ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 2, Issue 11, DOI: 10.29027/IJIRASE.v2.i11.2019, 362-367 , May 2019

A Research on Big Data Clustering with Improvisation in K-Means Clustering using Semi Supervised Clustering

[1]Bindu Rani,*bindu.var17@gmail.com,Sharda University,Greater Noida,India*
[2]Shri Kant, *shri.kant@sharda.ac.in.,Sharda University, Greater Noida,India*

***Abstract*— The rapid revolution and adoption of big data by organizations has changed the approaches for using sophisticated information technologies as well as to gain insight knowledge for proactive decisions making. This data-oriented concept is remarkable as data is generated and available easily via various living (normal users) as well as non living media (sensors, web media etc) also and is increasing exponentially at rapid pace. Due to advancement in technologies, data storage is not trouble but how to analyze data is a major issue. Taking into account analysis of data, considerable data mining techniques are association, classification, clustering and regression analysis. These techniques have position in the design phase of Decision making process. Clustering have the property to acquire knowledge from data and can be considered the best technique to improve decision making process. Existing clustering algorithms are appropriate for small data sets but for big data or real life data it is challenging task, no unique algorithm for clustering can be applied directly. Scaling, correct parameterization, parallelization, cluster validity are some problems in using clustering techniques. In consideration of all aforementioned problems, continuous efforts are being made by data mining researchers. Big data Clustering techniques are discussed in this paper with main focus on unsupervised K-means clustering algorithms and their limitations. In addition with unsupervised clustering, semi-supervised clustering methods are also reviewed and**

I. INTRODUCTION

In the view of enormous data available, big data analytics is becoming need for every organization. Most famous social media networks as Twiiter, Google, Facebook etc, health organizations, media & entertainment, Government organizations are most common organizations that are in the field of exploiting and analyzing big data generated by various sources and of different variety to find valuable knowledge . Hence there is need to extract as much as value from this big data in making informed decisions. Every second data is produced by various sources and data mining tools and technologies can be used to process and analyze this data. Consistent patterns and relationship between variables are the key characteristics for data mining process. Then these hidden patterns are validated on new subset of data and converted into structured knowledge. Structured knowledge and human knowledge are combined to make intelligent decisions[1].

In consequences of different V's(Volume, Variety, Velocity)[2] of big data, the major issues are to load, store and perform analytical queries on big data. Instead of having structured and unstructured data both, now a day's storing of huge data is not a big problem. One approach, Data lake [3] store all the data in single storage system either structured or unstructured. Some popular techniques to store big data are Hadoop framework, NoSQL databases, Cassandra. Since unstructured data do not have predefined structure as structured data and not easy to analyze, unstructured data is first converted into structured form. Even structured data can consist of high dimensional data items as suppose structured data X consists of n number of data items X1,X2, . . . , Xn. Each data item is defined by the m number of features x1, x2, . . . , xm, then xij, i = 1, . . . , n, j = 1, . . . , m where xij = value of jth feature for ith data item. For big data, the values of n and m are large enough and the data is called the high dimensional data if m is high.

Hence challenges in big data mining starts from pre-processing of data till the result we get. Advanced techniques and algorithms are developed and are continuously in process of improvisation for big data analytics. The focus of this paper is directed towards overview on big data clustering methods with main focus on widely used k-means clustering algorithm. Section 2 covers background for clustering Section 3 discusses big data clustering methods. Section 4 is about K-Means Clustering and its limitations. In Section 5 semi supervised clustering is contended. Analysis and Findings are discussed in Section 6. In Section 7, Proposed Methodology is defined and Conclusion with Future Aspects is drawn in Section 8.

II. BACKGROUND

Today's organizations demand accurate and efficient process to analyze big data by producing decisions more informative and useful. Since it is not possible to manage and process the big data only using traditional data mining techniques, advanced tools and techniques are continuously used and improved. Some advanced techniques such as machine learning, fuzzy systems and artificial intelligence play vital role with big data analytical techniques.

As most of data is noisy, inconsistent and incomplete, before applying data mining techniques, pre-processing of data is required to clean, slice and normalize [4] the data. Cleaning [4] is the process to remove outlier, inconsistent and missing values in the data. By Slicing [4] process, dimensionality of data is reduced by converting cleaned data

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 2, Issue 11, DOI: 10.29027/IJIRASE.v2.i11.2019, 362-367 , May 2019

into horizontal and vertical format. PCA, subspace clustering, Vertical segment are some techniques to reduce data from high dimensionality to low. Generally it is frequent to scale the data in [0,1] or [-1,1] scale. Normalization is for scaling data to bring within specified range. Through survey it is found that min-max technique, Z-score and Decimal scaling technique are used for normalization.

With Min max normalization data transformation takes place at [0,1] scale This type of transformation is linear transformation that preserve relationship among the original data. It also produces smaller standard deviations that overcome the effect of outliers.

Let the original value V to scale from range [m, M]to [m', M'] where m'=0 and M'=1

As per Min-Max normalization, scaled value is:

val'=(val-m)/(M-m)*(M'-m')+m'

As in min-max normalization, transformation of the data takes place at scale [0, 1] by Z-score normalization. Transformation is performed by mean and standard deviation of data. The unstructured data value can be normalized by Z-score normalization as follows:

v_i'=v_i-mean/STD

Where v_i= original data value

v_i'=z-score value of original data value of row j of ith column

Mean = $\frac{1}{n}\sum_{i=1}^{n} v_i$

STD= $\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(v_i - mean)^2}$

Decimal scaling is used to provide scaling between [-1,1] by using formula

$$v_i = \frac{v}{10^j}$$

Where v_i = normalized value

v = range of values

J= smallest integer Max($[v_i] < 1$

To drive decision making, big data is processed and analysed after normalization to unlock values inherited by data. In this paper we are taking into account most extensively used method -unsupervised clustering [5]. But after pre-processing, no unique algorithm for clustering can be applied for analysis of big data.

Data clustering is generally used for:

Hidden pattern: to find observations into data, formulate hypotheses, detect anomalies, and identify pertinent features.

Inherent classification: to identify the degree of similarity among forms or organisms

Compression: as a process for manage the data and summarizing it through cluster prototypes.

Main application areas for clustering are social media, market research, healthcare, image processing etc. Hidden patterns can be achieved by classifying data into uniform groups consisting of homogeneous data while different groups contains heterogeneity in data as much as possible by decreasing intra-distance and increasing inter-cluster distance among data. Two varieties of clustering: Hard clustering versus soft clustering. In Hard clustering, each data point lie in one and only one cluster but in soft clustering, each data point may lie to one or more cluster according to probability of each data.

The distance between data is calculated by similarity measures. This parameter is used to group data points by maximizing similarity measures in same group and minimizing between different groups and to obtain high quality clusters. Similarity measure parameters are used to group data points. Different types of similarity measures are:

Cosine similarity: Cosine distance is calculated by the cosine of angle between two vectors. If A and B are two n-dimensional vectors then the angle between two vectors is

$$\alpha = Cos\ \frac{A.B}{|A||B|}$$

Euclidean Distance: Euclidean is most widely used standard distance metric. It is simply distance between two data points and is calculated by root of square differences between the coordinates of a pair of objects as

$$d_{xy} = \sqrt{\sum_{i=1}^{k}(x_{ik} - x_{jk})^2}$$

Manhattan Distance: It calculates the absolute difference the coordinates of a pair of objects as

$$d_{xy} = max_k|x_{ik} - x_{jk}|$$

Minkowiski Distance: It is generalized distance metric and data oriented approach. It is calculated by

$$d_{xy} = \lambda\sqrt{\sum_{i=1}^{k}(x_{ik} - x_{jk})^2}$$

If λ=1 then it is City Block Distance

If λ=2 then it is Euclidean Distance

III. BIG DATA CLUSTERING

Since big data cannot be processed on single machine hence

clustering techniques in terms of big data [6] is studied into two forms: single machine clustering techniques multi-machine clustering techniques. For single machine clustering techniques, five main types of clustering algorithms are Partitioning, Hierarchical, Density Based, Grid based and Model Based Clustering[6]. These algorithms still are facing technological challenges that proliferate with increasing complexity of algorithms.

• **In Partitioning Clustering**, data objects are divided into initially specified no. of clusters. Algorithm aims to minimize the distance of data objects from the centroid of nearest cluster. This process is repeated until no change in distance is found. One of most generally used partitioning algorithm is K-means algorithm .Its main characteristics are very efficient and logically simple. But it is not without having drawbacks that outcome are repeated and produce different results. Other partitioning algorithms are K-means++, PAM, CLARA, CLARANS[6].

• **In Hierarchical Clustering**, clusters are generated hierarchically means one cluster is dependent on another clusters. At starting point each data point is considered as an individual cluster. Then two nearest clusters are merged into single cluster. This terminates when there is only single cluster. But it is not found very efficient for big data clustering. Some Hierarchical clustering are BIRCH[7][8], AGNES, CURE,CLUB[9],CLUB+[10].

• **In Grid Based Clustering**, data sets are partitioned into cells to form grid and clusters are generated on the basis of grid structure. Subspace and hierarchical clustering is used to form clusters. Some grid based clustering are STING, CLIQUE,MAFIA

• **In Density Based Clustering**[11], Clusters are generated in dense area. This clustering is very efficient in developing clusters of any irrational shape and in detecting noise as well as outliers. DBSCAN(Density Based Spatial Clustering of Application with Noise), DENCLUE and OPTICS are density based algorithms .

• With difference as above algorithms, Clusters are formed in Model Based Clustering using Statistical approaches and Neural network approaches. EM, CLASSIT, SOM and COBWEB are some popular Model Based algorithms.

In contrast to single machine clustering techniques, Multi-machine clustering techniques are universally used. There are different machines to store and process the data and become efficient in case of big data analytics. As big data is huge data, it is splitted into small chunks and loaded on different machines for processing. But they have problem of scalability and requires more computing time and memory. The multi machine clustering can be performed using parallel clustering as well as Map reduces Based clustering. Parallelism can be obtained by independent parallelism; task level parallelism and Single program multiple data parallelism. Some types of Parallel clustering are K-mean algorithm, PBIRCH, PDBSCAN , ParMETIS. Map reduced Based clustering has two phases Mappers and Reducers, Mappers to create a set of intermediate key pairs and Reducer takes these intermediate key pairs as input, process them and produce final output which is stored in HDFS[12](Hadoop Storage file System).

IV. TRADITIONAL K-MEANS CLUSTERING

Clustering is exploratory due to finding pattern in the data. Most popular and oldest clustering algorithm is K-mean clustering algorithm. Traditional K-means clustering [13] is efficient in producing clusters for small data sets but currently data are produced by social media, sensors and geographical data. To handle this data, traditional K-mean is not sufficient to form clusters and find meaningful pattern insight. There is need to enhance and improve the algorithm so that it can work efficiently and produce result effectively.

There are some points of matter for enhancing the capability of K-means;
1) Determination of no. of desired clusters k
2) Way to select initial centroids
3) Assignment of data objects to nearest cluster by using some distance metrics
4) Only local search optimization hence solution to Minimization problem is NP-hard.
5) Sensitivity to presence of outliers.

Inputs: X={x1,x2.....xn} data items and No. of clusters k
Steps :

1. Define k centroids (equal to no. clusters) from datasets randomly.
2. Repeat steps until no more changes in centroid value
   - Calculate distance of each data item to each centroid by using Euclidean distance metric and assign each data item to that cluster having minimum distance with that cluster.
   - Update the centroid value of cluster by calculating the mean of each data items value and old centroid value.

The objective is to minimize squared error function as

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - C_j||^2$$

Where J=objective function
k= no. of clusters
n= no. of data items
C_j= centroid for cluster j

Variation of K-Means Clustering

In literature, researchers have proposed and implemented algorithms for betterment the performance of k-means clustering in terms of determination in no. of clusters, quality of clusters, cluster validity etc. There are numerous research papers and some has been discussed here. We

limited our research to find different methods to improve the selection procedure of number of clusters, cluster centroids and clustering for big data.

In view of solving problems for big data clustering, Author[14] used Hadoop for storing large data sets. A new algorithm Hadoop MapReduce standard K-means clustering algorithm KH-HMR is proposed to govern the large data sets up to peta level bytes and a new metric (KM-I2C) K-means inter and intra clustering is given for similarity measurements to increase similarity between intra-cluster objects and decrease similarity inter-cluster objects.

Suryawanshi et al.[15] focussed on way to find initial cluster size. They modified original k-mean algorithm by proposing the idea to divide total number of attributes by number of clusters to calculate initial centroids. Data sets are also normalized followed by sorting of data sets to improve the accuracy and reduce execution time.

Abdul Nazeer et al. [16] [improved k-means] proposed an enhanced algorithm to calculate initial centroids but no. of clusters k is still given as input to algorithm.

FAHIM A et al. [17] proposed a systematic method in which some heuristic approach is used for less calculation in next iteration to find initial centroids. i.e. in each iteration the centroid closer to some data objects remain in the cluster and those far apart from the other data objects, may change their cluster. So there is no need to find its distances to other cluster centroids. This is simple and efficient clustering algorithm based on the k-means algorithm.

Aletti et al.[18] here propose a local metric based on Mahalanobis distance to find clusters in data which is calculated in real time.

Unnati R. et al.[19] proposed a methodology to improve the efficiency and accuracy of k-mean algorithm. The methodology is based on two phases by deriving initial centroids and assigning data points to nearest clusters.

V. SEMI SUPERVISED CLUSTERING

This section is related to only semi supervised clustering [14]. In many cases, it may be the possibility to have large amount of unlabeled data but some labelled data. Labelled data can be either partially labelled data or known relationship between some observations. Semi-supervised clustering in contrast with unsupervised clustering exploit some supervised data either in the form of class labels or pair wise constraints. Three features of semi supervised clustering algorithm are partially labelled data, known relationship between observations and clusters associated with outcome variable.

Clustering based on Partially Labelled data:

In the case of partially labelled data, unlabelled data is classified into clusters using known set of data cluster assignment. Basu et al. [20][21] modified k-means algorithm as they calculate initial cluster mean of each feature including cluster value and then by assigning labelled data to known cluster in spite of having minimum distance with other clusters and unlabelled data is assigned only according to minimum distance measure. This algorithm is known as constraint k-mean clustering algorithm. But in this case if any labelled data is misclassified, it can not be corrected in any way. Seeded k-mean is similar to tradition k-mean clustering. It uses labelled data for selecting initial clusters. Clustering using partially labelled data is useful in gene classification.

Clustering based on Known constraints among observations:

In this method, complex relationships in the form of constraints among observations are considered and Must-link constraint and Cannot-link constraints are two pair wise constraints that are included in clustering algorithms. Observations having must-link constraints must be allocated same clusters, while having can-not link constraints must not be allocated same clusters. This can be applied with repeated measurement taken on some subset of data of any experimental unit. Wagstaff et al.[20][22] proposed an algorithm COP-KMean that nominates each observation to nearest cluster with no violation in constraints. But there may be the possibility of constraint violation. To solve this problem Basu et al. recommended PCK-Means algorithm that exploit pair wise constraint with granting some violation. COP-Kmean and PCK-Means are modified algorithms of K-Means algorithm with constraint satisfaction and called as 'Constraint based Methods'.

Clustering based on outcome variable:

Traditional clustering do not have any option to cluster associated with an outcome variable of interest. Thus special techniques are needed to identify the cluster of data having outcome variable. Only few methods are in existence to cluster associated with outcome variable. One of the earliest method is supervised clustering given by Bair and Tibshirani[20][23]. Another method proposed by Koestler et al.[20][24] is semi-supervised recursively partitioned mixture models (RPMM). In continuation of this Gaynor and Bair [20][25] propose a method called "supervised sparse clustering," which is a modification of the "sparse clustering" method of Witten and Tibshirani[20][26]. Known outcome variable is used as seeding in the first step of the sparse clustering method and the remainders of the sparse clustering algorithm are repeated without further consideration of the outcome variable.

VI. ANALYSIS

We applied traditional K-means clustering on well known Iris data set. The data set contains 3 classes of iris plant –iris satosa, iris-virginica and iris-versicolor with 50 instances of

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 2, Issue 11, DOI: 10.29027/IJIRASE.v2.i11.2019, 362-367 , May 2019

each and tried to find out the improper labeling due to unsupervised clustering.

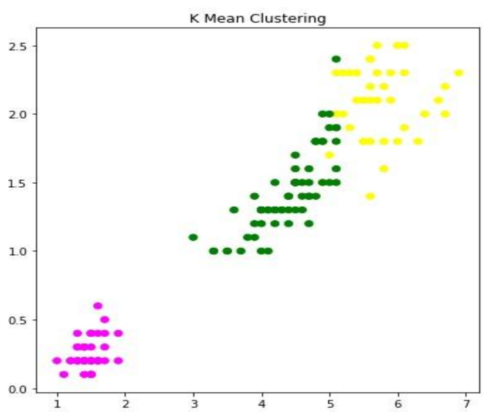
The result is shown as below

The recorded accuracy is 0.8933333333333333 and the confusion matrix is

[[50 0 0]
 [0 48 2]
 [0 14 36]]

This matrix shows that all 50 classes of label 0(Type-Satosa) have been classified correctly. 48 classes of label 1(Type-Virginica) have been classified correctly but 2 classes are misclassified as label2. 36 classes of label 1(Type-Versicolor) have been classified correctly but 14 classes are misclassified as label1.

```
In [8]: runfile('C:/Users/Toshiba/.spyder-py3/iris1.py', wdir='C:/Users/Toshiba/.spyder-py3')
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 2 2 2 2 0 2 2 2 2
 2 2 0 0 2 2 2 2 0 2 0 2 0 2 2 0 0 2 2 2 2 2 0 2 2 2 2 0 2 2 2 0 2 2 2 0 2
 2 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2
 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2
 2 1]
0.8933333333333333
[[50  0  0]
 [ 0 48  2]
 [ 0 14 36]]
```



This misclassification of data can be improved by using semi-supervised clustering –seeded k-means algorithm. Traditional k-means clustering picks initial centers randomly but by using seeded K-means we can provide some labeled data as initial centers and can improve the performance of k-means clustering.

VII. PROPOSED METHODOLOGY

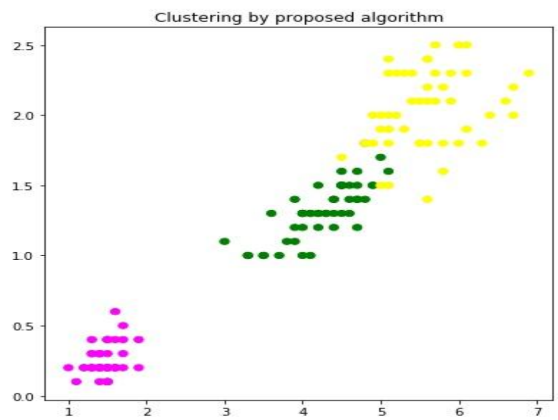
The proposed methodology has based on both semi-supervised and unsupervised clustering to improve the above misclassifications. It has two aspects as Improvisation of K-Mean(as using sorted data) and Selection of initial centroids(as Seeded K-mean)

Step 1: Input data sets
Step 2: Sort the input dataset to process fast
Step 3: Select subset from set of inputs as labeled data
Step 3: Provide these values as initial seed for selection of centroids in K-mean algorithm
Step 4: Repeat the steps of K-means clustering to optimize the objective function.

After applying this algorithm, the accuracy was 0.9243433333333333 and confusion matrix as below:

```
0.9243433333333333

[[50 0 0]
[0 48 2]
[0 10 40]]
```



VIII. CONCLUSION AND FUTURE WORK

Semi supervised clustering is achieving greater attention among data mining and machine learning community. It provides better cluster results through incorporation of some labeled data and background information. Several semi supervised algorithms have been developed to intensify the quality of clusters by accepting supervision either as constraints or known labels but still there is scope for improvements in this field. Some aspects but not limited for improvements are automatic selection of number of clusters, model selection criteria using both unsupervised data and limited supervised data, usage of common and uncommon features, generating overlapping clusters etc. In our future work, we will apply proposed methodology on different big data sets and try to improve the accuracy and other parameters.

REFERENCES

1. Elgendy,Nada., Elragal,Ahmed.,: Big data analytics in support of decision making process . Procedia Computer Science(100) , 1071 – 1084 (2016).
2. Gartner.: Big data [Online]. Available: http://www.gartner.com/it-glossary/big-data/. (2016).
3. Khine, Phyu, Pwint., Wang, Shun, Zhao.,:Data lake:A New Ideology in Big Data Era, https://www.researchgate.net/publication/321825490 (2017).
4. Golov,Nikolay., Ronnback, Lars., : Big Data Normalization for massively processing databases Science Direct, Computer Standards & Interfaces 00, 1–12 (2017).
5. Venkatkumar, Aurobind,Iyer., Shardaben, Kondhol, Jayantibhai,Sanatkumar.,: Comparative study of Data Mining Clustering algorithms., 978-1- 5090-1281- 7/16/$31.00 IEEE (2016).
6. Chen,Min., Ludwig , A .,Simone., Li, Keqin.,: Clustering in Big Data Big Data Management and Processing Book Chapter 16“ pp. 333-345 (2017).
7. Zhang,Tian., Ramakrishnan,Raghu., Livny,Miron.,: BIRCH: A new Data clustering algorithm and applications, Kluwer Academic Publishers, Boston, pp.1-40 (1997).
8. Lorbeer,Boris., Kosareva,Ana., Deva,Bersant., Softic,Dženan., Ruppel,Peter., Küpper,Axel.,: Variation on clustering algorithm BIRCH Elseviser Big data research ,(2017)
9. Mazzeo,M., Giuseppe., Zaniolo,Carlo.,: The Parallelization of a Complex Hierarchical Clustering Algorithm: faster unsupervised learning on larger data sets UCLA CSD Technical Report No. 160001 (2016).
10. Mazzeo,M., Giuseppe., Masciari, Elio., Zaniolo, Carlo.,: A fast and accurate algorithm for unsupervised clustering around centroids Information Sciences 400–401,pp. 63–90 (2017).
11. Sunder, Reddy, Shyam, K., Bindu, Shoba ,C.,: A Review on Density-Based Clustering Algorithms for Big Data Analysis International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC),IEEE pp.123-130 (2017)
12. Hajeer ,Mustafa., Dasgupta ,Dipankar., :Handling Big Data Using a Data-Aware HDFS and Evolutionary Clustering Technique IEEE transactions on big data, (2016).
13. Jain, K., Anil .,: Data clustering: 50 years beyond K-means Pattern Recognition Letters 31, 651–666 (2010).
14. Sreedhar, Chowdam., Kasiviswanath, Nagulapally., Reddy, Reddy,Pakanti., : Clustering large datasets using K means modified inter and intra clustering (KMI2C) in Hadoop, DOI 10.1186/s40537-017-0087-2, Journal of Big Data (2017).
15. Suryawanshi, Rishikesh., Puthran ,Shubha.,: A Novel Approach for Data Clustering using Improved K-means Algorithm International Journal of Computer Applications (0975 – 8887)(142), (2016).
16. KA Abdul, Nazeer., Sebastian., P.,M.,: Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, In Proceedings of the World Congress on Engineering,(1), pp. 1-3. (2009).
17. Fahim, A. M., A. M. Salem, F. A. Torkey, and M. A. Ramadan. "An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University SCIENCE A 7, (10),1626-1633, (2006).
18. Aletti,Giacomo., Aletti , Micheletti ,Alessandra., Micheletti, : A clustering algorithm for multivariate data streams with correlated components Journal of Big Data , https://doi.org/10.1186/s40537-017-0109-0, (2017).
19. Raval, R.,Unnati., Jani, Chaita.,: Implementing & Improvisation of K-means Clustering Algorithm,IJCSMC, (5), Issue., pp.191 – 203, (2016).
20. Bair,Eric., Semi-supervised clustering methods, http://arxiv.org/abs/1307.0252v1 (2013).
21. Basu, S., Banerjee, A., R, Mooney,.: Semi-supervised clustering by seeding. In Proceedings of the 19th International Conference on Machine Learning (ICML-2002) 19–26, (2002).
22. Wagstaff, K., Cardie, C., Rogers, S., Schrodl, S.,: Constrained k-means clustering with background knowledge. In Proceedings of the 18th International Conference on Machine Learning (ICML-2001) 577–584 (2001).
23. Bair, E., Tibshirani, R.,: Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol. 2(4):e108. doi:10.1371/journal.pbio. 0020108, (2004).
Koestler, DC., Marsit, CJ., Christensen, BC., Karagas, MR., Bueno, R., Sugarbaker, DJ., Kelsey, KT., Houseman, EA.,: Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. Bioinformatics. 26(20):2578–2585. doi:10.1093/bioinformatics/btq470. (2010)
24. Gaynor, S., Bair, E.,: Identification of biologically relevant subtypes via preweighted sparse clustering. ArXiv e-prints. arXiv:1304.3760. http://arxiv.org/abs/1304.3760 (2013).
25. Witten, DM., Tibshirani, R.,: A framework for feature selection in clustering. Journal of the American Statistical Association. 105(490):713–726. doi: 10.1198/jasa.2010.tm09415 (2010).