IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-11, Nov 2019 E-ISSN: 2347-2693

Implementation of K-Means Clustering in Big Data Environment

Ayush Gupta[1*], Pratik Gite[2]

[1]Computer Science and Engineering, IET DAVV, Indore, India
[2]Computer Science and Engineering, IES-IPSA, Indore, India

*Corresponding Author: ayushgupta9800@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i11.3844 | Available online at: www.ijcseonline.org

Accepted: 10/Nov/2019, Published: 30/Nov/2019

***Abstract*—**In recent years the digital data is grown much frequently. Handling and processing of such bulky data are much complex and need the attention of a human. Moreover, the existing techniques and methods are not much suitable to deal with this complex nature of computation. To deal with such a complex nature of computation, the big data analytics played an essential role. In this presented work the unsupervised learning technique namely k-means clustering is implemented initially and their performance is measured. During this to enhance the performance of the system a new modified k-means clustering algorithm is proposed by improving the centroid selection technique and using the RBF kernel. The comparative performance analysis of both the versions of k-means clustering demonstrate the modified k-means clustering is efficient and has the low algorithm run time. Therefore it is a promising approach for analytics, thus it's a future extension that is also presented in this work.

***Keywords*—**Big Data, Big Data Analytics, Unsupervised learning, Clustering Algorithm, improvements**.**

I. INTRODUCTION

Big data analytics is a complex process of analysis of huge and heterogeneous data, which is sometimes also known as big data. That is helpful to uncover information -- such as hidden patterns, unknown relations, trends, and preferences. It helps organizations in making decisions. Data analytics techniques provide a way for analysis of data and find a model for decision making or forecasting. Big data analytics is an advanced feature for data analytics. That involves applications i.e. prediction, statistics, and other analysis techniques. Driven by analytics systems and high-powered computing, big data analytics provide different advantages, such as [1]:

- New opportunities
- Marketing
- Customer service
- Competitive trend analysis

Big data analytics enables us, to analyze growing volumes of data and other forms of data. This data is a mix of semi-structured and unstructured data i.e., click-stream data, web server logs, social media, and others. Unstructured and semi-structured data have not fit properly in a traditional manner. Additionally, data warehouses are not able to handle the processing of large volumes of data. That needs to be updated continuously. As a result, various organizations that collect, process and analyze big data usages NoSQL, as well as Hadoop and its data analytics tools [2] [3]:

- **YARN:** a cluster management technology and it is a feature in second-generation Hadoop.
- **MapReduce:** a framework that allows developers to write programs that process massive amounts of unstructured data across a distributed cluster.
- **Spark:** an open-source, parallel processing framework that enables users to run large-scale data analytics applications.
- **HBase:** a column-oriented key/value data store built to run on top of the Hadoop Distributed File System (HDFS).
- **Hive:** an open-source data warehouse for querying and analyzing large data stored in Hadoop files.
- **Kafka:** a distributed messaging technique that is implemented for bypassing the message brokers.
- **Pig:** that technology is used for parallel programming for processing the MapReduce jobs.

However, big data analytics are utilizing the potential of Hadoop data storage for developing the repository to host the incoming data streams. In this architecture, data are directly analyzed in a Hadoop cluster or executed using a processing engine. Data is stored in the HDFS, is prepared, configured and partitioned to get effective performance for the cases of data extraction, transformation, and load integration processes and other user queries [4]. When the data is prepared or preprocessed using these techniques it may be used with the different analytical techniques such as:

- Data mining helps recover the essential patterns and relationships among data;
- Predictive analytics, allow us to develop different computational models for predicting customer behavior and others;
- Machine learning, that can also be used for dealing with the huge dataset instances with fewer efforts
- Deep learning, is a comparatively new technique for learning on essential patterns using filters and others.

Text mining and statistical analysis techniques can also play a role in big data analytics. For ETL and analytics applications, MapReduce helps to write queries, using programming languages for relational databases that are used with SQL-on-Hadoop [5].

II. RELATED WORK

This section provides different research articles and important contributions that are developed to provide guidelines for newer system development.

Advanced unsupervised learning techniques are an emerging challenge in big data due to the increasing requirements of extracting knowledge from a large amount of unlabeled heterogeneous data. LingyunXiangi et al [6] address a fast unsupervised heterogeneous data learning algorithm, namely two-stage unsupervised multiple kernels extreme learning machine (TUMK-ELM). TUMK-ELM extracts information from multiple sources and learns the heterogeneous data with closed-form, which enables its extremely fast speed. TUMK-ELM has low computational complexity, and the iteration of its two stages can be converged within finite steps. As an experimental demonstration on 13 real-life data sets, it gains a large efficiency as compared with three state-of-the-art unsupervised heterogeneous data learning methods.

Deep network architectures such as cortical algorithms are inducing the big data issues such as lengthy and complex training. Nadine Hajj et al [7] present a distributed cortical algorithm for big data using unsupervised learning. That data is combined node-data parallelization. A data sparsity technique is employed for partitioning data before distributing attributes over the network for various computational nodes. The results of multiple datasets showed an average speed-up.

Big data allow Machine learning (ML) algorithms to uncover fine-grained patterns to make accurate predictions. Lina Zhou et al [8] introduce a technique based on ML for big data to offer opportunities and challenges. That is centered on ML and follows all the processes such as preprocessing, learning, and testing. This technique also includes the key components, i.e. big data, user, domain, and system. The system provides directions for the identification of relevant opportunities and challenges.

Deep learning is currently an extremely active research area. It has gained huge successes in a broad area of applications. With the sheer size of data, big data brings big opportunities and transforming potential. Additionally, it demonstrates extraordinary challenges to connect with data. As the size of data is being bigger, deep learning is playing an essential role in predictive data analytics. Xue-Wen Chen et al [9] provide an overview of deep learning, and current research efforts and challenges to big data, with future trends.

Intelligent fault diagnosis is a tool to deal with mechanical big data due to its ability to rapidly and efficiently processing collected signals and providing an accurate diagnosis. In traditional methods, the features are manually extracted and diagnostic expertise. Such processes take advantage of human ingenuity. By the motivation of unsupervised learning which learns from raw data, a two-stage learning process introduced by Yaguo Lei et al [10]. In the first stage, an unsupervised two-layer neural network with sparse filtering is used. In the next phase, soft-max regression is used for classifying the health circumstances based on learned attributes. The technique is evaluated using bearing datasets. The results show that the method obtains high diagnostic accuracy and superiority over traditional techniques. Because the learning features are adaptive, and it reduces human labor and offers easy fault diagnosis.

Abdelkarim Ben Ayed et al [11] propose to give a review of the most used clustering methods. First, give an introduction to clustering methods. Second, present the clustering methods with some comparisons mainly partitioning clustering methods like k-means, Gaussian Mixture Models and variants, the hierarchical clustering like agglomerative algorithm, fuzzy clustering, and Big data clustering methods. They offered examples of clustering techniques and present ideas to make scalable and noise insensitive clustering based onfuzzy type-2.

In recent years a significant amount of data is collected from different data sources. Additionally, it is represented using different views. The term views are used describing the perspectives or utility of data. Each utility used for finding patterns using clustering techniques. The core issue is in combining heterogeneous features for unsupervised big data clustering. Xiao Cai et al [12] propose a big data multi-view clustering algorithm for the integration of heterogeneous data. They evaluate the new method using six different publically available data sets and conduct a comparative study among several clustering techniques. The results, of proposed methods, consistently achieve superior performances.

The main aim of data categorization is to prepare clusters of similar objects. A major issue in the clustering of big data is confusion due to a lack of consensus in the definition of features. AdilFahad et al [13] introduces theory and algorithms related to clustering. Additionally, they are providing a comparative study of theoretical and practical perspectives of the clustering algorithm. They conducted experiments and compared them with the different algorithms using various big data sets. The clustering algorithm's performance is measured using various internal and external metrics.

III. PROPOSED WORK

The proposed work is motivated to study and explore the domain of big data analytics. In addition to designing the efficient and accurate unsupervised learning model for big data analytics. In this context, the aim is to implement and enhance the traditional k-means clustering algorithm for big data analytics. The classical k-means algorithm for clustering of the input dataset is described first. In this context, the following algorithm is used as given table 3.1. [14]

Table 3.1 k-means clustering

```
Input: N objects to be a cluster$(x_1, x_2 \dots x_n)$, number of clusters k;
Output: k clusters and the sum of dissimilarity between objects and nearest centroid;
Process:
    1. Randomly choose k objects as initial centroids $(m_1, m_2, \dots, m_k)$;
    2. Calculate distance between each object $x_i$ and centroid, then assign an object to the nearest centroid, distance calculated as:
       $\boldsymbol{d}(\mathrm{x_i}, \mathrm{m_i}) = \sqrt{\sum_{\mathrm{j=1}}^{\mathrm{d}} (\mathrm{x_i} - \mathrm{m_{j1}})^2}, \mathrm{i} = 1 \dots \mathrm{N}, \mathrm{j} = 1 \dots \mathrm{k}$
       $\boldsymbol{d}(x_i, m_i)$is the distance.
    3. Calculate mean of objects in each centroid to update centroid,
       $m_i = \frac{1}{N} \sum_{j-1}^{n_i} x_{ij}, i = 1,2, \dots, K$
       N is the number of samples;
    4. Repeat step 2 and 3 until the error is minimizing
```

In this existing clustering technique, we identified two major issues first random selection of the initial centroid can increase the running time of the algorithm. Second is the distance measuring technique help to improve the performance of the system. Therefore we include both aspects to improve the above-given system.

In this context first, we need to find the optimal centroids for clustering. Therefore the following process is used for centroid selection.

Table 3.2 centroid selection

```
Input: dataset D, number of clusters k
Output: k number of centroids C
Process:
    1. $R_n = ReadDataset(D)$
    2. $R_n = R_n.SortData$
    3. $\boldsymbol{min} = GetMin(R_n)$
    4. $\boldsymbol{max} = GetMax(R_n)$
    5. $\boldsymbol{I} = \frac{max+min}{k}$
    6. $\boldsymbol{temp} = 1$
    7. $\boldsymbol{for}(i = 1; i \leq k; i++)$
        a. $\boldsymbol{for}(j = temp; j \leq I; j++)$
            i. $S_i.Add(R_j)$
        b. $\boldsymbol{end}\ for$
        c. $\boldsymbol{temp} = temp + I$
        d. $\boldsymbol{I} = I + I$
    8. End for
    9. $\boldsymbol{for}(l = 1; l \leq k; l++)$
        a. $C_l = \frac{1}{S_l.length} \sum_{m=1}^{S_l.length} S_l$
    10. $\boldsymbol{end}\ for$
    11. Return C
```

According to the process given in table 3.2, the algorithm accepts two input parameters first the dataset to find the centroid and second the number of clusters k. first, the data is read and we recover max and min data object based on these two values first we compute the interval for the data. Using the computed interval the k number of sub-list is created. After the creation of sub-lists, the mean values of each sub-list are used as the centroids.

The next improvement is proposed by modification of distance function therefore to compute the distance we replace the existing distance function with the RBF (radial basis function). The RBF kernel function can be described as [15]:

$$k(x_i, y_j) = exp\left(\frac{1}{2\sigma^2} \|x_i - y_j\|^2\right)$$

In the above given function $\|x_i - y_j\|^2$ is the Euclidean distance, and σ^2 is variance of input data.

Therefore the modified algorithm can be described using the process as given in table 3.3.

Table 3.3 proposed k-means

```
Input: Dataset D, Number of clusters k
Output: clustered data

Process:
1. $R_n = ReadDataset(D)$
2. $C_k = calculateCentroid(R_n, k)$
3. $\mathbf{for}(i = 1; i \leq k; i++)$
    a. $\mathbf{for}(j = 1; j \leq n; j++)$
    b. $\mathbf{compute}\ kernel\ distance\ k(C_i, y_j)$
    c. $\mathbf{end}\ for$
4. End for
5. Compute mean of cluster to update centroid
6. $\mathbf{if}(error \rightarrow 0)$
    a. Assign objects to cluster centre
    b. Exit
7. Else
    a. Go to step 3
8. End if
9. Return clustered data
```

According to the above-given algorithm, the algorithm accepts similar parameters as the traditional k-means algorithm. After that first, the process of centroid selection is initiated as described in algorithm table 3.2. After computing the centroids the kernel-based distance matrix is prepared for assigning the clusters. Otherwise, the process is continuous until error→0.

In order to analyze the ability of the proposed system and to compare the performance with the classical k-means algorithm, the model is given in figure 3.1. the dataset of different data instances is hosted in the HDFS file system which is read first and spilled into parts first is training dataset which contains 70% of data attributes additionally the second part is containing 30% of the dataset which is used for testing of the system. Both the implemented algorithms process the training dataset for clustering and obtaining the optimally distributed cluster centers. The classical k-means algorithm is provided in table 3.1 and the second proposed modified k-means clustering technique is given in figure 3.2. After that, the test data is applied to the trained model for categorizing them and for performance evaluation. This section provides the efforts placed in this work, additionally, the next section provides a detailed understanding of the performance evaluation and their variation during the different sizes.

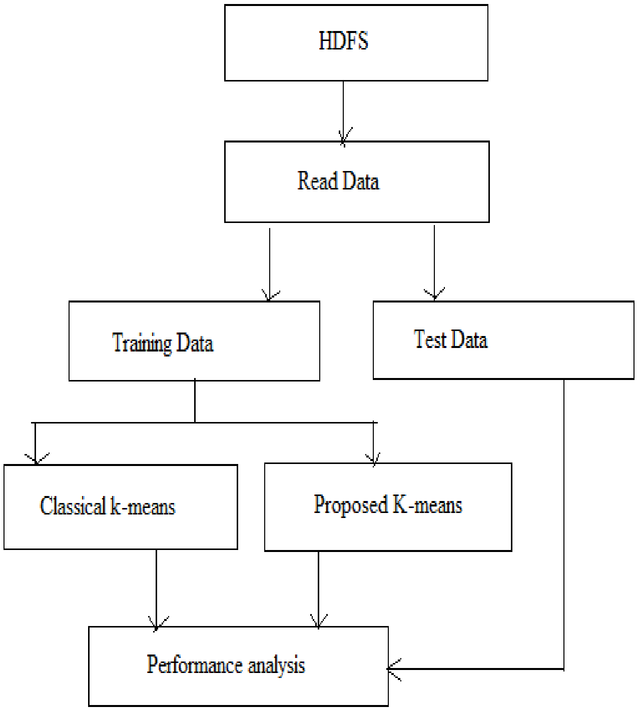

Figure 3.1 Proposed Model

IV. RESULTS AND DISCUSSION

This section provides the evaluation of the proposed method of big data clustering using the modified k-means clustering algorithms. The following performance parameters are used.

Accuracy

The accuracy of a machine learning algorithm is the ratio between total correctly recognized patterns and the total patterns are provided for classification. That can be measured using the following function.

$$accuracy = \frac{total\ correctly\ classified}{total\ input\ samples} X100$$

Figure 4.1 contains the line graph for demonstrating the performance of the k-means algorithm and its variant. The accuracy of both the algorithms is measured in terms of percentage (%). The blue line of the graph shows the performance of the proposed k-means clustering algorithm additionally the red line shows the performance of the traditional k-means clustering algorithm.

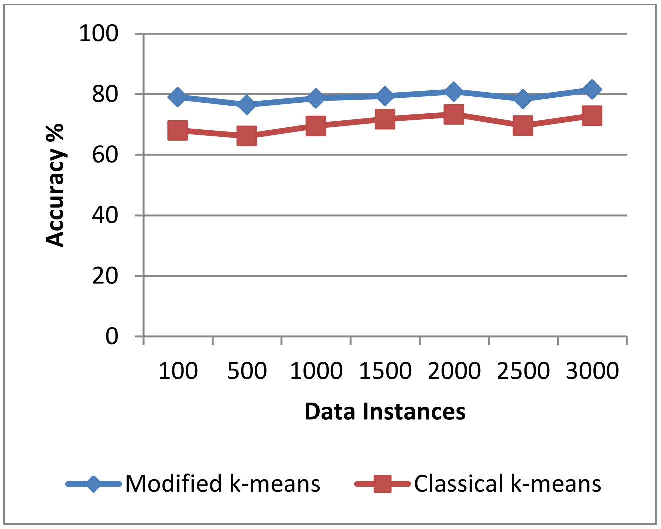

Figure 4.1 accuracy

According to the obtained experimental results, the proposed modified clustering algorithm shows improved performance as compared to the traditional algorithm.

Time usages

The time consumption is also known as the time complexity of the algorithm. That is the amount of time which is utilized during the execution of the algorithm. This parameter is computed based on the following function.

$$time\ consumption = end\ time - start\ time$$

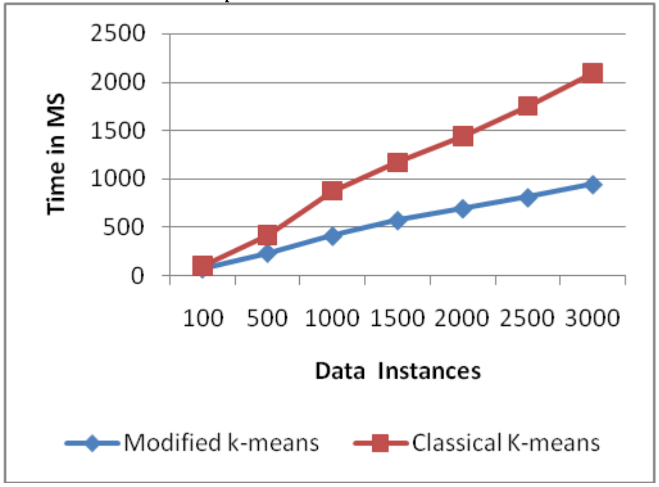

Figure 4.2 time complexity

The performance of both the implemented algorithms is measured here in terms of milliseconds (MS). The proposed modified k-means clustering algorithm is given here using a blue line and the traditional technique's performance is described using the red line. Figure 4.2 contains the performance line graph between the implemented techniques. According to the obtained results the proposed technique outperforms as compared to the traditional technique.

Memory usages

Memory usages are an essential parameter of the algorithm's performance measurement. The memory usages are also known as space complexity. Thus that is the amount of main memory which is utilized during the algorithm execution. The following function can be used for measuring the memory usages of an algorithm.

$$memory\ usages = total\ assigned\ memory \\ - free\ memory$$

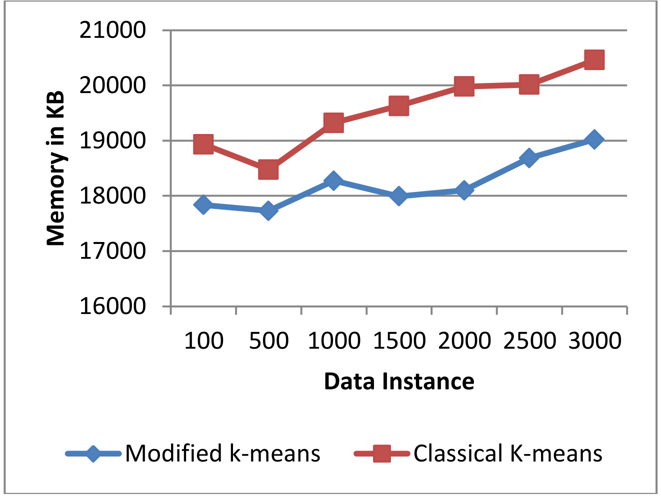

Figure 4.3 memory usages

The memory usages of the proposed work and the traditionally available k-means clustering algorithm in a big data environment are demonstrated in figure 4.3. According to the obtained results, the proposed technique's performance is given using the blue line and the traditional technique is demonstrated using the red line. The X-axis of the diagram shows the total number of data objects in the dataset and the Y-axis shows the obtained memory consumption for both the techniques in terms of KB (kilobytes). According to the observed results, the main memory usages of the proposed modified algorithm are low as compared to traditional techniques.

V. CONCLUSION AND FUTURE SCOPE

The data mining techniques enable us to analyze the data and recover the essential patterns from data. These techniques when clubbed with the big data technology then they are termed as the big data analytics. The big data technology supports significant kinds of data and their formats. Additionally, it also supports the supervised as well as unsupervised learning techniques. The proposed work is currently focused on the unsupervised learning technique more specifically the k-means clustering for categorizing the unlabelled data. The k-means clustering algorithm is a partition-based clustering algorithm which clusters according to their similarity and differences in automatic mode.

In this work, a modified k-means algorithm is proposed for design and implementation. Therefore the two key modifications on the existing k-means clustering technique are adopted, first the centroid selection process and second the distance measuring function. In order to find centroids, we proposed a centroid selection algorithm, additionally for computing the distance matrix for clustering the Euclidean distance is replaced with the kernel-based distance. Thus the RBF (radial basis function) is employed for measuring the distance in the clustering algorithm.

The proposed technique is implemented in the Linux operating system and with the help of JAVA technology. Additionally for preserving the measured performance parameters, the MySql database is used. After the implementation of the required system, the performance is measured which is summarized in table 5.1 as the performance summary.

Table 5.1 Performance Summary

S. No.	***Parameters***	***Improved k-means***	***Classical k-means***
1	***Accuracy***	***High***	***Low***
2	***Memory usage***	***Low***	***High***
3	***Time consumed***	***Low***	***High***

According to the given performance summary the proposed modified k-means clustering technique produces optimal results as compared to the traditional clustering approach.

Future work- The proposed technique is used for big data analytics purposes for handling the bulk amount of data with minimum efforts. Therefore an unsupervised learning algorithm is modified for improving the performance for big data analytics. According to the measured performance, the proposed modified k-means clustering algorithm improves the performance with respect to the traditional algorithm. Thus it is a promising approach and in the near future, it can be extended for the following aspects.

1. Implementing an heuristics-based unsupervised learning technique
2. Improving the technique for stream data mining for big data analytics

ACKNOWLEDGMENT

I am grateful to Dr. Pratik Gite, Assistant Professor, IPS Academy, Institute of Engineering and Science, Indore, for his encouragement and suggestions in carrying out this research work. I like to remember the motivation initiated by my respected Father and mother. . I wish to thank faculty members of the CSE department of IET DAVV, Indore for their constant co-operation, directly inspiration and encouragement.

I deeply express my heartfelt thanks to the IJCSE and Publications team for publishing this paper in such a beautiful getup and well in time.

REFERENCES

[1] R. H. Hariri, E. M. Fredericks, K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges", J Big Data (2019) 6:44, https://doi.org/10.1186/s40537-019-0206-3

[2] A. Patel, M. Jaiswal, R. K. Chawda, "An Approach to Predict Train Delay Using Big Data Analytic Approaches", International Journal of Advanced Research in Computer and Communication Engineering, ISO 3297:2007 Certified, Vol. 7, Issue 3, March 2018

[3] Z. P. Reddy, P.N.V.S. P. Kumar, "Comparing the Word count Execution Time in Hadoop & Spark", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 10, October 2016, ISSN (Online) 2348 – 7968

[4] F. C. Yayah, K. I. Ghauth, C. Y. Ting, "Adopting Big Data Analytics Strategy in Telecommunication Industry", Journal of Computer Science & Computational Mathematics, Volume 7, Issue 3, September 2017, DOI: 10.20967/jcscm.2017.03.002

[5] C. L. P. Chen, C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences 275 (2014) 314–347

[6] L. Xiangi, G. Zhao, Q. Li, W. Hao, F. Li, "TUMK-ELM: A Fast Unsupervised Heterogeneous Data Learning Approach", VOLUME 6, 2018, 2169-3536, 2018 IEEE

[7] N. Hajj, Y. Rizk, M. Awad, "A MapReduce Cortical Algorithms Implementation for Unsupervised Learning of Big Data", Procedia Computer Science, Volume 53, 2015, Pages 327–334, 2015 INNS Conference on Big Data

[8] L. Zhou, S. Pan, J. Wang, A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges", Neurocomputing 237 (2017) 350–361

[9] X. W. Chen, XIAOTONG LIN2, "Big Data Deep Learning: Challenges and Perspectives", Vol. 2, 2014, 2169-3536, 2014 IEEE

[10] Y. Lei, F. Jia, J. Lin, S. Xing, S. X. Ding, "An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data", 0278-0046 (c) 2015 IEEE.

[11] A. B. Ayed, M. B. Halima, A. M. Alimi, "Survey on clustering methods: Towards fuzzy clustering for big data", 978-1-4799-5934-1/14/$31.00 ©2014 IEEE

[12] X. Cai, F. Nie, H. Huang, "Multi-View K-Means Clustering on Big Data", Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence,

[13] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", Vol. 2, No. 3, Sep. 2014, 2168-6750 2014 IEEE

[14] S. S. Chouhan, R. Khatri, "Data Mining based Technique for Natural Event Prediction and Disaster Management", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.14, April 2016

[15] B. Feizizadeh, M. S. Roodposhti, T. Blaschke, J. Aryal, "Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping", Arab J Geosci (2017) 10:122, DOI 10.1007/s12517-017-2918-z

Authors Profile

Mr. Ayush Gupta pursued Bachelor of Engineering from Institute of Engineering and Technology, Indore in Information Technology. His main research work focuses onBig Data Analytics, Data Mining, machine learning and cloud computing based education. He has 18 months of professional experience and 1 year of Research Experience.

Destiny drew Dr. Pratik Gite towards the computer education in 2012. He has completed his B.E. and M.E. from RGPV University Bhopal (M.P.). He holds a Ph.D. in Wireless Mobile Ad-hoc Network from Pacific Academy of Higher Education & Research University, Udaipur, Rajasthan. He started his academics at LKCT Indore (MP). Dr. Pratik Gite has published 18 research papers in various national, international journals and conferences. He has a passion for writing computer engineering books. He is an expert of various computer technologies. He has worked with many computer programming languages and open source technologies. His areas of interest include Mobile Ad-hoc Network, Software Engineering, Computer Network, Basic Computer Engineering, Big Data Analytics and Information Technology. His current affiliation includes Asst. Prof. at IES-IPS Academy, Indore (M.P.). His contribution includes active participation in various research projects and research papers (IEEE). He can be reached at pratikgite135@gmail.com.