IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Development of Privacy Preserving Clustering Process with Cost Minimization for Big Data Processing

S. Chitra[1*], R. Bharanidharan[2]

[1,2]Dept. of Computer Science and Engineering, Vinayaka Missions's Kirupananda Variyar Engineering College Salem – 636 308, Tamil Nadu, India

Available online at: www.ijcseonline.org

Accepted: 14/Nov/2018, Published: 30/Nov/2018

***Abstract* -** Unfathomable quantum of comprehensive private data is habitually gathered as the mutual exchange of the corresponding information has come as a shot in arm for a multitude of data mining applications. The related data extensively encompass the shopping trends, criminal records, medical history, credit records and so forth. It is true that the corresponding information has proved its mettle as a vital asset to the business entities and governmental organization for the purpose of taking prompt and perfect decisions by means of assessing the pertinent records. However, it has to be borne in mind that harsh privacy. With an eye on effectively addressing the corresponding thorny issues, in this document, an earnest endeavor is made to kick-start a novel clustering Probabilistic Possibility Fuzzy C Means Clustering (PFCM) approach viz. The Big data processing, in fact, involves the explosive expansion of demands on evaluation, storage, and transmission in data centers, thus leading to incredible working expenses to be borne by the data center providers. To achieve this, we introduce VSSFA and Map Reduce Framework in Cloud environment. In this thesis we deeply develop a privacy preserving clustering process with cost minimization for big data processing.

***Keywords*-** Probabilistic Possibilistic Fuzzy C Means Clustering Algorithm, Neural Network, Privacy Preserving Data Mining, Fuzzy C-Means.

I. INTRODUCTION

Data over the internet has been rapidly increasing day by day. Automatically mine useful information from the massive data has been a common concern for the organizations having large dataset. Here, the privacy preserving is one of the larger concerns. The aim of privacy preserving calculations is to extricate significant learning from a lot of information while ensuring in the meantime delicate data. The unprecedented advancement in the Information and Communications technology brings with it the accelerated requirement for the safe storage and sharing of electronic data without being thrown open to the whims and fancies of the mischievous miscreants. The voluminous quantity of data, when made publicly accessible, may be effectively employed for the purpose of carrying out a host of intensive investigations. The Data Mining is considered as one of the modern techniques extensively employed to extort fruitful data from the gargantuan compilations of data. However, when it is published, it leads to the undesirable trend of exposing the susceptible data on the individuals concerned, culminating in the gross violation of ethical or privacy codes.

The cloud computing had its humble origin in the good old days of the 1960‘s when john McCarthy came up with the stunning declaration that _computation may someday be arranged as a public utility‘. The onset of the year 2006 witnessed the giant Amazon spearheading the cloud computing advancement by the dynamic activation of the Amazon web service on a utility basis [1]. In the current scenario , the cloud computing continues to hold sway as the shining star in the galaxy of the sophisticated technologies in the amazing arena of Silicon Valley fondly named the _Information Technology‘ and also in the thrilling realm of the _Research & Development‘. It has assumed office as an appealing technology well-equipped with the innate skills of accessing the network and segmenting the computing resources with the minimum possible executive effort. It is, indeed, one of the graceful techniques slated to take the cosmos by storm by redesigning the world and facilitating the information technology infrastructure to third party to be made available to the clients in the garb of commodities [2]. Figure 1 explains the integrated big data process. By means of the World Wide Web the computing scenario of the cloud computing can be hired to third parties to enable them effectively exploit the computing power or resources. The vital merit of the novel technology is the fact that it enables the relocation of the computing power and data from the personal computer and portable devices into titanic data centers. The clients are offered the convenience of accessing and making the fullest use of the entire services, totally unaware of the physical location and the organization of the system at the provider‘s end [3]. The cloud computing has been offered a red carpet welcome all through the universe as the most appealing candidate with the magical tool to present on a platter a cost-conscious and convenient access to the externalized IT (Information Technology) resources to the _byte-thirsty' mouse potatoes .

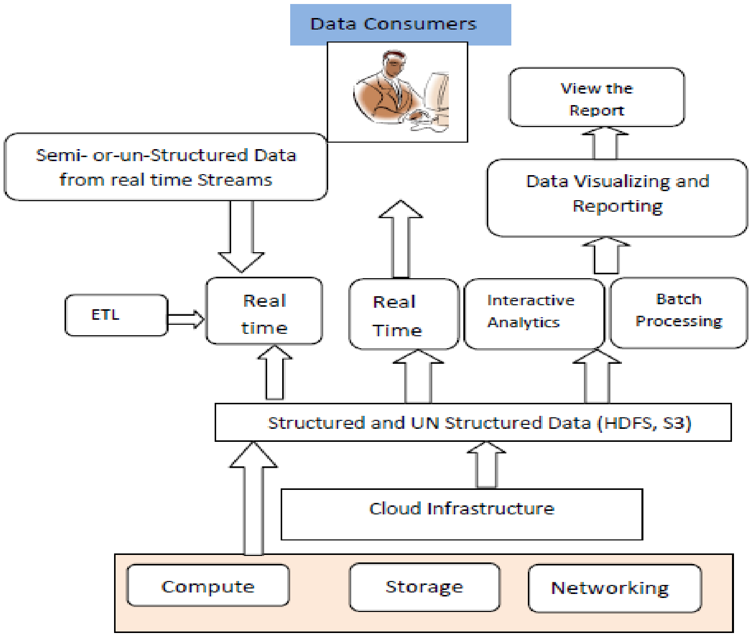

Figure 1: Integrated view of big data process

The zooming number of entities such as the investigation centers and firms take huge benefit out of the Cloud computing to host their varied applications. In fact, the cloud computing is well-geared to successfully tackle the identical physical infrastructure with an awe-inspiring client base with diverse computational requisites [4] by means of virtualization. Chapter 2 introduces the literature survey of Privacy preserving clustering process with cost minimization for Big Data Processing and chapter 3 explains Privacy Preserving Probabilistic Possibilistic Fuzzy-C Means Clustering approach. Chapter 4 explains the result and explanation and chapter 5 explains conclusion for proposed system.

II. LITERATURE REVIEW

A number of researches have been proposed by researchers for privacy preserving in big data. A detailed survey has been carried out to identify the various research articles available in the literature in all the categories of privacy preserving in big data, and to do the analysis of the major contributions and its advantages. Following are the literatures applied for assessment of the state-of-art work on the privacy preserving in big data. Here, forty five works has been analyzed.

Asmaa *et al.* [5] have explained the Protect Privacy of Medical Informatics using K-Anonymization Model. Here, they displayed a structure and model framework for de-distinguishing health data including both organized and unstructured information. They exactly examine a straightforward Bayesian classifier, a Bayesian classifier with an inspecting based strategy, and a contingent irregular field based classifier for removing distinguishing traits from unstructured information. They, convey a k-anonymization based system for de-recognizing the removed information to save most extreme information utility.

In this section we explained the big data classification. Among the big data classification Pekka Paakkonen and Daniel Pakkala proficiently designed the [6] reference architecture and classification of technologies, products and services for big data systems. A supplementary contribution was the classification of corresponding execution techniques and technologies and products/services, which was dependent on the assessment of the published use instances and survey of the associated work. Similarly, Isaac Triguero *et al.* [7] intelligently tabled the Map Reduce Solution for Prototype Reduction in Big Data Classification. The technique aimed at characterizing the original training data sets as an abridged number of instances. Their vital objective was to accelerate the classification procedure and cut down the storage requisites and susceptibility to noise of the nearest neighbour rule.

Similarly, Bingwei Liu *et al.* [10] brilliantly launched the Scalable sentiment classification for the Big Data analysis employing the Naive Bayes Classifier (NBC). In the technique without resorting to the deployment of a standard library (e.g., Mahout), they performed the NBC to realize the fine-grain regulation of the appraisal process. Further, they were able to design the Big Data analyzing system for the purpose of the current investigation. In [11], Shan Suthaharan have explained specified problem of Big Data classification of network interruption traffic. Specifically, the document effectively explains the challenges associated with the integration of the supervised learning approaches, representation-learning methods, machine lifelong learning algorithms and Big Data techniques such as the Hadoop, Hive and the Cloud for successfully addressing the network traffic classification issues.

Similarly, Chhaya *et al.* [12] charismatically designed the Privacy Preservation Enriched Map Reduce for Hadoop Based Big Data Applications, and the corresponding techniques included the privacy characterization model, anonymizer for datasets, dataset update and privacy preserved data management. The innovative method empowered the data users with the skills to regain the datasets in its unrevealed versions which facilitates the user task dispensing with the need for publishing vital detail particulars regarding the original data. Moreover, Rongxing Lu *et al.* [13] have explained the privacy preserving in big data. They first formulate the general architecture of big data analytics, identify the corresponding privacy requirements, and introduce an efficient and privacy-preserving cosine similarity computing protocol as an example in response to data mining's efficiency and privacy requirements in the big data era. Moreover, Mehdi Sookhak *et al.* [14] have explained the securing big data storage in cloud.

III. PROPOSED SYSTEM

The probabilistic possibilistic clustering technique (PPFCM) is initiated on each and every segment, which generated the number of clusters for each segment. The PPFCM represents the amalgamation of the probabilistic clustering algorithm [15] with the possibilistic fuzzy c means clustering technique (PFCM) [16]. In addition, the PFCM constitutes a hybridization of Possibilistic c-means (PCM) and fuzzy c-means (FCM) which habitually keeps at bay several thorny issues of the PCM, FCM and the FPCM. The PFCM is equipped with the requisite skills of overwhelming the noise sensitivity deficiency of the FCM, and the coincident cluster challenge of the PCM in addition to removing the row sum restraints of the FPCM.

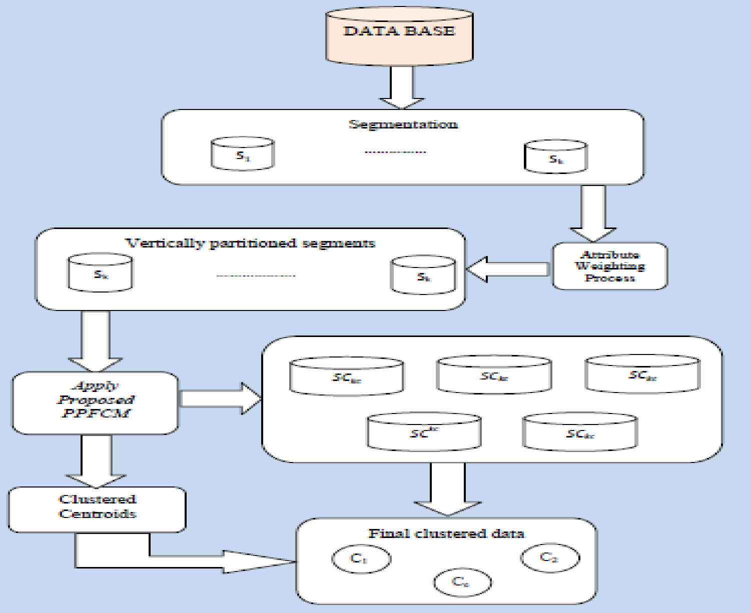

Figure 2: Overall diagram of the PFCM based privacy preserving

Fuzzy c-means (FCM): In the fuzzy clustering, data components are put into various clusters, and related with every component is an arrangement of membership levels. The FCM exquisitely utilizes the fuzzy partitioning in such a manner that a data point is competent to be a part of the entire groups with diverse membership grades ranging between 0 and 1. The objective function of the fuzzy c-means (FCM) clustering is represented in equation (6) shown below.

$$F_M = \sum_{i=1}^{c} \sum_{k=1}^{n} (V_{ik})^m \left\| Y_k - v_i \right\|^2 \quad \text{--- (1)}$$

Where;
c-The Number of clusters
n- The Number of data points.
V_{tk}- The relative fuzzy membership functions of Xk in class i.
m- Degree of fuzziness of the technique.

Possibilistic c-means (PCM):
The PCM algorithm has emerged as one of the important clustering techniques extensively employed to cluster the data. In the PCM technique, each cluster does not depend on another cluster which goes a long way in locating the noise points. Therefore, the objective function related to the cluster is configured as illustrated in Equation 2 appearing below

$$F_M = \sum_{i=1}^{c}\sum_{k=1}^{n} T_{ik}^{m} D^2\left(x_k, v_i\right) + \sum_{i=1}^{c}\eta_i \sum_{k=1}^{n}\left(1-T_{ik}\right)^m \quad \text{--- (2)}$$

Fuzzy possibilistic c-means algorithm (FPCM):
The FPCM technique brought to light by encompasses the possibility (typicality) as well as the membership values. It is a deft blend of a fuzzy partition and a possibilistic partition Equation 3.

$$F_M = \sum_{i=1}^{c}\sum_{k=1}^{n}\left(V_{ik}^{m} + T_{ik}^{n}\right)\left\| Y_k - u_i \right\|^2 \quad \text{--- (3)}$$

c □□Number of clusters
n □□Number of data points
A solution of the objective function is effectively accomplished by an iterative technique where the degrees of membership, typicality and the cluster centers are adjusted by method for the accompanying comparisons.

Possibilistic fuzzy c means clustering algorithm (PFCM):
In this type of clustering process, the PFCM representing an amalgam of the FCM and the PCM effectively employs the membership and typicality features from both the clustering techniques. Further, it is well-equipped with the skills of overwhelming the hurdles faced by both the FCM and the PCM approach, turning out a superb clustering process in an efficient manner. The objective function of PFCM is illustrated by means of the following Equation 4.

$$J_{PFCM}(U,T,V;Y) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left(aV_{ik}^{m} + bT_{ik}^{\eta}\right)\times\left\|Y_k - u_i\right\|_A^2 + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}\left(1-T_{ik}\right)^{\eta} \quad \text{--- (4)}$$

Probabilistic possibilistic fuzzy c means clustering (PPFCM):
The PPFCM represents the amalgamation of the probabilistic clustering algorithm [15] and the possibilistic fuzzy c means clustering algorithm (PFCM) [16]. In this method, at the outset, the number of cluster (C) is spelt out by the user, which is identical for each and every segment. After determining the number of cluster, it is followed by the evaluation of the distance between the centroids and data tuples for each and every segment. The steps involved in PPFCM are given in following steps;

Step 1: Calculation of distance matrix:
The Euclidian distance function as illustrated in Equation 20 is entrusted with the task of effectively estimating the distance between the centroids and data tuples to determine the distance matrix, which is calculated for each and every segmentation.

$$d\left(d_i, v_j\right) = \sqrt{\sum_{i,j=1}^{i=n,\, j=n}\left(d_i - v_j\right)^2} \quad \text{--(5)}$$

$$D(ss_k) = \begin{bmatrix} d_{11} & d_{12} & d_{1j} & d_{1C-1} & d_{1C} \\ d_{21} & d_{22} & d_{2j} & d_{2\,C-1} & d_{2C} \\ d_{i1} & d_{i2} & d_{i\,j} & d_{i\,C-1} & d_{i\,C} \\ d_{n-11} & d_{n-12} & d_{n-1\,j} & d_{n-1\,C-1} & d_{n-1\,C} \\ d_{n1} & d_{n2} & d_{n\,j} & d_{n\,C-1} & d_{n\,C} \end{bmatrix}$$

In an identical way, the distance function is utilized to determine the distance between the each data tuple with every centroid value and in the long run, the distance matrix is configured, as elegantly exhibited in the captioned figure 3.

Step 2: Calculation of probability matrix:
Subsequent to the evaluation of the distance matrix for each and every segmentation, the probability matrix is generated. The probability matrix possessing values of probability of data tuple in relation to each centroid is gathered from [15]. The probability value is effectively evaluated by means of Equation (22) shown below.

$$p_{i\,j} = \frac{\sum_{k=1}^{C} e^{d_{ik}}}{\sum_{l=1}^{C} e^{d\,l}} \quad (k \neq j) \quad \text{------- (6)}$$

By means of Equation 22 shown above, the probability value of each data tuple in relation to every centroid is determined and thereafter the probability matrix is built up as illustrated in equation (23).

$$P(ss_k) = \begin{bmatrix} p_{11} & p_{12} & p_{1j} & p_{1C-1} & p_{1C} \\ p_{21} & p_{22} & p_{2j} & p_{2\,C-1} & p_{2C} \\ p_{i1} & p_{i2} & p_{i\,j} & p_{i\,C-1} & p_{i\,C} \\ p_{n-11} & p_{n-12} & p_{n-1\,j} & p_{n-1\,C-1} & p_{n-1\,C} \\ p_{n1} & p_{n2} & p_{n\,j} & p_{n\,C-1} & p_{n\,C} \end{bmatrix}$$

The equation 23 characterizes the probability matrix of sanitized segment (Pss_k) where the value of P_{ij} indicates the probability of ith data tuple chance to move towards the jth centroid. Similarly, the distance function is employed to determine the distance between the each data tuple with every centroid value and in the long run the distance matrix is formed as depicted in Figure 3 exhibited above.

Step 3: Calculation of typicality matrix:
After the evaluation of the probability matrix, the typicality matrix is determined which is gathered from [16]. As illustrated in Equation 22 shown below, the probability value of each data tuple in relation to every centroid is determined and it is followed by the creation of the probability matrix which is furnished in equation (25).

$$t_{ij} = \frac{1}{\sum_{j=1}^{n}\left(\frac{d_{ik}}{d_{ij}}\right)} \quad 1 \le i \ge n, 1 \le k \ge C$$

---- (7)

$$T(ss_k) = \begin{bmatrix} t_{11} & t_{12} & t_{1j} & t_{1C-1} & t_{1C} \\ t_{r21} & t_{22} & t_{2j} & t_{2\,C-1} & t_{2C} \\ t_{i1} & t_{i2} & t_{ij} & t_{i\,C-1} & t_{i\,C} \\ t_{n-11} & t_{n-12} & t_{n-1\,j} & t_{n-1\,C-1} & t_{n-1\,C} \\ t_{n1} & t_{n2} & t_{nj} & t_{n\,C-1} & t_{n\,C} \end{bmatrix}$$

The equation (25) symbolizes the probability matrix of sanitized segment T(ss$_k$)where the value of T$_{i,j}$ indicates the probability of ith data tuple chance to move towards the jth centroid. In an identical way, the distance function is employed to evaluate the distance between every data tuple with every centroid value and at last, the distance matrix is created which is illustrated in Figure 3 appearing above.

Step 4: Calculation of membership matrix:
The evaluation of the membership matrix *V(ss$_k$)* is carried out by means of estimation of the membership value of data tuple which was determined with the help of Equation (26) given below, where the value of d_{ij} represents the distance of ith data tuple with respect to the jth centroid. The value of e^{dij} characterizes the exponential value of d$_{ij}$ and p_{ij} corresponds to the probability of d_{ij} . The clustering the data tuple is performed in relation to the membership value of the data tuple.

$$v_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{(p_{ij})^2\left(e^{d_{ij}}\right)+(d_{ij})}{d_{ik}}\right)} \quad 1 \le i \ge n, 1 \le j \ge C$$

--- (8)

$$V(ss_k) = \begin{bmatrix} v_{11} & v_{12} & v_{1j} & v_{1C-1} & v_{1C} \\ v_{r21} & v_{22} & v_{2j} & v_{2\,C-1} & v_{2C} \\ v_{i1} & v_{i2} & v_{ij} & v_{i\,C-1} & v_{i\,C} \\ v_{n-11} & v_{n-12} & v_{n-1\,j} & v_{n-1\,C-1} & v_{n-1\,C} \\ v_{n1} & v_{n2} & v_{nj} & v_{n\,C-1} & v_{n\,C} \end{bmatrix}$$

Step 5: Updation of centroid:
When the clusters are created, it is followed by the process of modernizing the centroids as illustrated in Equation (28) shown below.

$$u'_j = \frac{\sum_{i=1}^{n}(v_{ij} + t_{ij})Y_i}{\sum_{i=1}^{n}(v_{ij} + t_{ij})} \quad 1 \le j \ge C$$

--- (9)

After the centroids are modernized for the every segment, the succeeding step is to commence the procedure of evaluating the distance with the newly modernized centroids; it continues the procedure till the estimation of the modernization of centroids. Further, the relative process is replicated till the modernized centroids of each segment are identical in successive iterations.

IV. RESULT AND DISCUSSIONS

This section shows the outcomes got from the experimentation and its point by point discussion about the outcomes. The proposed methodology of PPFCM is explored different avenues regarding the Adult Datasets, mushroom dataset, plants datasets and seeds dataset. The result is evaluated with the probabilistic clustering [15] and possibilistic fuzzy c means clustering [16], accuracy and computation time. We have implemented our algorithm using Java (jdk 1.6) with cloud Sim tools and a series of experiments were performed on a PC with Windows 7 Operating system at 2 GHz dual core PC machine with 4 GB main memory running a 64-bit version of Windows 2007. The system is experimented with the four widely applied datasets namely, Adult Datasets, mushroom dataset, plants dataset and seeds dataset. These four benchmark datasets are taken from the UCI machine learning repository.

Here, the performance of the approach is explained based on the clustering accuracy. The methods proposed by probabilistic clustering [15] and possibilistic FCM [16] are the best known among existing schemes for clustering based privacy preserving. Furthermore, they characterize local details of the database such as clusters and centroids. Therefore, we have chosen to compare the performance of our approach against that of these ones.

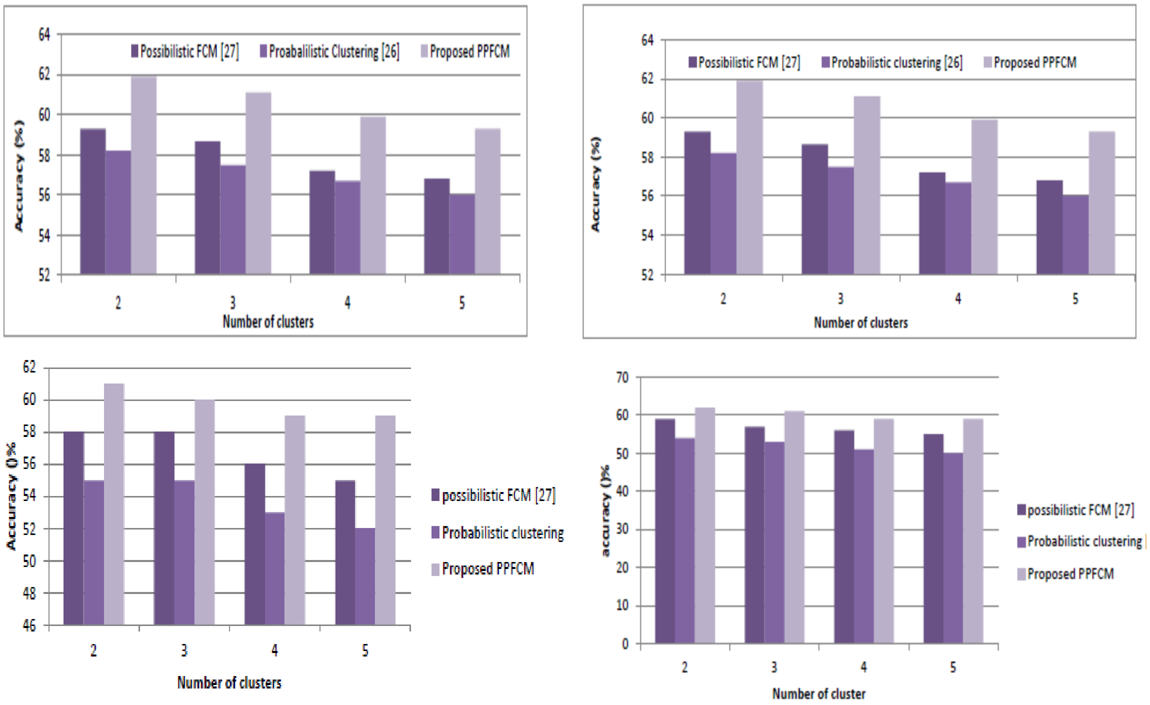

Figure 3: comparative analysis of clustering accuracy a) adult dataset b) mushroom dataset c) plants dataset d) seeds dataset

The above figure 3 a), b), c),d) shows the performance of the privacy preserving based on the clustering accuracy. Here the proposed approach PPFCM is the hybridization of the probabilistic clustering algorithm [15] and possibilistic fuzzy c means clustering algorithm [16]. The individual clustering algorithm is not performing well. Therefore we hybrid the two clustering algorithm. The above figure 3 shows the comparative analysis of clustering accuracy of adult dataset based on number of clusters.

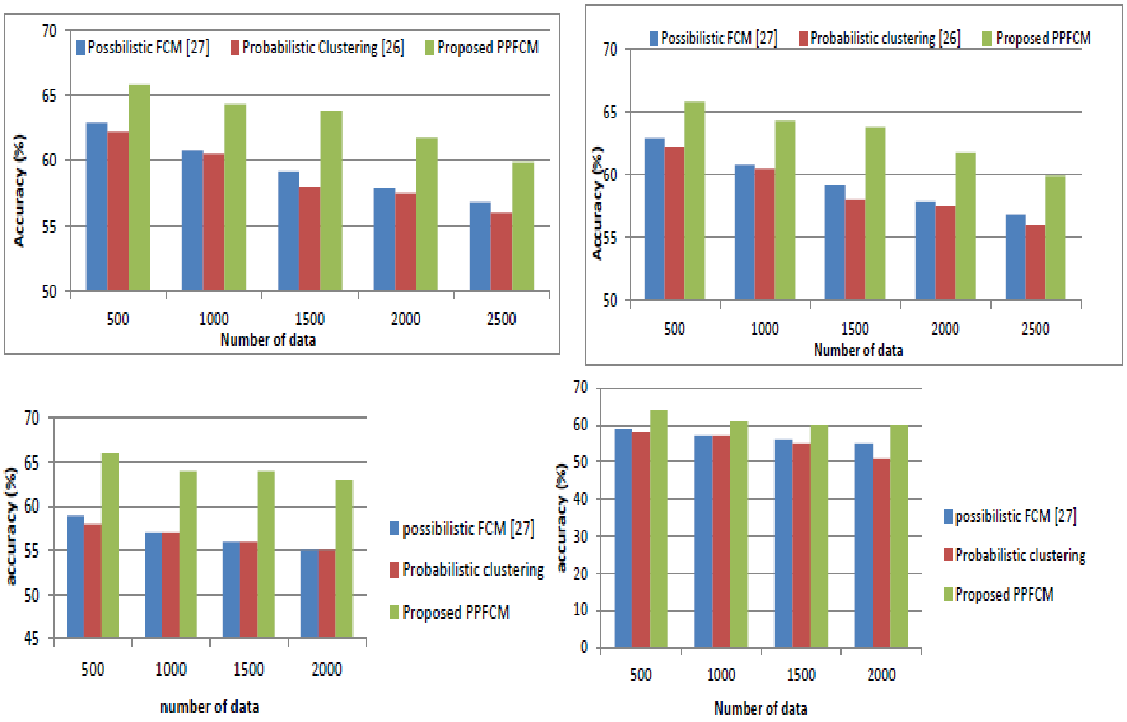

Figure 4: comparative analysis of clustering accuracy of a) adult dataset b) mushroom dataset c) plants dataset d) seeds dataset

From the above figure 3, the maximum accuracy is attained by proposed PPFCM is 61.9% for number of clusters as two and the minimum accuracy is attained by proposed PPFCM is 59.3% for number of clusters as five. The above figure 3.4 represents the accuracy of possibilistic clustering algorithm, probabilistic clustering algorithm and proposed PPFCM clustering algorithm. From the above figure 3, the maximum accuracy is attained by proposed PPFCM is 61.9% for number of clusters as two and the minimum accuracy is attained by proposed PPFCM is 59.3% for number of clusters as five. In figure 3 shows the performance of proposed approach using plants dataset.

The above figure 3 represents the accuracy of possibilistic clustering algorithm, probabilistic clustering algorithm and proposed PPFCM clustering algorithm by varying data. By analysing the above figure 4, when the number of clusters increased, the accuracy of clustering process is decreased gradually for three clustering algorithms as used for evaluation process. Here, the maximum accuracy is attained by proposed PPFCM is 65.8% for number of data as 500 and the minimum accuracy is attained by probabilistic is 59.8% for number of data as 2500. The above figure 3 represents the accuracy of possibilistic clustering algorithm, probabilistic clustering algorithm and proposed PPFCM clustering algorithm. In figure 4, the maximum accuracy is attained by proposed PPFCM is 65.8% for number of data as 500 and the minimum accuracy is attained by probabilistic clustering is 59.9% for number of data as 2500. Similarly in figure 3 we obtain the maximum accuracy of 66% and minimum accuracy of 63%. Similarly, in figure 4 we obtain the maximum accuracy of 64%. From the above graph we clearly understand our proposed approach achieves the maximum accuracy compare to existing approaches.

V. CONCLUSION

The current proposal offers a divergent technique of effectively employing the privacy preserving clustering procedure with added emphasize on the incredible cost reduction for gigantic data processing. The amazing accomplishment of the novel PPFCM method is assessed, analyzed and contrasted with those of the possibilistic FCM and probability-clustering approaches for the yardstick datasets. Second-in-succession is the new-fangled Privacy Preserving Clustering procedure with considerable Cost reduction for the gargantuan Big Data Processing which emerges in flying colors in successfully addressing the most vital challenges such as the detection of clusters in multi-dimensional data sets, the multi-faceted hassles relating to secrecy and safety, and the drastic cut in the time complication and overheads of the total task. In this regard, the Hadoop plagued with an assortment of deficiencies, represents an extremely inferior accomplishment to appropriately assess the varied needs such as the Map Reducing, overall skills so as to enable the developer to function various versions of PLATFORA.

REFERENCES

[1] YogitaChawla and MansiBhonsle, (2012). A Study on Scheduling Methods in Cloud Computing‖, International Journal of Emerging Trends and Technology in Computer Science, vol. 1(3).

[2] Xun Xu, (2012). Cloud Computing to Cloud Manufacturing‖, Robotics and Computer-Integrated Manufacturing, vol. 28(1), pp. 75-86.

[3] Jadeja and Kirit Modi, (2012). Cloud Computing Concepts, Architecture and Challenges‖, International Conference on Computing, Electronics and Electrical Technologies.

[4] P. Garbacki and V. K. Naik,(2007). Efficient Resource virtualization and sharing strategies for heterogeneous Grid environments‖,/IEEE, pp. 40–49.

[5] Asmaa H.Rashid and Abd-Fatth Hegazy, (2010). Protect Privacy of Medical Informatics using K-Anonymization Model‖, IEEE Explore,

[6] Pekka Paakkonen and Daniel Pakkala, (2015). Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems‖, Big Data Research,

[7]Isaac Triguero, Daniel Peralta, Jaume Bacardit, Salvador Garcia and Francisco Herrera, (2015). MRPR: A Map Reduce Solution for Prototype Reduction in Big Data Classification‖, Neuro computing, vol. 150, pp. 331-345.

[8] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, (2014). Data Mining with Big Data‖, IEEE Transactions on Knowledge and Data Engineering vol. 26(1).

[9] Xingjian Li,(2015). An Algorithm for Mining Frequent Itemsets from Library Big Data", journal of software, vol. 9, no. 9.

[10] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen,(2014). Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier. IEEE International Conference on. IEEE.

[11] Shan Suthaharan, (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning, ACM SIGMETRICS Performance Evaluation Review, vol. 41(4), pp. 70-73.

[12] Chhaya S Dule,H.A. Girijamma and K.M Rajasekharaiah,(2014). Privacy Preservation Enriched MapReduce for Hadoop Based BigData Applications, American International Journal of Research in Science, Technology, Engineering & Mathematics,.

[13] Rongxing Lu, Hui Zhu; Ximeng Liu; Liu and J.K.Jun Shao, (2014). Toward efficient and privacy-preserving computing in big data era‖, IEEE Communication socity, vol.28 (4).

[14] Mehdi Sookhak, Abdullah Gani, Muhammad Khurram Khan and Rajkumar Buyya, (2015). Dynamic remote data auditing for securing big data storage in cloud computing‖, Information Sciences,

[15] Cem lyigun. Probabilistic Distance Clustering‖, Proquest, pages. 137.

[16] Nikhil R. Pal, Kuhu Pal, James M. Keller and James C. Bezdek, (2005). A Possibilistic Fuzzy c-Means Clustering Algorithm‖, IEEE transactions on fuzzy systems, vol. 13, (4), pp. 517-530,