International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Parallel Computation Performingkernel-Based Clustering Algorithm using Particle Swarm Optimization for the Big Data Analytics

E. Laxmi Lydia, B. PRASAD, Gogineni HimaBindu, K.Shankar, K.Vijaya Kumar

Abstract: ***Digital data has been accelerating day by day with a bulk of dimensions. Analysis of such an immense quantity of data popularly termed as big data, which requires tremendous data analysis scalable techniques. Clustering is an appropriate tool for data analysis to observe hidden similar groups inside the data. Clustering distinct datasets involve both Linear Separable and Non-Linear Separable clustering algorithms by defining and measuring their inter-point similarities as well as non-linear similarity measures. Problem Statement: Yet there are many productive clustering algorithms to cluster linearly; they do not maintain quality clusters.Kernel-based algorithms make use of non-linear similarity measures to define similarity while forming clusters specifically with arbitrary shapes and frequencies.***
Existing System:Current Kernel-based clustering algorithms have few restraints concerning complexity, memory, and performance. Time and Memory will increase equally when the size of the dataset increase. It is challenging to elect kernel similarity function for different datasets. We have classical random sampling and low-rank matrix approximation linear clustering algorithms with high cluster quality and low memory essentials. Proposed work: in our research, we have introduced a parallel computation performing Kernel-based clustering algorithm using Particle Swarm Optimization approach. This methodology can cluster large datasets having maximum dimensional values accurately and overcomes the issues of high dimensional datasets.

I. INTRODUCTION PREFACE

Advancement in data creation, data acquisition, and data storage applications. This lead to an explosion in digital data. Corporations like IDC and ECM has anticipated that 44 ZB of digital data generated by the year 2020(Fig 1).Real-timeprocessing of high-dimensional digital data both structured and unstructured as a part of big data.

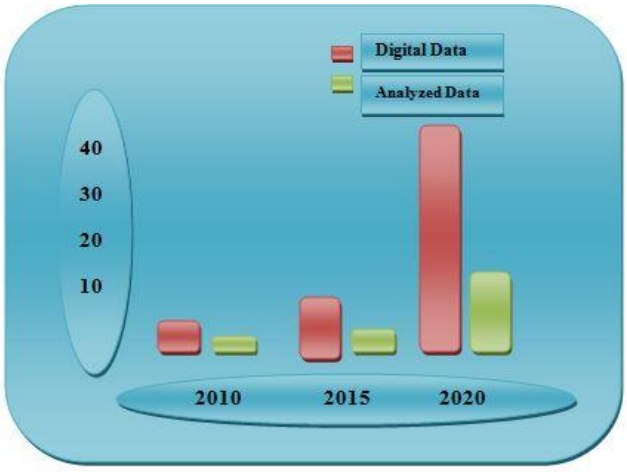

Fig.1 Data creation from past years to future years

Big Data Analytics margin new applications for different sectors, recent beneficiary services were provided to medical healthcare systems. Any data analysis is normally classified as finding data and process the data effectively without any noise. It identifies various patterns and designs the data accordingly. Most of the digital data deal with Statistical methods. It follows the internal process of pattern recognition like the representation of data, learning, and interpretation of data.

Set of features represent data objects among the datasets. These features from the data sets can be mathematical numbers, unambiguous. A text document words can be represented as x_p. Image pixels can be represented using intensity values. Representation of data plays a major domain to proper and accurate result analysis. With respect to the new applications, deep learning has employed effective approaches for automatic representation of data.

Hidden structured data in unsupervised does not need labeling, therefore it is easy to avoid searching of data. Tasks performed by unsupervised learning are estimating density, dimensionality reduction, finding features and extracting them, finally cluster features.

A. Unsupervised Classification

Prominent approaches for unsupervised classification is Clustering. The process of clustering is easier and reliable. Clustering has found its place in popular applications like web search, retrieval of information, market analysis etc. Clustering algorithms are developed based on the datasets.

Hierarchical clusteringalgorithm, creates a hierarchy of clusters i.e, groups to the subdivisions.

Revised Manuscript Received on 30 July 2019.
* Correspondence Author
E. Laxmi Lydia, Associate Professor, Department of Computer Science Engineering, Vignan's Institute of Information Technology,Andhra Pradesh, India.
Dr. B. PRASAD, Professor, Department of Computer Science Engineering, Vignan's Institute of Information Technology,Andhra Pradesh, India.
Gogineni HimaBindu, Asst. Prof., MCA, Vignan's Institute of Information Technology(Autonomous), Visakhapatnam, India.
K.Shankar, Assistant Professor,School of Computing,KalasalingamUniversity,Krishnankoil, – 626126, Tamil Nadu, India.
K.Vijaya Kumar, Associate Professor, Department of Computer Science Engineering, Vignan's Institute of Information Technology for Women, Andhra Pradesh, India.

- *Agglomerative hierarchical clustering algorithm* initiates clusters with each point and merges them to similar points.
- *Divisive hierarchical clustering algorithm* follows top-down procedurestarting from thebase point finds all the data points and recursively share the data points to accurate clusters

Partitional clustering algorithm, instantly classify the data into "C" clusters encompass centroid based mechanism.

- *Centroid-based algorithms* are implemented using K-means clustering and K-medoids clustering algorithms.
- *Model-based algorithms* are implemented using Mixture models and Latent Dirichlet Allocation,
- *Graph-based theoretic algorithms* are implemented using Minimum Spanning trees, Normalized-cut, Spectral Clustering
- *Density and Grid-based algorithms* are implemented using DBSCAN, OPTICS, CLIQUE.

For statistical data, clustering is characterized into parametric and non-parametric. For parametric clustering, it tries to identify the component densities by all the densities p(x), algorithms like centroid-based and model-based come into parametric clustering. For non-parametric, it evaluates all the high-densityneighborhood regionsand creates cluster trees.

Complications during data clustering

The process of unsupervised clustering has to undergo some challenges in generating clusters. They are

→ Appropriate Representation of Data
→ Assigning the number of clusters
→ Choosing suitable Clustering algorithm
→ Accurate calculation of similarity measures
→ Specifying parameters for clusters like quality, scalability, validity measures.

B. Kernel-Based Clustering

Large-scale clustering algorithms have found scalability issue. They predict that in the input space of clusters are linearly separable and interpret the inter-point similarity measures using distance calculations. The two main deficiencies are

→ Clusters that cannot be separated by hyperplane input space and also cannot be clustered alone by linera clustering when compared to other clustering algorithms.
→ Arbitrarily shaped clusters are determined using Non-linear similarities.
→ Kernel-based clustering algorithms accomplish high-quality clusters.

C. Necessity of Clustering in Big Data

Acknowledging the issues in Big Data, aim to have high scalability in picking appropriate clustering algorithms.Various clustering algorithms focusing on large datasets try to compress the data in the preprocessing phase ahead in creating clusters. Following are the major clustering algorithm's runtime and memory complexities.

Clustering Algorithms	Running time	Memory complexity
Hierarchical clustering algorithm	$O(n^2 d + n^2 \log(n))$	$O(n^2)$
Partitional Clustering algorithms	$O(nlog(n))$	$O(n^3)$

II. BACKGROUND

Pradeep Singh et, al.[1] have proposed a machine learning technique for clustering techniques concentrating on the estimation of densities in spatial datasets. They have worked and processed on the removal of noise in the benchmark datasets that has worst run-time complexity. M.Parimala et, al[18] in their research they have also concentrated on the same density based algorithms working on different clustering algorithms by scrutinizing their essential parameters in developing quality clusters.

Jian-Sheng Wu et al, [2] have conducted they study on large-scale datasets for nonlinear clustering. They preferred already existing kernels for mapping data to high dimensional space directly and analyzed that kernel matrix consuming lot of memory and initiated Euler clustering. Its main target is to get rid of existing kernels and never to rely on random sampling i.e, kernel matrix. A positive Euler kernel, Euler spectral clustering, kernel k-means are introduced for fast computation.

Radha Chitta et al, [3][8][19] worked on sparse kernel k-means clustering algorithms. Combinations of Sampling and Sparsity have achieved in reaching minimum run time and memory complexity i.e, $O(NCd)$. Top Eigen vectors are converted using k-means algorithm. Distinct high dimensional text data and pixel data become faster with this kernel-based clustering.

Charu C. Aggarwal et al, [4][20] focused on different streaming algorithms for clustering using micro-clustering framework, Pyramidal Time Frame, Online Clustering in CluStream, Density-based stream clustering, Grid-based stream clustering, Probabilistic streaming algorithms, clustering Discrete and Categorical streams, text stream clustering. They have concentrated on the natural applicability of summarizing the input dataset data in heterogeneous data streams. In their research, they have found that maximum complexity over graphs and dynamic affiliations on various data streams.

Bing Liu et al, [10][7][16] proposed an unsupervised non-parametric kernel learning algorithm by the rise of the problem in kernel-based algorithms. They emerged into the subject that significantly leads to effective maximum margin clustering to enrich its performance. Predicting data on both supervised and unsupervised using multiple kernel learning algorithms.

Dimitris Achlioptas [17] have popularized atechnique related to the lower rank matrices by considering the most frequent methods while estimating the approximations independently. They performed computational calculations on sampling and quantization. They minimized and discarded the lower rank matrix and achieved a linear structure in faster approximations.

Wen-Yen Chen et, al.[15][14] recommended spectral clustering algorithms parallelly in distributed systems in achieving scalability problems on large datasets. They have implemented the strategy by performing nearest neighbours with parallelization.

Itslike parallel algorithm to handle large dataset problems.

III. METHODOLOGY

Kernel-based clustering has lead data to the simplest path in classifying unsupervised data. This has achieved computative geometrical parameters in high-density zones. To achieve quality clustering, the following conditions need to be satisfied.

→ *Data must be Scalable*
→ *Dealing with various attributes of the data*
→ *Dealing with all indifferent noise and outliers.*
→ *Identify exact cluster data*
→ *Dealing with high dimension data.*

Kernel k-means clustering algorithm

To cluster the data, Clustering algorithms are classified into two divisions mainly,unsupervised linear clustering algorithms and unsupervised non-linear clustering algorithms. Kernel k-means clustering algorithm is similar to k-means but kernel-based method calculates Euclidean distance. The analysis is performed based on

1. The computational complexity that performs uniform sampling approach evaluated in O(n)
2. Approximation error calculated between the kernel k-means and approximate kernel k-means

Algorithm Procedure:

Step1: Suppose if the dataset contains n data points, assume that D= $\{a_1, a_2, a_3, \ldots . a_n\}$

Step2: Initialize clusters randomly with its center "c".

Step3: Calculate the distance among the data points and the defined cluster center by using$D(\{\pi_c\}_{c=1}^{k}) = \sum_{c=1}^{k}\sum_{a_i \in \pi_c} ||\emptyset(\ a_i) - m_c||^2$

Step4: check the minimum distance of every data point to the cluster center and sign the point to the nearest cluster.

Step5: continue the process until all the data points are assigned to the clusters.

Kernel k-means clustering is a more advantageous for real-time data sets and it can easily find the non-linear structures.

Procedure for Approximate kernel K-means

Step1: Let the given dataset m with data points D= $\{a_1, a_2, a_3, \ldots . a_n\}$ and kernel function as k (.) and clusters as $c_k(.)$ Where $c_k(.) = \sum_{i=1}^{n} \hat{U}_{k,i}$ k (a_i), k∈ $[c]$

Step2: Compute matrices for kernel function

Step3: Randomly initiate cluster membership matrix U, with optimal cluster centers.

Step4: Compute normalized membership matrix.

Step5: continue the process until the cluster membership matrix remains same.

PSO (Particle Swarm Optimization)

Step1: Consider high dimensional dataset containing N data points (particles).

Step2: Estimate the density function

Step3: check for the number of clusters and initial cluster centers and go to step2

Step4: Assign velocity and position of the particles with Radom initializations.

Step5: Find fitness functions and set values for p_{best} *and* g_{best}.

Step6: Find inertia weight factor (w) value.

Step7: Update the procedure values of velocity and position according to the boundary restrictions.

If the position is less than or equal to the upbound and if the position is greater than or equal to the position then position is equal to the positionElse the position is calculated as position - velocity.

Step8: Now group the particles with the nearest cluster center and calculate the cluster center

Step9: repeat the process until all the particles get into clusters.

IV. CONCLUSION

The main objective of this paper is to produce accurate clusters for the large datasets by using clustering algorithms. As per the inspection and analysis of previous implementations kernel-based clustering algorithms were widely accomplished for high-quality clusters. The framework of this paper address the scalability issue in large-scale datasets using a proper selectionof *approximate kernel-based clustering algorithm*and *Particle Swarm Optimization* techniques. This has illustrated the kernel based clustering analytically and experimentally to manage the trade-off within scalability and capability with respect to high dimensions. Therefore, we conclude that this way has reduced the runtime and memory complexity by sampling the data points and limiting the memory necessities by maintaining quality clusters.

REFERENCES

1. Pradeep Singh, Prateek A. Meshram, Survey of density-based clustering algorithms and its variants, 2017 International Conference on Inventive Computing and Informatics (ICICI), May 2018.
2. Jian- Sheng Wu, Wei-Shi Zheng, Jian- Huang Lai, Ching Y. Suen, Euler Clustering on Large-scale Dataset, IEEE Transactions on Big Data, 2017.
3. R. Chitta, A. K. Jain and R.Jin, Sparse kernel clustering of massive high-dimensional datasets with large number of clusters. Technical report MSU-CSE-15-10, Department of Computer Science, Michigan State University, 2015.
4. C.C. Aggarwal, A survey of stream clustering algorithms. In Data Clustering: Algorithms AI and Applications, pages 231-258,2013.
5. Erte Pan, Husheng Li, Lingyang Song and Zhu Han, Kernel-based non-parametric clustering for load profiling of big smart meter data, IEEE Wireless Communications and Networking Conference (WCNC), 10.1109/WCNC.2015.7127817, 2015.
6. Arshiya Mubeen, N. D Abhinav, C.V. Shanmuka Swamy, K.H. Swetha, H. Rakesh, Reducing the risk of customer migration by using Bigdata clustering algorithm, 2nd IEEE International Conference on recent Trends in Electronics, Information & Communication Technology (RTEICT), Jan 2018.
7. R. Chitta, R.Jin, T.C. Hevens, and A.K. Jain. Scalable kernel clustering: Approximate kernel k-means. Arxiv preprint arXiv:1402.3849,2014.
8. R. Chitta, R. Jin, T.C. Havens, and A.K.Jain Approximate kenel k-means: Solution to large-scale kernel clustering. In proceedings of the International Conference on knowledge Discovery and Data mining, pages 895-903, 2011.
9. R. Abbasifard, B.Ghahremani, and H.Naderi. A survey on nearest neighbor search methods. International Journal of Computer Applications, 95(25):39-52, 2014.
10. B. Liu, S.X. Xia, and Y. Zhou, Unsupervised non-parametric kernel learning algorithm knowledge- based systems, 44:1-9,2013.

11. Q.Le, T.Sarlos, and A.Smola, Fastfood- Approximating kernel expansions in loglinera time. In proceedings of the International Coference on Machine Learning, Pages 16-21,2013.
12. R. Hamid, Y. Xiao, A.Gittens, and D. DeCoste. Compact random feature maps.arXiv preprint arXiv:1312.4626,2013.
13. Mathias Rossignol, Mathieu Lagrange, Cont A (2018) Efficient Similarity-based clustering by optimal object to cluster reallocation. PLoS ONE 13(6):e0197450.
14. A.Gittens, P. Kambadur, and C. Boutsidis. Approximate spectral clustering via randomized sketching. arXiv preprint arXiv:1311.2854, 2013.
15. W.Chen, Y. Song, H. Bai, C. Lin, and E.Y. Chang. Parallel spectral clustering in distributed systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(3):568-586, 2011.
16. C. Cortes, M. Mohri, and A.Talwalkar. On the impact of kernel approximation on learning accuracy. Journal of Machine Learning Research, 9:113-120,2010.
17. D. Achiloptas and F. McSherry. Fast computation of low-rank matrix approximations. Journal of the ACM,54(2), 2007.
18. M. Parimala, Daphne Lopez, N.C.Senthil Kumar, A survey on Density Based clustering Algorithms for mining large spatial Databases, International Journal of Advance Science and Technology, vol.31, june 2011.
19. Radha Chitta. Kernel- Based Clustering of Big Data, Michigan State University, 2015.
20. Maryam Mousavi, Azuraliza Abu Bakar, Mahammadmahdi Vakilian, Data Stream clustering Algorithms: review, Int. J. Advance soft Compu. Appl, Vol.7, No.3, November 2015 ISSN 2074-8523.