ISSN: 2321-7782 (Online)
Volume 2, Issue 2, February 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Paper / Case Study
Available online at: www.ijarcsms.com

A Review paper on Parallel Power Iteration Clustering for Big Data

Darji Ankit[1]
Research Scholar
CSE Department
Parul Institute of Technology
Vadodara – India

Dinesh Vaghela[2]
Assistant Professor
CSE Department
Parul Institute of Technology
Vadodara – India

Abstract: ***In todays Distributed Data Mining is most popular topic in research area because there are so many issues and there are also a solutions for that but still they are not as per satisfaction, there are some problems already there, mainly we are focusing in this papers that about reducing communication cost and computational cost of our data transfer. Here this paper contains some methods for clustering and from those traditional methods are still popularly used, they generally lack robustness and suffer from the so-called ''curse of dimensionality''. Here Power Iteration Clustering is newly developed method. It present a simple and scalable graph clustering method called power iteration clustering (PIC).But still it is not scalable for large dataset so here, another one is parallel pic which is parallel approach of PIC in which is focus on matrix vector multiplication and based on that clusters are arranged and then measures communication cost and computational time.***

Keywords: Clustering, Data Mining, Big Data. Parallel Computing, Spectral clustering, normalized cuts, nearest neighbors, Nystrom approximation.

I. INTRODUCTION

Clustering is a process of organizing data into groups within which the elements are similar in some way. As an unsupervised learning technique mainly for discovering natural groups or underlying structure of a given dataset, clustering has been an active research subject in many fields including statistical analysis, image analysis, pattern recognition, machine learning, and data-mining. Clustering applications also span wide range of domains, including text-mining , social network analysis , bioinformatics, market research, and scientific and engineering analysis, to name a few.. Traditional clustering methods include hierarchical methods (e.g., single link) and partition methods (e.g., k-means). Although those traditional methods are still popularly used, they generally lack robustness and suffer from the so-called ''curse of dimensionality''. Most of those methods are also computational expensive especially when applied to large fundus images helps the ophthalmologists to decide on the severity of the DR and advise the required treatment to the patients.

The paper is organized as follows: Section II gives a review of DBSCAN Algorithm and Spectral Clustering Algorithms that shown to be more effective for finding Clustering.

In Section III, Contains introduction of newly developed power iteration clustering, Section IV, Contains another research on power iteration cluster which is then converted into parallel power iteration clustering. Section V, Define the conclusion of this paper.

II. DBSCAN AND SPECTRAL CLUSTERING

In 1996, **M.Ester, H.P., J.Sander...DBSCAN** [1], it presents the new clustering algorithm DBSCAN which is based on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. DBSCAN algorithm requires only one input parameter and DBSCAN supports the user in determining an appropriate value for it. Here this paper shows an experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data and real data of the SEQUOIA 2000 bench-mark. The results of their experiments demonstrate that, (1) DBSCAN outperforms CLARANS by a factor of more than 100 in terms of efficiency, and that (2) DBSCAN is significantly more effective in discovering clusters of arbitrary shape than the well-known algorithm CLAR-ANS. Future research will have to consider the following issues. First, they have only considered point objects. Spatial data-bases, however, may also contain extended objects such as polygons. They have to develop a definition of the density in an Eps-neighborhood in polygon databases for generalizing DBSCAN. Second, applications of DBSCAN to high dimensional feature spaces should be investigated.

In 2011, **W.Y. Chen, Y. Song, H. Bai** [4], they presents Spectral Clustering algorithm. Spectral clustering algorithms have been shown to be more effective in finding clusters than some traditional algorithms which are used for finding clusters, such as k-means. However, spectral clustering suffers from a scalability problem in both memory use and computational time when the size of a data set is large. To perform clustering on large data sets, they investigate two representative ways of approximating the dense similarity matrix. They compare one approach by sparsifying the matrix with another by the Nystrom method. They then pick the strategy of sparsifying the matrix via retaining nearest neighbors and investigate its parallelization. They parallelize both memory use and computation on distributed computers.

In 2001, M. Meila, J. Shi [2]**,** they present a new view of Clustering and segmentation by pairwise similarities. They interpret the similarities as edge owns in a Markov random walk and study the eigenvalues and eigenvectors of the walk's transition matrix. This view shows that spectral methods for clustering and segmentation have a probabilistic foundation. They prove that the Normalized Cut method arises naturally from our framework and they provide a complete characterization of the cases when the Normalized Cut algorithm is exact. Then they discuss other spectral segmentation and clustering methods showing that several of them are essentially the same as NCut.

III. POWER ITERATION CLUSTERING

In **2010, F. Lin Present** [3], It represents PIC finds a very low-dimensional embedding of a dataset using truncated power iteration on a normalized pair-wise similarity matrix of the data. PIC is very fast on large datasets which are running over 1,000 times faster than an NCut implementation based on the state-of-the-art IRAM eigenvector computation technique. In spectral clustering the embedding is formed by the bottom eigenvectors of the Laplacian of a similarity matrix.In PIC the embedding is an approximation to an eigenvalue-weighted linear combination of al l the eigenvectors of a normalized similarity matrix. The main focus of PIC is its simplicity and scalability. They demonstrate that a basic implementation of this method is able to partition a network dataset of 100 million edges within a few seconds on a single machine, without sampling, grouping, or other preprocessing of the data.

Table 1. Runtime comparison (in milliseconds) of PIC and Spectral clustering algorithms on several real datasets. [3]

Datasets	Size	NCutE	NCutI	PIC
Iris	150	17	61	1
PenDigits01	200	28	23	1
PenDigits17	200	30	36	1
PolBooks	102	7	22	1
UBMCBlog	404	104	32	1
AGBlog	1,222	1095	70	3
20ngA	200	32	37	1

20ngB	400	124	56	3
20ngC	600	348	213	5
20ngD	800	584	385	10

Table 2. Runtime comparison (in milliseconds) of PIC and spectral clustering algorithms on synthetic datasets.[3]

NODES	EDGES	NCUTE	NCUTI	PIC
1K	10K	1,885	177	1
5K	250K	154,797	6,939	7
10K	1,000K	1,111	44142	34
50K	25,000K	--	--	849
100K	100,000K	--	--	2,960

Perhaps one of the greatest advantages of PIC lies in its scalability. Space-wise it needs only a single vector of size n for v^t and two more of the same to keep track of convergence. Speed-wise, power iteration is known to be fast on sparse matrices and converges fast on many real-world datasets; yet PIC converges even more quickly than power iteration, since it naturally stops when vt is no longer accelerating towards convergence.

IV. Parallel Power Iteration Clustering (PPIC)

In this section they describe methods in detail the parallel implementation of the PIC algorithm. They start this section with a brief discussion of different parallel programming frameworks.

The message passing interface (MPI) is a message passing library interface and is the defacto standard for performing communications in parallel programming environments. It has been traditionally used and is still the dominant model for high performance computing. OpenMP as well as POSIX Threads (Pthreads) are other parallel programming strategies, but both of these only work well on shared-memory multiprocessor systems. Due to its efficiency and performance for data communications in distributed cluster environments, they choose MPI as the programming model for implementing the parallel PIC algorithm. In the future it may consider a hybrid approach where Pthreads are used to exploit fine-grained parallelism within each node while MPI is used to exploit the parallelism across the available nodes.

This paper describes implementation of a parallelization of power iteration clustering–a recently developed clustering algorithm in this paper, they address the memory issue for storing the similarity matrix by splitting the data into small chunks and distributing these small chunks of data to multiple machines. They implemented the parallel PIC algorithm by using the message passing interface in short it is called (MPI).

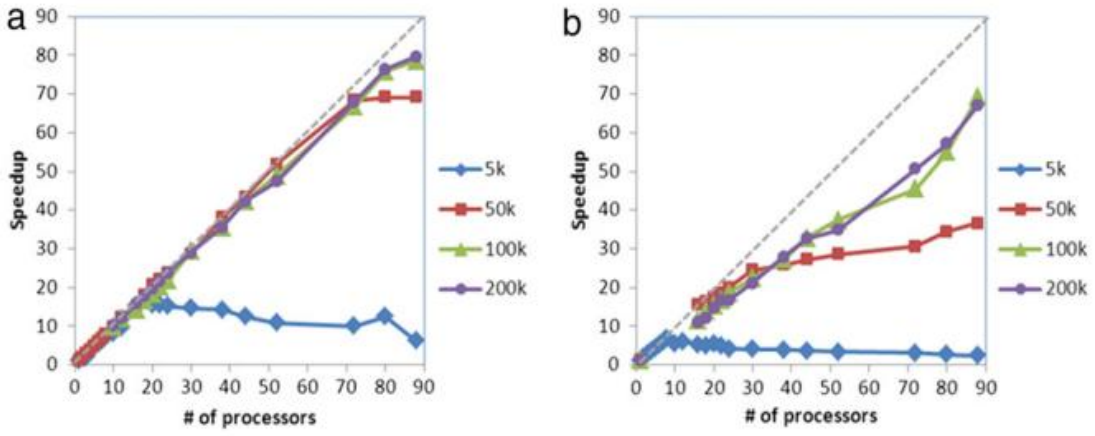

Fig.1 Speed up results for the pPIC algorithm on local cluster (a) and on Amazon EC2 machines (b).[5]

Their experimental results demonstrate that their parallel PIC implementation does scale well with respect to the data size. Fig. 1(a) shows the speedup of the p-PIC algorithm when run on a local cluster while Fig. 1(b) is on Amazon EC2. Overall the p-Pic algorithm achieves an almost linear speedup on all datasets, except for the 5 K dataset which shows decreased speedup as the number of processors increases. The reason for this is: the 5 K dataset is too small to have the processing speedup gained Overcome the communications and preprocessing.

V. Conclusion and Future Scope

From the review of the above papers, it can be concluded that many traditional different methods are there for the clustering and among all those methods spectral clustering suffers from a scalability problem in both memory use and computational time when the size of a data set is large. So in **2010, F. Lin** [3] develop power iteration method which uses matrix vector multiplication but still it is not good for large dataset and there was a memory issue.

So Parallel Power Iteration Clustering was developed and due to its parallel implementation it reduces the memory storage issue and its communication cost is also reduced but still there is an issue of node failure. So in this we can use Map reduce to solve this issue by implementing on Hadoop platform and we can use Nystrom Approximation for the clustering of our dataset which has a similar value.

Acknowledgement

I thanks to Assistant Professor Dinesh Waghela sir who is Head Of Department of computer and he is also my guide who inspire me for doing such work and due to their support here I am able to write this review paper and he also provide me good knowledge about big data and Guide me for doing this work.

References

1. M.Ester, H.P.Kriegel, J.Sander, X.W.Xu. ."DBSCAN" Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD-96, AAAI Press,1996,pp.226–231
2. M. Meila, J. Shi, "A random walk's view of spectral segmentation", AI and Statistics, AISTATS, 2001
3. F. Lin, W.W. Cohen, "Power iteration clustering",27th International Conference on Machine Learning, pp. 655–662, 2010.
4. W.Y. Chen, Y. Song, H. Bai, C. Lin, E.Y. Chang, "Parallel spectral clustering in distributed systems", IEEE Transactions on Pattern Analysis and Machine Intelligence33 (3) (2011)568–586.
5. WeizhongYana, Umang Brahmakshatriyaa, YaXuea, MarkGilderb, BowdenWisec"p-PIC: Parallel power iteration clustering for big data" J.ParallelDistrib.Comput. 73(2013)352–359

Author(s) Profile

Ankit Darji received the B.E. degree in Computer Engineering from Veer Narmad South Gujarat University in 2011.

Currently he is doing his M.E. Degree from Parul Institute of Technology of Gujarat Technological University and his interesting area of research is in Distributed Data Mining.

