,x_2,x_3,...,x_i,...,x_m$ and the new dimensionality d where n×d matrix which is ($d \leq\leq D$)
2. Compute the mean $m = 1/n\sum^{n}_{i=1} x_i$
3. Subtract mean value m from each row x_i then go to step 6.
4. Compute the Eigen vectors and Eigen values.
5. Sort the Eigen vectors and Eigen values according to decreasing order.
6. Select some subset of Eigen vectors (R^p) based on principal attributes, where p is the dimension of Eigen vector ($p \leq\leq d$).Complexity :$O(p^{2n}+p^3)$, where p is the features and n is number of data points.

The Principal Component Analysis is an important nonlinear dimensionality reduction algorithm. The implementation process of PCA is compute the Eigen vector and Eigen values then select the subset (R^p) vectors from Eigen vectors (R^d). The subset
vector is selected based on principal attributes.

The implementation process of partition K-Means clustering algorithm is represented in Algorithm 2.

Algorithm 2. Partition K-Means Clustering

Input : $d_i = (x_1,x_2,x_3,...,x_i,...,x_n)$ // where d is n×d Matrix. k // Number of centroids .
Output : set of k clusters.

1. In the given data set d, if the data points contain the both positive and negative attribute values then go to step 2, otherwise go to step 4.
2. Find the minimum attribute value in the given data set d.
3. For each data point attribute, subtract with the minimum attribute value.
4. For each data point calculate the distance from origin. Divide the distance from sum of square error, and take the average distance.

5. Sort the distances obtained in step 4.
6. Compare the average with the sorted data points
7. If average distance ≤ distance then assign the average distance as a initial centroid. Compute the distance between each data point di($1 \le i \le n$) for all the initial centroids cj($1 \le j \le k$).
8. Else
9. For each cluster j($1 \le j \le k$), recalculate the centroids. Compute the distance from each centroid of the present nearest cluster. Divide the distance from sum of square error and take the average distance.
10. End for. Repeat until the convergence criterion is met.

The implementation process of K-Means algorithm described into different steps, first to select the initial centroids then calculate the distance. Next to divide the distance from sum of square error and take the average. The average distance compare with data points distance. The average distance is less than the data point distance then the average distance is an initial centroid, otherwise recalculate the centroids. This process is continued until the convergence criterion is met.

IV. Proposed Principal Component Analysis with Fuzzy C-Means.

The proposed Partition Based Principal Component Analysis with Fuzzy C-Means (PCAFC-Means) algorithm used for the orthogonal transformation. It converts the set of observation is possibly correlated attributes into a set of values in uncorrelated data variables. The FC-Means identifies the patterns and reduce the dimensions of the data with minimal loss of information. The proposed algorithm is used to identify which attributes has been converted into principal attributes, and which data objects are divided into a number of partitions. The partition represents a cluster and each object must belong to exactly one cluster. The existing reduction algorithms struggling to prove its efficiencies while reduce the dimensions of big data. The FC-Means is a representative algorithm of fuzzy clustering which is based on K-means with square error concepts to partition dataset into clusters. It improve the partition of data objects in more than one clusters based on fuzzy computing techniques and also produce good clustering accuracy.

The implementation process of partition Fuzzy C-Means clustering algorithm is represented in Algorithm 3.

Algorithm 3. Partition Fuzzy C-Means Clustering

Input : $d_i = (x_1,x_2,x_3,...,x_i,...,x_n)$ // where d is n×d Matrix. k // Number of centroids .
Output : set of k clusters.

1.Partition matrix W= w_{ij} . [0,1], i=1,2, …, n and j=1,2,..,c. Where each element w_{ij} tells the degree to which element X_i belongs to cluster C_j .
2. Avg min $S(i=1)S(j=1)\ w_{ij}m \parallel X_i - C_j\parallel 2$,
$W_{ij}=(1/S_c(k=1)(\|X_i-C_j\|/\|X_i-C_k\|)(2/m-1))$
3. $J=S_n(i=1)S_c(k=1)\mu_{mik}|p_i-v_k|2$ where J=objective function , n=number of objects, c=number of defined clusters, μ_{ik}=likelihood values by assigning the object
i to the cluster k, m=fuzziness factor (a value<1),
4. $|p_i-v_k|$=Euclidean distance between i^{th} object p_i and the k^{th} cluster center. The centroid of the k^{th} cluster $v_k=((S_n(i=1)\mu_m(ik)\ p_i)/(S_n(i=1)\mu_m(ik)))$.
5. Do calculate the cluster centroids and the objective value J.
6. Compute the membership values stored in the matrix.
7. If the value of J between consecutive iterations is less than the stopping condition, then stop=true.
8. While(!stop)

The implementation process of FC-Means algorithm described into different steps, first to partition the

reduced dataset in to matrix then calculate the average minimum. Next set the objective function using fuzziness factor then compute the membership value. This process is continued up to the value of objective function is less than the stopping condition.

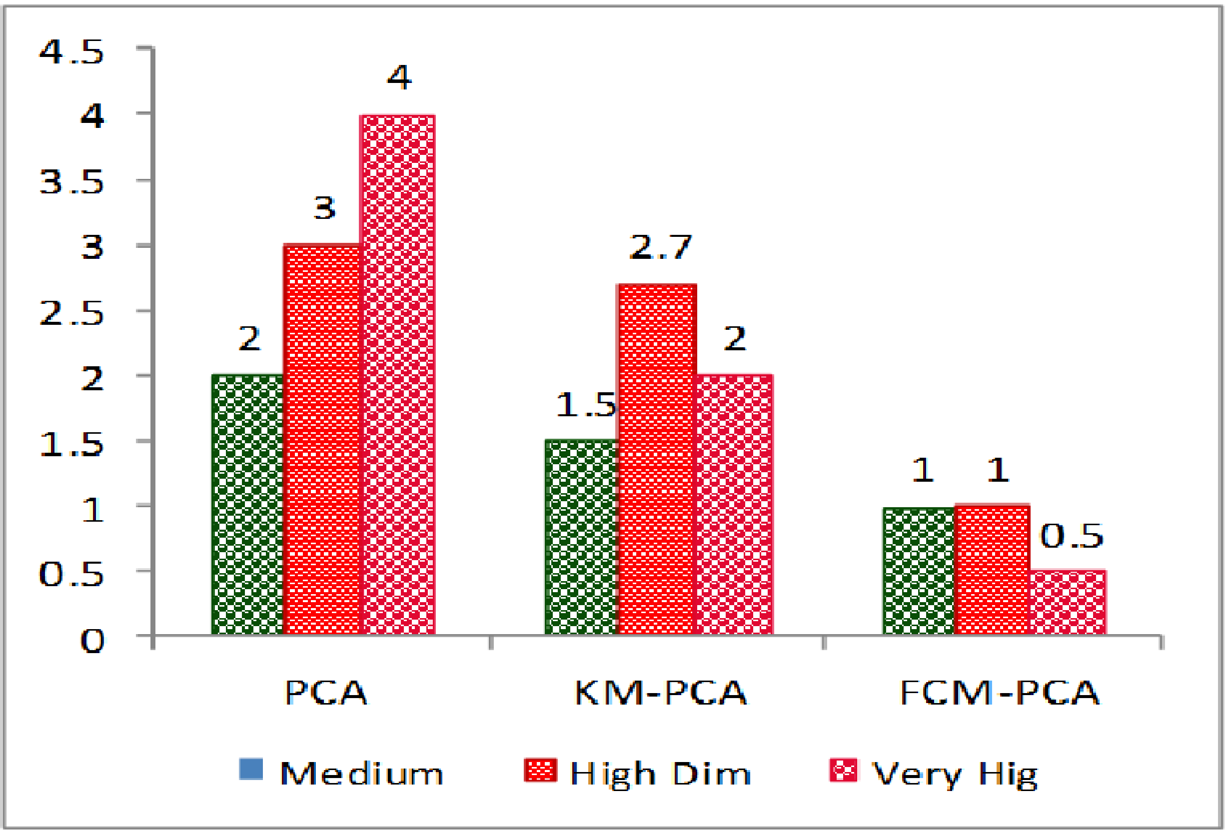
V. Experimental Results

In this section, the Principal Component Analysis with Fuzzy C-Means (PCAFC-Means) algorithm is tested with CancerRNA, HealthNews, and Diabetes UCI repository datasets. The CancerRNA [19, 20] dataset contains 805 instances and 20531 attributes. The HealthNews [17,18] is a twitter dataset. It contains the 58000 instance and 25000 attributes. The Diabetes [16, 21] data set contains 100000 instances and 55 attributes. This data has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes The result of PCAFC-Means algorithm is compared with PCA and PCAK-means. The performance of PCAFC-Means algorithm and PCAK-Means algorithm tested with CancerRNA, HealthNews, and Diabetes datasets. The result of reduction status and clustering efficiency of PCAFC-Means algorithm compared with PCAK-Means algorithm. The results shows that PCAFC-Means algorithm has outperformed. It is represented in Table 1.

Table.1: Clustering Status for PCA, PCAK-Means and PCAFC-Means Algorithm

Name of Data Sets	**Number of Attributes and Instances**	***Reduced with Clustering Status***		
		PCA %	*Kmean-PCA %*	*FC-mean PCA%*
Cancer RNA Sequence	20531/805	98	98.5	99
Health News	25000/58000	97	97.3	99
Diabetes	55/100000	96	98	99.5

From Table 1 it is observed that the clustering status of algorithms for three big datasets. The clustering status percentage with CancerRNA dataset of PCAK-Means is 98.5 and PCAFC-Means is 99. The clustering status percentage with HealthNews dataset of PCAK-Means is 97.3 and PCAFC-Means is 99. The clustering status percentage with diabetes dataset of PCAK-Means is 98 and PCAFC-Means is 99.5. The results indicates that the minimum loss of data is PCAFC-Means algorithm. The PCAFC-Means algorithm has outperformed compared with existing PCAK-Means algorithm. The clustering status of PCAK-Means and PCAFC-Means algorithm is represented in Figure 1



From the Figure 4.1, it is observed that the PCAFC-Means clustering algorithm is outperformed.
The processing time, error rate and accuracy of PCAFC-Means algorithm is compared with PCAK-means.

The performance accuracy of PCAFC-Means algorithm and PCAK-Means algorithm is tested with various parameters.

The proposed model is validated using four parameters namely the Accuracy of the clustering, through Area under ROC Curve, Sensitivity and Specificity.

The accuracy of algorithm is measured by using the Equation (1).

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN} * 100 \quad \ldots\ldots(1)$$

The Sensitivity of algorithm is measured by using the Equation (2).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad \text{------}(2)$$

The Sensitivity of algorithm is measured by using the Equation (3).

$$\text{Specificity} = \frac{TN}{TN + FP} \quad \ldots\ldots(3)$$

Where TP(True Positive): The number of examples correctly classified to that class.

TN(True Negative): The number of examples correctly rejected from that class.

FP(False Positive): The number of examples incorrectly rejected from that class.

FN(False Negative): The number of examples incorrectly classified to that class.

Table 4.2: Processing Time and Clustering Accuracy for PCAK-Means and PCAFC-Means

Classification Performance	***Reduction and Clustering Algorithm status***		
	PCA	*K-PCA*	*FC-PCA*
Classification Accuracy(%)	98	98.35	99.14
Processing Time(msec)	0.51	0.50	0.49
Error Rate(%)	$\pm$ 1.50	$\pm$ 0.7	$\pm$ 0.5

From Table 2 it is observed that the clustering accuracy, processing time and error rate for PCAFC-Means algorithm outperformed with existing PCAK-Means Algorithm. The performance of algorithm is measured through ROC curve. Receiver Operating Characteristic(ROC) describes the tradeoff between Sensitivity and Specificity, as well as the performance of the clustering can be visualized and studied to the proposed PCAFC-Means Algorithm has outperformed. The ROC curve is represented in Figure 2.

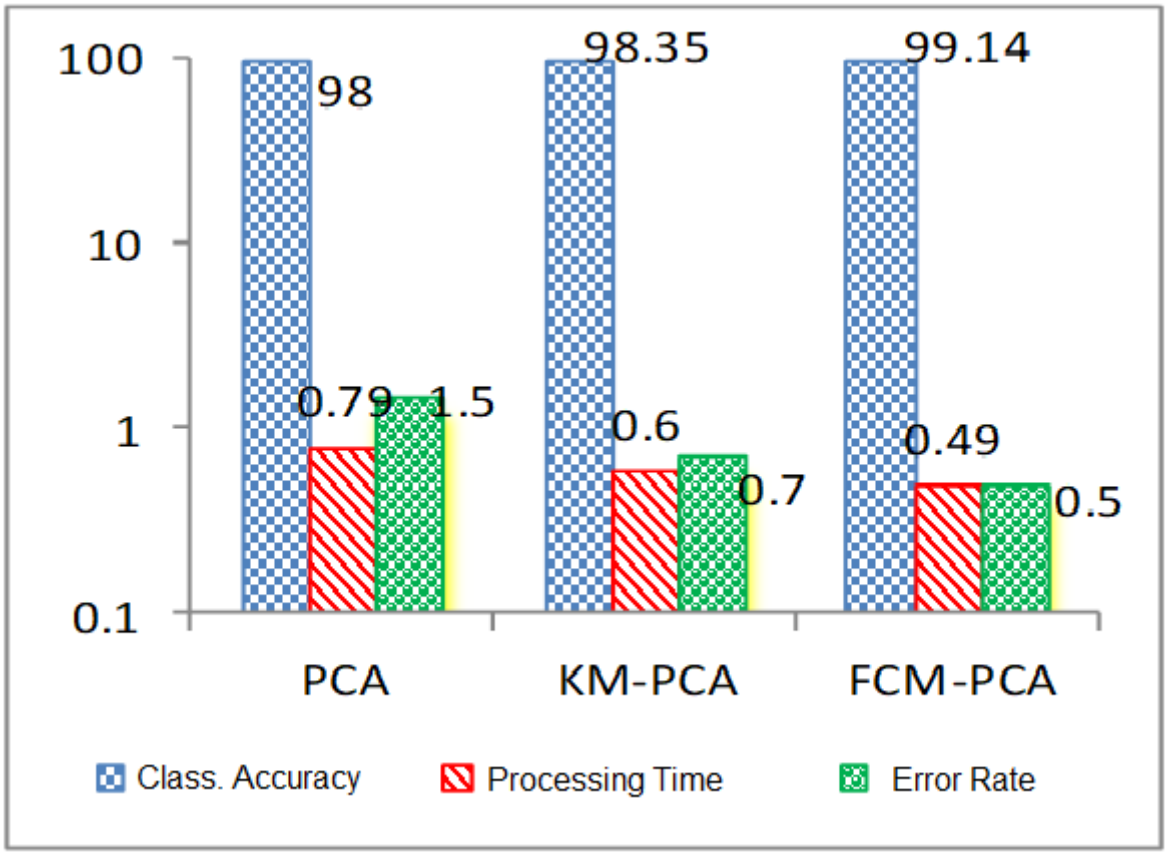


Fig. 2 ROC Curve for Dimensionality Reduction Techniques Vs Classification Accuracy

From the Figure 4.2 it is observed that the ROC indicates the high accuracy and low error rate with minimum processing time for PCAFC-Means algorithm compared with PCAK-Means algorithm. It indicates that the PCAFC-Means algorithm has outperformed.

IV. CONCLUSION

The basic concepts and related literature reviews of non linear dimensionality reduction PCAK-Means based clustering techniques are presented in this chapter. The PCAFC-Means clustering algorithm is proposed. The proposed PCAFC-Means algorithm with its soft computing techniques provide the solution for transformation problems. The proposed PCA-FC-Means algorithm test with HealthNews, Diabetes Datasets. The experimental results evaluate the performance of existing PCAK-Means algorithm and proposed PCAFC-Means algorithm. The result was observed that the proposed PCAFC-Means algorithm outperformed the existing PCAK-Means algorithm.

VI. 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, 2014.

[2] Che.D, Safran.M and Peng.Z, From big data to big data mining: challenges, issues, and opportunities. In: Database systems for advanced applications, 2013.

[3]. Zhai.Y, Ong.Y.S and Tsang.I.W (2014) The emerging ‘‘big dimensionality’’. Comput Intell Mag IEEE .:PP..14-26, 2014.

[4]. Fan. J, Han. F and Liu. H (2014) Challenges of big data analysis. Nat Sci Rev 1(2): PP.293-314, 2014.

[5]. Chandramouli.B, Goldstein.J and Duan.S, Temporal analytics on big data for web advertising. In: 2012 IEEE 28th international conference on data engineering (ICDE), 2012.

[6]. Ward R.M et al. Big data challenges and opportunities in high-throughput sequencing. SYST-Biomed Vol.1, PP.29-34, 2013.

[7]. Weinstein.M et al, Analyzing big data with dynamic quantum clustering. arXiv preprint arXiv:PP.1310.2700, 2013.

[8]. Hsieh.C.J et al, BIG & QUIC: sparse inverse covariance estimation for a million variables. In: Advances in neural information processing systems, 2013.

[9]. Vervliet.N et al, Breaking the curse of dimensionality using decompositions of incomplete tensors: tensor-based scientific computing in big data analysis. IEEE Signal Process Mag, Vol. 5, PP.71–79, 2014.

[10] Abbasi.A.A and Younis.M, ``A survey on clustering algorithms for wireless sensor networks", Vol. 30, PP. 2826-2841, Oct. 2007.

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

[12] Ranjan Maitra, Anna.D. Peterson, Arka.P and Ghosh.A systematic evaluation of different methods for initializing the k-means clustering algorithm IEEE Transactions on knowledge and data engineering, Vol.1, No.2, Jan.2011.

[13] Madhu Yedla. et al. “Enhancing k-means clustering algorithm with improved initial centers,” International Journal of Computer Science and information Technologies. Vol.1(2), 2010.

[14]. Weiling Cai, Songcan Chen and Daoqiang Zhang- Fast and Robust Fuzzy C-Means Clustering Algorithms Incorporating Local Information for Image Segmentation. 2010.

[15]. Srinivasalu Asadi, Dr.Ch.D. Subba Rao.V and Saikrishna.V, “A Comparative Study of Face recognition with Principal Component Analysis and Cross-Correlation Technique”, International Journal of Computer Applications (0975 – 8887), Vol. 10, No.8, PP. 17 – 25. Nov. 2010.

[16]. Ann Kellett.J.D and R.N, International Diabetes Center Diagnosis and classification of diabetes mellitus, American Diabetes Association, Diabetes Care, Vol. 36, pp. S67-S74, 2013.

[17] Ismkhan.H, Pattern Recognition, An iterative clustering algorithm based on an enhanced version of the k-means, Elsevier, 2018 .

[18] Karami.A, Gangopadhyay.A, Zhou.B and Kharrazi.H, Fuzzy approach topic discovery in health and medical corpora, International Journal of Fuzzy Systems, pp.1-12, 2017.

[19] Jian-Rong Li, Chuan-Hu Sun, Wenyuan Li, Rou-Fang Chao, Chieh-Chen, Huang Xianghong, Jasmine Zhou and Chun-Chi Liu, Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells, Nucleic Acids Research, Vol. 44, Iss.D1, pp. D944-D951, 2016.

[20] Shah.S.A.A, Nadeem.U and Bennamoun.M, Efficient image set clustering using linear regression based image reconstruction, openaccess.thecvf.com, 2017.

[21] Takayasu Moriyaa, Holger.R. Rotha, Shota Nakamurab, Hirohisa Odac, Kai Nagarac, Masahiro Odaa and Kensaku Moria, Unsupervised pathology image segmentation using representation learning with spherical k-means, Medical Imaging, spiedigitallibrary.org, 2018.