International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 3, June 2018, pp. 1711~1719
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i3.pp1711-1719

A Novel Approach for Clustering Big Data based on MapReduce

Gourav Bathla[1], Himanshu Aggarwal[2], Rinkle Rani[3]
[1,2]Departement of Computer Engineering, Punjabi University Patiala, India
[3]Departement of Computer Science and Engineering, Thapar University Patiala, India

Article Info

Article history:

Received Dec 11, 2017
Revised Mar 20, 2018
Accepted Mar 28, 2018

Keyword:

Big data
Clustering
Kmeans
Kprototype
Mapreduce

ABSTRACT

Clustering is one of the most important applications of data mining. It has attracted attention of researchers in statistics and machine learning. It is used in many applications like information retrieval, image processing and social network analytics etc. It helps the user to understand the similarity and dissimilarity between objects. Cluster analysis makes the users understand complex and large data sets more clearly. There are different types of clustering algorithms analyzed by various researchers. Kmeans is the most popular partitioning based algorithm as it provides good results because of accurate calculation on numerical data. But Kmeans give good results for numerical data only. Big data is combination of numerical and categorical data. Kprototype algorithm is used to deal with numerical as well as categorical data. Kprototype combines the distance calculated from numeric and categorical data. With the growth of data due to social networking websites, business transactions, scientific calculation etc., there is vast collection of structured, semi-structured and unstructured data. So, there is need of optimization of Kprototype so that these varieties of data can be analyzed efficiently.In this work, Kprototype algorithm is implemented on MapReduce in this paper. Experiments have proved that Kprototype implemented on Mapreduce gives better performance gain on multiple nodes as compared to single node. CPU execution time and speedup are used as evaluation metrics for comparison.Intellegent splitter is proposed in this paper which splits mixed big data into numerical and categorical data. Comparison with traditional algorithms proves that proposed algorithm works better for large scale of data.

Corresponding Author:

Gourav Bathla,
Departement of Computer Engineering,
Punjabi University,
Patiala, India.
Email: gouravbathla@gmail.com

1. INTRODUCTION

Large scale of data are produced by social networking websites, data servers daily. Digital traces are also left by users on web space. This information can be very useful if it is extracted and analyzed properly. This large scale of data i.e. Big data cannot be processed with traditional computing. Management of this huge volume of data is very time consuming. There is need of mining algorithm which is distributed on different nodes. Clustering can be used on big data to combine large scale of data in compact format which will be highly informative [1]. Inter Cluster objects should have high dissimilarity and intra cluster objects should have high similarity. Traditional clustering algorithms are categorized based on their formation of clusters. These are divided into categories like partition based –KMeans, PAM, CLARA and FCM, hierarchical based- BIRCH, density based- DBSCAN, OPTICS, grid based- CLIQUE and model based-

EM,COBWEB. Big data clustering is analyzed by various researchers. Many techniques and frameworks are given in these studies for combining similar data objects in one cluster [2]. In this paper, we have highlighted parallel clustering algorithms – Kmeans and Kprototype on Mapreduce. Kmeans is the most important partition based algorithm. Kmeans clustering algorithm is widely used to combine objects with similarity based on distance metrics [3]. It uses distance measures like cosine distance, manhattan distance etc. Its main advantage is the simplicity. It calculates distance with k clusters and based on centroid value it starts making clusters. When the clustering is initialized, k is chosen before the clustering. These selections of k clusters have effect on the running time and efficiency of this algorithm. This paper proposes an algorithm which calculates k based on the information update while calculating clusters.

When Kmeans is migrated to big data, it does not perform well as compared to other clustering algorithms [4]. The main reason is that Kmeans is sequential and it computes the clusters in iterations. Kmeans works on numerical data with good accuracy. With categorical attributes, this algorithm can not calculate the centroid directly. Big data is combination of numerical and categorical data [2]. Kmeans can analyze numerical dataset with its proven accuracy. But this algorithm can not cluster categorical data. Kprototype algorithm is used to remove this drawback of Kmeans. Kprototype algorithm can handle numeric as well as categorical data effectively. We have also implemented Kprototype on Mapreduce so that it can handle large scale of data as well. As per our knowledge, very few research works have been carried out to focus on enhance the effectiveness of Kprototype algorithm.

Big data is combination of structured, unstructured and semi structured data. This research work covers big data characteristics like volume, velocity and variety. Volume is important characteristic of big data as this requires changes in storage architecture [1]. Velocity is another characteristic which should be managed by clustering algorithm as data flows in speed and response time should be accurate. Variety is third characteristic which is combination of structured, semi structured and unstructured data. This research work covers these characteristics with the use of big data technologies. In Kmeans algorithm with the use of hadoop platform big data can be processed effectively. (Key,Value) pairs of clustered data is processed with the use of Map [5]. Reduce combines the result of these pairs of different clusters. This approach reduces time complexity of clustering. When Kmeans is distributed on different clusters, running time for calculating clusters reduces significantly. In Section 2, several research works are described as literature survey. Clusering algorithms with Kmeans and Kprototypes detailed description is in Section 3. Proposed technique is presented in Section 4. Experimental analysis is elaborated in Section 5. Paper is concluded in Section 6 with future directions.

2. LITERATURE SURVEY

There are a lot of research works which are being carried out in clustering of big data. A Fahad et al [1] introduces a categorization framework for clustering algorithms. In this research work, authors have categorized different clustering algorithms based on designer perspective. Partition based, hierarchical based, density based, grid based and model based algorithm are explained in this paper. M. Haj Kacem et al [2] improves big data clustering by proposing Mapreduce based K-Prototypes (MR-KP). In this work, it is defined that big data is collection of numerical and categorical data. In various research works, few clustering methods can deal with mixed type.Proposed MR-KP can process numerical as well as categorical data. Experiments were conducted on many instances of chess dataset. It is proved in this research work that proposed K-Prototype shows good accuracy and scalability. X. Wu et al [3] demonstrates 10 algorithms which are most influential by IEEE International Conference on Data Mining (ICDM).C4.5, K-means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, and CART algorithms are explained in this research work. W. Zhao et al [4] have proposed a parallel Kmeans algorithm based on MapReduce. In this research work speedup, sizeup and scaleup is shown as better with the use of PKMeans algorithm. Mapreduce is used to implement machine learning and data mining algorithms in [6]. Hadoop and Mapreduce framework are explained in this paper. Kmeans, EM etc. data mining algorithms are implemented in parallel using Mapreduce in this paper. X Cui et al [7] proposed a novel processing model to remove the dependence on iterations. In Mapreduce there is limitation of restarting jobs. In this work, this is removed and resulted in high performance. In [8], authors have used dissimilarity measures between prototype of clusters and data objects. Four datasets are used for comparison of proposed method and traditional techniques. A. Ahmad et al [9] proposed a cost function based on co-occurrence of values. This cost function improves the cluster center accuracy for k-means clustering.

3. CLUSTERING ALGORITHMS

There are different clustering algorithms having specific applications in the field of data mining. In Figure 1, categories of clustering algorithms are explained with example.

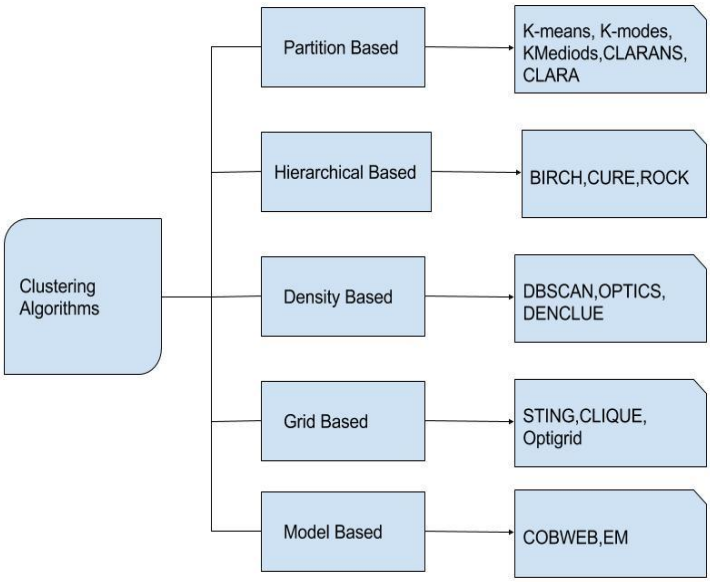

Figure 1. Categories of clustering algorithm

Partition Based: In these types of algorithms, data objects are divided into different partitions. These different partitions are clusters where data objects are having high intra-similarity. Kmeans which is a partition based algorithm which decides cluster membership by calculating centroid values.
Hierarchical Based: Complete data set is assumed as one cluster. This data set is divided into clusters in hierarchical manner (upto k number of clusters).
Density Based: Data objects are assigned into clusters based on density or connectivity.
Grid Based: In these types of algorithms, clusters are assigned to data objects based on statistical values.
Model Based: In these types of algorithms, clusters are assigned to data objects based on predefined model.

In our work, partition based algorithm is used. Kmeans which is partition based algorithm assigns data attributed to different clusters based on cost function. Euclidean distance is used for calculating distance function. The drawback of kmeans is that it can work only for numeric data. For categorical data, Kmeans can not work as there is no Euclidean space for this type of data. K-prototype is used for calculating cost function for categorical data, which is explained in next subsection.

3.1. Kmeans and kprototype

Kmeans is the most popular clustering method to check objects similarity [1]. The objects with in a cluster have high similarity and different clusters have high dissimilarity. It classify objects based on k value which are fixed before clustering, In various research work it is proved that results converge to local solution and not on global solution [4]. This algorithm calculates centroid value in iterative way. In first step, random objects are assigned to clusters. Then next step calculates new centroid value based on previous step. The value of k centroids change until last step when there is no change in value of centroid. This is final centroids value and objects assigned to clusters [6]**.**

In algorithm, step 2 takes maximum time. In this step, data is traversed for assigning to cluster. The running time can be reduced by using our technique. Only some i dimensions changes value after some iteration. There is no need to calculate K dimensions in every iteration. In this optimized approach, only i dimensions out of K dimensions are selected. These i dimensions are relevant. These dimensions are given the fixed priority. Only dimensions which are given priority are used in calculation of Euclidean distance from centroid. This reduces time complexity of deriving clusters from big data. Our technique is choosing k clusters and after some iteration selecting only objects which changes clusters. There is no need to compute the centroid value for the objects who remain in same clusters after some fixed iterations. It reduces the computation when there is a large scale of data- structured as well unstructured data.

Algorithm 1: Kmeans
Data : Data set N= {n_1……..n_n);K number of clusters
Result : Cluster Centroids : C_1…….C_K
begin
    Select K points in n- Euclidean space for intial centroids
    repeat
        Place data objects in these K points using distance measures.
        Recalculate centroids value by taking mean of data objects
    until there is no change in centroid positions.
end

In unstructured data there are numerical attributes as well as categorical attributes. In this proposed work, cost function will be defined as the combination of distance measures of numerical values as well as categorical values. Categorical values are not calculated as binary values or discrete values, rather it is calculated based on overall distribution or co-occurrence with other attributes. The similarity and dissimilarity of objects depend on how close their values are for all attributes. For numerical data it is easier to calculate the distance between objects based on Euclidean distance. It is difficult for categorical data to compute the closeness between objects. Binary distance measures is not appropriate for categorical data, it should give some value to categories of data [5].
Conversion of categorical data to numerical data:

a. The numerical distance can be applied after conversion of categorical attributed into numerical, attributes but it is very difficult.
b. Numerical data can be discretized to categorical data.

The distance between a data object and a cluster center is the summation of the distances between its numeric and categorical attribute values. For numeric attributes, we take the Euclidean distance between the object's attribute value and the mean value of the center. For categorical attributes, all values have a proportional presence in the definition of cluster center.

It is presented in many studies that Kmeans can process numerical data only. Kprototype is able to remove this limitation [2]. Kprototype is proposed in [10] to remove the limitation of Kmeans algorithm. Kprototype is combination of Kmeans and Kmodes algorithms. Kprototype algorithm can handle numerical and categorical data [11]. Euclidean distance is used for calculating similarity for numerical attributes. Hamming distance is used for calculating similarity for categorical attributes. Split the data D into numerical and categorical value.

$$d(x_i, Y_l) = \sum_{r=1}^{m_r}(x_{ir} - y_{lr})^2 + \sum_{c=1}^{m_c}\partial(x_{ic}, y_{lc}) \quad (1)$$

In this equation, distance between attribute value x_{ir} and numeric center y_{lr} is calculated. Numerical values distance is calculated by using means of data objects allotted to a cluster. Then these clusters are updated based on iterations. Also, distance between attribute value x_{ic} and categorical center y_{lc} is calculated. Categorical values distance is calculated by using most frequently occurring value as cluster center. Different distance measures can be used for numerical as well as categorical data.

Numerical and categorical data is separated as shown in Algorithm 2. Initial values are selected and then similarity is calculated using Equation (1). These iterations are carried out until there is no change in clusters values i.e Old center C is equal to updated center Cu.

Kmeans and Kprototype algorithm works on small scale of data with good accuracy. But when it is deployed on big data, it takes unrealistic duration to process this large scale of data. We have deployed Kprototype on Mapreduce in this paper. In this work, intelligent framework is also proposed for big data clustering. Different varieties of mixed data are separated into numerical and categorical data. Then these different data objects are assigned clusters on different Map and Reduce phase. Detailed proposed framework is explained in next section.

Algorithm 2: Kprototype
Data: Dataset D = {x1 …….xn) and cluster yn for numerical and yc for categorical data
Result: Cluster Center C for numerical and categorical data

```
begin
    Select K points in n- Euclidean space for
    intial centroids for numerical data and K points for hamming distance for categorical data
    repeat
        Compute distance using cost
        function proposed.
        Update Cluster values C to Cu
    until   C=Cu
end
```

4. PROPOSED TECHNIQUE

Big data is collection of numerical and categorical data. Traditional Kmeans can not work on these types of data efficiently. It works on numerical data with proven accuracy. It calculates centroid value of objects for clustering. Distance is calculated between n-dimensional vectors using Euclidean distance. Then center is calculated for different clusters c1, c2 …ck. and average distance is measured using sample points.Cosine distance, Euclidean distance and Pearson correlation are used for calculation of similarity [13]. These distance measures works for numerical data with accuracy because numerical data have origin in Cartesian coordinate's value. Mapreduce [12] can process data in parallel by the use of map and reduce phase. Kmeans is deployed on Mapreduce with parallel calculation of clusters for processing large scale of data [4], [14], [15]. Similarity between data objects and clusters are different for every object. So, distance can be calculated in parallel by the use of map and distance from each nodes is combined to form global result in reduce.

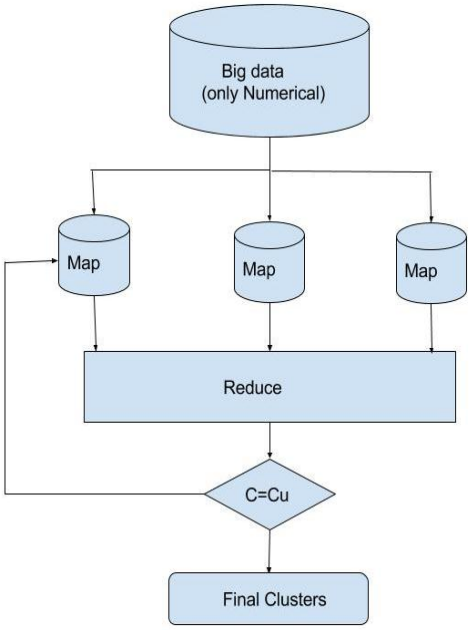

Figure 2. Clustering of numerical data

As explained in Figure 2, only numerical part of big data can be assigned in one cluster using Kmeans on Mapreduce. If Clusters values do no change then clusters are finalized, otherwise map and reduce phase is repeated in next iteration. In this work, the limitation of Kmeans using numerical data only is removed by using proposed framework.

Categorical data can be converted into numerical form as proved by various research works. But it results in a lot of time consumption and loss of information. In this proposed framework, intelligent algorithm is used which check type of data in first phase. Then dataset is deployed on map only after deciding type of data. Splitter proposed in this work separate mixed dataset and then assign it to correct split as shown in Figure 3. Kprototype algorithm can calculate similarity between objects for big data by using euclidean distance and hamming distance. This algorithm removes the drawback of kmeans algorithm which

is working only on numerical data. This algorithm produces very interesting results on mixed data. In our proposed framework, this algorithm is deployed on MapReduce model to manage large scale data. In this algorithm, calculation of object with cluster center is independent of another object calculation of distance with relevant cluser center. So, Kprototype algorithm fits well to be implemented in parallel on Mapreduce. In Figure 4, it is clearly explained that mixed data is distributed in numerical and categorical data. On numerical part, Euclidean distance is used for calculating distance with center. On categorical part, hamming distance measure is used. Then results of both numerical and categorical data are combined to form cluster centers.

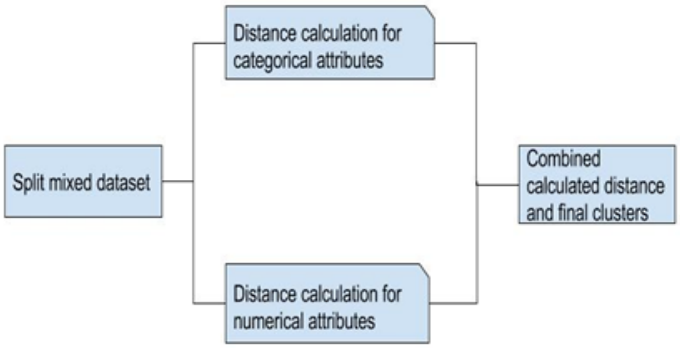

Figure 3. Splitter for distributing numerical and categorical data on clusters

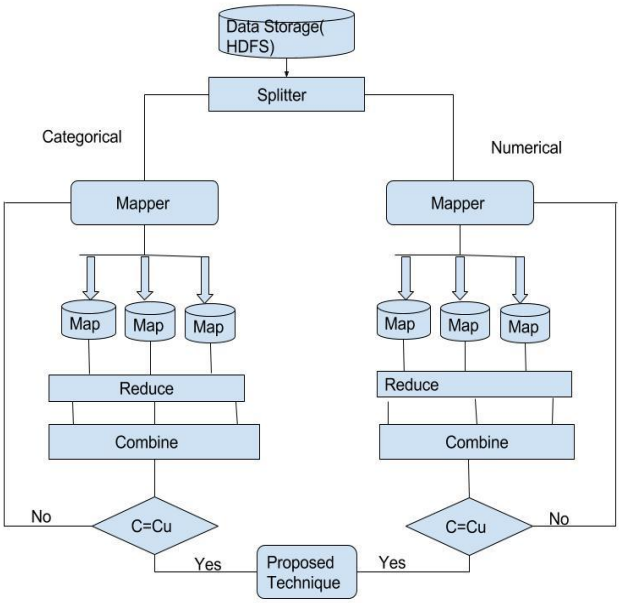

Figure 4. Proposed framework for clustering mixed dataset on Mapreduce

In map phase object distance with cluster center is calculated and in reduce phase results are combined. Existing approaches use map and reduce for numerical data and then after getting input from it, distance is calculated for categorical data. Our proposed approach runs parallel for parallel computation. Numerical and Catrgorical data clusters are calculated in parallel and in these processes, cluster centers are calculated individually using map and reduce also. Splitter checks the mixed dataset and sends data to cluster set which is appropriate. This reduces processing time as initially mapper has correct data to process. On mapper, distance calculation is separate for categorical and numerical data. When clusters are finalized after using Equation (1), proposed technique combines the cluster.

5. RESULTS AND ANALYSIS

A key motivation for this experiment is to prove that Kprototype works better with the use of hadoop and Mapreduce. This section proves that on big data, proposed work gives accurate results for clustering. The important parameters for checking the performance are scaleup, speedup and CPU utilization. Experiments prove that proposed algorithm satisfies these parameters with cluster accuracy. When this proposed algorithm is deployed on multiple nodes then performance improves in terms of response time. This scaleup is improved by comparing K_1 with K_m. We have used Chess dataset which has combination of numerical as well as categorical attributes. This dataset is combination of chess positions as shown in Figure 5.

Dataset Statistics	
Number of data objects	28056
Numerical Attributes	6
Categorical Attributes	10

Figure 5. Chess dataset statistics

In our experiment, hadoop 1.2.1 using VMWare is used. Results show that cluster accuracy is very good when our proposed technique is implemented on Mapreduce. Using hadoop platform, the input data is processed on Map. Then using HDFS, Kprototype works on semistructured and unstructured data. Part-00000 file contains the final clusters from big data.

Mapreduce process this library as follows [16], [17]:

Input - This library is divided into several data blocks for working on map function. abstract classes are defined at this step of processing.
Key-Value pair- In this step <key,value> pair is defined for each key-value pairs.
Shuffle- In this step all input of <key,value> pairs are sorted.
Reduce- In reduce step, <key,{list}> pairs are traversed to <key,value>.
Output - This step combines the output of different clusters and combines final output.

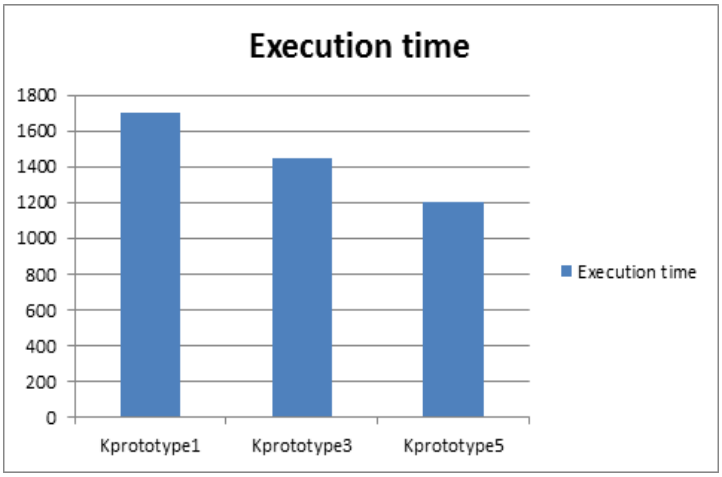

Figure 6. Comparison of Kprototype on 1, 3 and 5 clusters

In Figure 6, it is elaborated that Kprototype is deployed on single node and multiple nodes to analyse the difference in CPU time. It is clear that when it is deployed on multiple nodes using intelligent splitter of our proposed approach, CPU time reduce significantly.
Speedup is also used to prove better results from our proposed approach.

$$\text{Speedup} = \frac{T1}{Tm} \quad (2)$$

where T_1 is speed on single node and T_m is speed on m nodes.

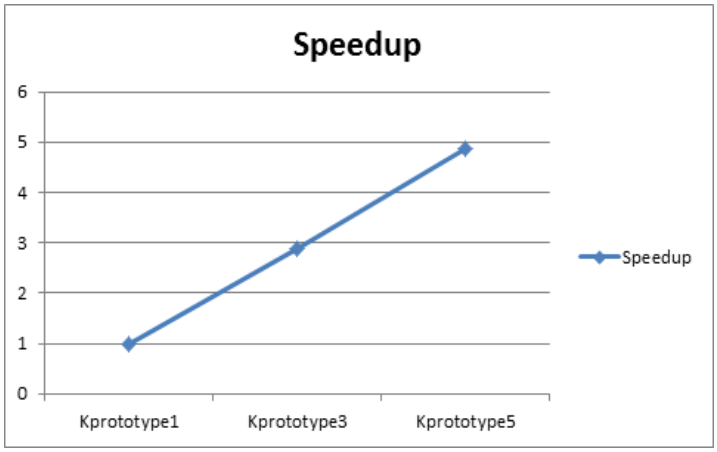

Figure 7. Kprototype speedup on multiple nodes

From Figure 7, it is clear that when this proposed technique is deployed on multiple nodes, speedup is gained with the increase of number of nodes. Experiments prove that linear speedup is not gained as some CPU time is consumed in data transfer and final result after merging of data from different nodes.

6. CONCLUSION

Clustering is used in our work to process big data efficiently. Different types of clustering algorithms are explained in this paper. Kmeans which is partitioning based algorithm is elaborated in this paper. Kmeans can work for numerical data but it can not work well for categorical data. Big data is combination of different varities of data like numerical and categorical. Kprototype is used in our work which can analyse numerical as well as categorical data. Kprototype is deployed on big data by using Mapreduce. CPU execution time and speedup are improved significantly which is also proved in experiment section. We have proposed intelligent splitter which checks the variety of data, splits it into numerical and categorical and deploy the data to its correct map and reduce. Using hadoop and Mapreduce, big data velocity, variety - structured, unstructured and semi structured, and volume – huge quantity of data, are managed and processed very effectively. In future work, more distance measures can be used that can be compared with this proposed technique of categorical data clustering.

REFERENCES

[1] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A.Y. Zomaya, S. Foufou and A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", *IEEE Transactions on Emerging topics in Computing*, vol. 2 no. 3, pp. 267-279, 2014.

[2] M. HajKacem, C. Ben N' cir and N. Essoussi, "*MapReduce-based K-Prototypes Clustering Method for Big Data*", In Proceedings of International Conference on DSAA, IEEE, pp. 1-7, 2015.

[3] X. Wu, V. Kumar, J. Ross Quinian, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng. B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D.J. Hand and D. Steinberg, "Top 10 algorithms in data mining", *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, 2007.

[4] W. Zhao, H. Ma and Q. He, "*Parallel K-Means Clustering Based on MapReduce*", in Cloud Com LNCS 5931, pp. 674-679, 2009.

[5] J. Heer and S. Kandel, "Interactive analysis of big data", *XRDS ACM*, vol. 19 no. 1, pp. 50-54, 2012.

[6] K. Shim, "MapReduce Algorithms for Big Data Analysis", *Databases in Networked Information Systems, LNCS*, vol. 7813, pp. 44-48, 2013.

[7] X. Cui, P. Zhu, X. Yang, K. Li and C. Ji, "Optimized big data K-means clustering using MapReduce", *Journal of Supercomputing Springer*, vol. 70 no. 3, pp. 1249-1259, 2014.

[8] M. HajKacem, C. N'cir and N. Essoussi, "Parallel K-Prototypes for Clustering Big Data", *ICCCI LNCS 9330*, pp. 628-637, 2015.

[9] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data", *Journal of data and knowledge engineering ACM*, vol. 63 no. 2, pp. 503-527, 2007.

[10] Z Huang, "Clustering large datasets with mixed numeric and categorical values", in *Proceedings of conference on Knowledge discovery and data mining*, pp. 21-34, 1997.
[11] J. Ji, T. Bai, C. Zhou, C. Ma and Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data", *Neurocomputing Elsevier*, vol. 120, pp. 590-596, 2013.
[12] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", *Communications of the ACM*, vol. 51, no.1, pp. 107-113, 2008.
[13] R. Xu and D. Wunsch, "Survey of Clustering Algorithm", *IEEE Transactions on Neural Networks*, vol. 16, no. 3, 2015.
[14] Z. Huang, "A fast clustering algorithm to cluster very large categorical datasets in data mining", *Research issues on data mining knowledge discovery*, pp. 1-8, 1998.
[15] K.R. Nirmal and K.V.V Satyanarayana, "Issues of K Means clusteing while migrating to Map Reduce paradigm with Big Data: A Survey", *IJECE*, vol. 6 no. 6, pp. 3047-3051, 2016.
[16] X. Yan, Z. Wang, D. Zeng, C. Hu and H. Yao, "Design and analysis of parallel MapReduce based knn-join algorithm for Big Data Classification", *IJEECS*, vol. 12 no. 11 , pp. 7927-7934, 2014.
[17] S.A.Thanekar, K. Subrahmanyam and A.B Bagwan, "A study on MapReduce: Challenges and Trends", *IJEECS*, vol. 4 no. 1, pp. 176-183, 2016.

BIOGRAPHIES OF AUTHORS

Mr. Gourav Bathla is PhD student in Department of Computer Engineering, Punjabi University, Patiala. He has 10 years of teaching and research experience.He has published papers in international conferences and journals. His area of interest is Big Data, Data mining, Programming languages. He is member of IEEE cloud computing, CSI and ISTE.

Dr. Himanshu Aggarwal, Ph.D., is currently serving as Professor in Department of Computer Engineering at Punjabi University, Patiala. He has more than 22 years of teaching experience and served academic institutions such as Thapar Institute of Engineering & Technology, Patiala, Guru Nanak Dev Engineering College, Ludhiana and Technical Teacher's Training Institute, Chandigarh. He is an active researcher who has supervised more than 30 M.Tech. Dissertations and contributed 80 articles in various Research Journals. He is guiding PhD to 8 scholars and Five has completed his PhD. He is on the Editorial Board of 9 Journals and Review Boards of 5 Journals of repute. His areas of interest are Software Engineering, Computer Networks, Information Systems, ERP and Parallel Computing. Himanshu Aggarwal can be contacted at: himanshu.pup@gmail.com

Dr. Rinkle Rani is working as Assistant Professor in Computer Science and Engineering Department, Thapar University, Patiala since 2000. She has done her Post graduation from BITS, Pilani and Ph.D. from Punjabi University, Patiala in the area of Computer Networks. She has more than 18 years of teaching experience. She has supervised 34 M.Tech. Dissertations and contributed 50 articles in Conferences and 41 papers in Research Journals. Her areas of interest are Computer Networks and Big data mining and Processing. She is member of professional bodies: ACM, IEEE, ISTE and CSI. She may be contacted at raggarwal@thapar.edu.