An Efficient Implementation of Chronic Inflation based Power Iterative Clustering Algorithm

G. Aarthy Priscilla1 and Dr. A. Chilambuchelvan

1 Research Scholar, Manonmaniam Sundaranar University, Tirunelveli.
2Profesor & Head, Dept of Electronics and Instrumentation Engineering, R.M.D. Engineering College, Kavaraipettai . 601 206.

---***---

Abstract -In recent days the software plays a major role in application writing areas such that big data, it is used to store the information which is efficiently processed. Managing large datasets is one of the challenging task and also time consumption is also increases in large databases. In existing several serial algorithms were discussed but since some of the problems occurred hence while going to parallel data mining the efficiency and the speed want to take care. The problem statement is to find the failure and improve the effectiveness of learning algorithm. For parallel computing MapReduce is used as a programming model. This model helps to convert and redesign the existing sequential algorithm, in that the chronic inflation Based Power Iterative Clustering Algorithm (CIPIC) is proposed here to compute the Eigen vectors. When handling large data sets the CIPIC performs better result in MapReduce framework.

Keywords---Bigdata, Hadoop, Power Iteration, Chronic inflation, clustering.

I. INTRODUCTION

In data mining clustering is one of the important task. It consists of several applications such as network analysis, image processing and biomedical applications. The clustering analyze is one of the critical task because it need tom identify the hidden structure which is present inside the data. If the clustering process successfully completed then it is a best representation and more flexible. Clustering is also defined as dividing the number of data points into clusters, since the given nearby data points are grouped together and produce a cluster and name a selected head as cluster head. Distributing computing is one of the technique used here to solve the computational problem over a network. For dimensionality still more approaches are using traditional clustering but it has some limitations mentioned in Steinbach *et al.,(*2004). The Power Iteration Clustering (PIC) mentioned by Lin & Cohen (2010) is one of the simple and scalable clustering method, it finds a very low dimensional embedding of dataset using power iterations on similarity matrix of data. PIC provides an effective clustering indicator and outperform on real datasets with low dimensional data embedding using truncated power iteration on a similarity matrix. Under the survey the proposed design helps to improve the algorithm in simple in nature. It also provides the approximation and programming relaxation. This paper focusses on the execution of the chronic inflation namely it tries to improve the running time in large data set is used. Clustering is applied to any type of domains such as marketing analyzing, medical image segmentation, e-Learning and so on. In order to provide the best algorithm, complexity is to be minimize. Hence in detail analyze and discussions are made in following sections. Some of the notations and its background details are mentioned below :

Given a dataset X = {x1, x2,X3, ..., xn},is a similarity function s(xi , xj) is a function,

where s(xi , xj) = s(xj , xi) and

s ≥ 0 if i ≠ j, and following previous work (Shi & Malik, 2000), s = 0 if i = j.

An affinity matrix A ∈ Rn×n is defined by Aij = s(xi , xj). The degree matrix D associated with A is a diagonal matrix with dii = P j Aij. The applications of power iteration is to solve the computational problems. Sometimes it is used for calculating the page rank of documents which is searching in search engine. Some of the advanced eigen value algorithms can be understand by power iteration.

In general inflation describes that there is an increase in demand and service in a period of time. With respect to the time the values will be boosted up and produce a reflection in real value. The value trend to increase while clustering takes place, hence the algorithm provides the detail data sets such as {X_1, X2, X3,......X_n), then split the whole datsets to the sub data sets such as split 1, split 2,….split n.. calculate the affinity matrix W = $D^{-1}A$. Compute the overall sum from the slaves and find the initial vectors by $v0 = R/||R||1$. Then calculate the sub vector $v^t = W * V^{t-1}$ until the criteria met for inflation. At last find the inflation K vector and get the output cluster point$(C_1, C_2, C_3, \dots, C_k)$. Hence inflated based power iteration method is suitable only for the spectral clustering. Moreover the inflation based PIC is discussed by dhanpal & Perumal (2016). The drawback of this inflated algorithm is nearby pairs are detected hence the accuracy varies. To overcome the drawback chronic inflation based PIC is presented. Chronic PIC states that when a clustering experiences high inflation for a several decades due to increase in the supply chain. The algorithm is defined based on the given data sets and find the Exact K value of clusters. The detail discussion is made with chronic inflation in following chapters.

The rest of the paper is organized as follows, section 2 shows the related works, Section 3 gives the background and present limitations based on methodology, Section 4 gives the results and discussion and in section 5 conclusion and future work is made.

II. LITERATURE REVIEW

Several studies were made on various technologies which is stated in Fowlkes *et al*.,(2004), it gives the classical method to integral Eigen value problem. The problems are rectified by solving the grouping of computational complexity comparable. They concentrated on image segmentation in pair wise clustering. Several factors are considered such as proximity similarity and common fate. The upcoming research is focusing in cloud computing and big data because in increasing demand the large data sets and the allocation is also create complexity. The Nystrom extension is considered here for finding numerical approximations. Poteraş (2014) presented an optimized k means and standard k means in running sequence of different clusters. Running time is directly proportional to the optimization hence it is due to the reduction of data space that re visited at each loop.

Comparing traditional methods, PIC provides fast, simple and scalable. It is similar to the traditional, it requires data and matrix fit into memory which provides the big data applications. Yan *et al.,* (2013). They attempts to expand PICs data scalability by power iteration clustering. Different parallelization procedures are made to minimize the computation and communication costs. Parallel Power Iteration Clustering (p-PIC) provides a proper results to both data and compute resources. Clustering and matrix powering described by Tishby & slonim (2000), it deals with the markov process and provide the decay of mutual information. Here pairwise distances and markovian relaxation is taken for process. Relaxation of the mutual information is indicated for assumed matrix. The pearson correlation is estimated between two expressions without any losses.

Deflated power iterative clustering applied in spectral clustering (SC), it made a change to Eigen values and its efficiency is boosted up in the comparison of real data. Instead of finding the Eigen vector, PIC finds pseudo-Eigen vector, it is one of the combination of the Eigen vectors in linear time. The collision problem is rectified with more Pseudo eigenvectors, then the results are made in classical and realistic document. Instead of detecting the Eigen vectors the PIC algorithm computes the pseudo Eigen vector as a linear combination of vectors. Hence there is no data loss using PIC and provide better accuracy. Heller *et al*.,(2008) focused the lagrangian multiplier in order to find the lower Eigen vector. The rate of convergence is calculated to grow fast and the computational time will be less. Its main focuses is based on the problem occurs in large data sets as a inter collision. For better solutions the lower Eigen values and vectors are considered. To aim these objective several algorithms were presented such as lanczos algorithm, Jacob/ Davidson technique and so on. To achieve the regulation lagrangain multiplier is used to regulate the inflation with low Eigen value.

K- Means is used to solve the clustering problem because it is simple learning algorithm. By means of approximation, the dataset is classified. The advantages of k means is fast robust and easy to understand mentioned by Kanungo *et al.,* (2002). The obtained result is distinct and well separated from each other. Since some demerits are there, the learning algorithm needs appropriate specification of the cluster centers detail. This algorithm is not invariant to nonlinear transformations. It is unable to handle noisy data and outliers. And this algorithm fails for nonlinear data set. The analysis and implementation is made by requiring a kd tree with a data structure. In the form of filtering algorithm the Lloyd's algorithm is discussed. Here two ways of practical efficiency is achieved such as data sensitive and empirical studies on generated data and real data sets from the application in compression, quantization and segmentation. In filtering algorithm the multidimensional data points are storing in kd tree. Here kd tree is one of the binary tree that represents the subdivisions of point sets using axis aligned splitting hyper planes. The node associated with the closed box is called as cell.

III. METHODOLOGY

This chapter deals with the Inflated Power Iterative Clustering, deflation Power Iterative Clustering and Proposed Chronic Inflation Power Iterative Clustering in detail.

[1] Inflated Power Iterative Clustering (IPIC)

In spectral clustering the time consumption is more to compute the Eigen values and vectors. In PIC the Eigen vectors are considered. In some case the largest and very lowest Eigen values are detected. It uses Lagrangain multiplier to regulate the inflation by exponentially grows relative to the neighbors. The algorithm is as follows; initially the set of data points from X1 to Xn with a summation of number of K clusters. By using graph method construct the function which is given as s(xi,xj), then build the affinity matrix A with a_{ij}=s ($x_{i,}x_j$) if I ≠j and a_{ij}=0 if i=j. find the diagonal matrix and normalize the obtained matrix. After these process the new vector and rate of convergence is made with incrementing the value of t until the algorithm is value equals to zero. At last stage the clusters are pointed in a k dimensional subspace with the help of K means algorithm.

[2] Deflated Power Iterative Clustering

In spectral partitioning the large data sets are used in unsupervised clustering, it leads in complexity and infeasible. Too compute the Eigen values the time consuming is the major issue present in the clustering. The algorithm is given below. Initially the data set X=$\{x_1, x_2, x_3, \dots, x_n\}$ is considered. The input matrix is considered to calculate the matrix and solve the overall row sum from all slaves. Then calculate the sub vectors until it meet the criteria deflation..

Calculate the K vector by deflation factor,W, it is given by

$$W = \frac{W_{l-1} - W_{l-1} V_l V_l^T W_{l-1}}{V_l^T W_{l-1} V_l}$$

For cluster point V_t the K-mean is used. The output of the cluster is$(C_1, C_2, C_3, \dots, C_k)$.

[3] Proposed Chronic Inflation based Power Iteration Clustering

To obtain the optimal solution the proposed design gives the detail description, it is similar to the inflation data sets { $X_1, X_{2, \dots\dots} X_n$} with number of K clusters. Construct the graph based design like Gaussian function on a given set s($x_{i,}x_j$), it is given by

$$s(xi, xj) = e^{(-\frac{||xi - xj||}{2\sigma^2}} ,$$

Where, σ is a scaling parameter.

Normalize the obtained matrix and find the arbitrary vector $v^t = \gamma W v^{t-1}$ & $\delta^{t+1} = |v^{t+1} - v^t|$. Calculate the rate of convergence of chronic inflation method by

$$\frac{(\varepsilon_0 (t))}{\varepsilon_n (t)} \sim - e^{\sqrt{e_1 - e_{ut}}}$$

Where λ (t) is the lagrangain multiplier at time t. the cluster points on K dimensional subspace using K means algorithm.

MapReduce

Map Reduce is a processing technique used in distributed computing in java. It consists of two tasks such as Map and Reduce. Map consists of set of data and converts it into another set of data. The merit of this technique is easy to scale data processing over a computing nodes. Due to scaling of applications the hundreds and thousands of machines in a cluster is changed. Hence mostly programmers selected this model. MapReduce consists of three stages such as MapStage, shuffle stage and reduce stage.

MapStage: The map stage is used to process the input data. The input data is in the form of file or directory stored in Hadoop file system (HDFS). Each line is processed by the mapper function. Finally mappers process the data and creates the data in the form of chunks.

Reduce stage: It is a combination of shuffle stage and reduce stage. It process the data that comes from the mapper. After processing the data it stores in the HDFS as new set of outputs.

In each task the hadoop sends the map and reduce tasks to sever in the cluster. To reduce the network traffic the computing takes place on nodes with data on local disks.

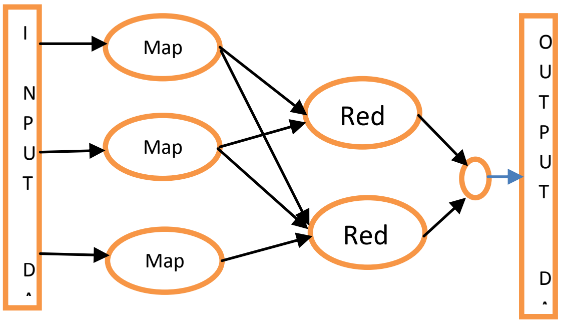

Figure 1 Map Reduce Algorithm Work flow

The execution model of the map reduce is shown in figure 1. It is explained in several steps, initially data is divided into M parts commonly known as chunks of default size. Hadoop is a fault tolerant technique since the data is being processed every time. In total nodes one node act as a master and it assigns the task to remaining nodes. The two tasks are Map Task (M) and a Reduce Task(R).

IV. RESULTS AND DISCUSSION

The experimental results were conducted based on cluster purity, accuracy and speedup ratio. These process made by MATLAB simulation tool.

1) Cluster Purity

To find the cluster purity, initially all clusters should arrange to a class which is most frequent in cluster. It is given by

$$Purity(\Omega + C)^n = \frac{1}{N} \sum_{k=0} \binom{max}{j} |\omega_k \cap c_j$$

Where $\Omega = \{ \omega_1 , \omega_{2 ,.......} \omega_k\}$ is the set of clusters and C is the set of classes.

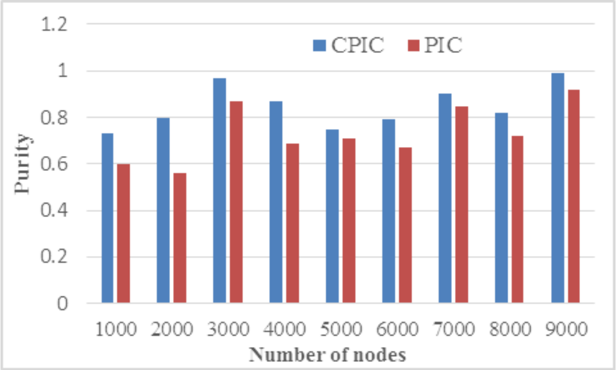

Figure 2 Clustering purity results for PIC and CIPIC

The purity results of PIC and the proposed Chronic Inflation Based Power Iterative Clustering (CIPIC) comparison are made Figure 2 .If there is decrease in t for large graphs, the number of edges does not increase too fast with respect to the number of nodes. Hence the PIC achieves lower purity than CIPIC in these cases. The results experimentally show that, in difficult cases, CIPIC produces a better solution PIC.

2) Accuracy

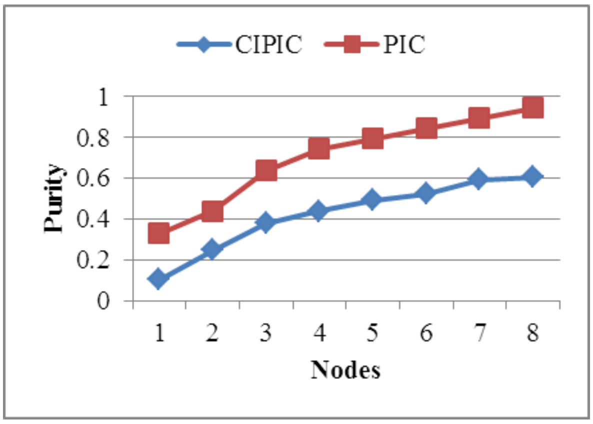

Figure 3 Accuracy for PIC and CIPIC Algorithm for Different Datasets.

From the figure 3, it shows the comparison between PIC and CIPIC, the accuracy of PIC is worst when compared it with proposed CIPIC algorithms. CIPIC computes the pseudo-eigenvector which is a linear combination of the k largest eigenvectors.

3) Test of speedup ratio

The performance is depends upon the execution time because it is an initial scale for consideration. The speedup is used to find how many times a parallel algorithm works faster than a serial algorithm. The speedup factor is inversely proportional to the time. The speedup is calculated using the formula

$$S=Ts/Tp$$

Where Ts is the execution time of the fastest sequential program and Tp is the execution time of the parallel program.

If a parallel program is executed on (p) processor, the highest value is equal to number of processors. In this system every processor needs Ts/p time of complete the job.

$$S=Ts/(Tp/p)=P$$

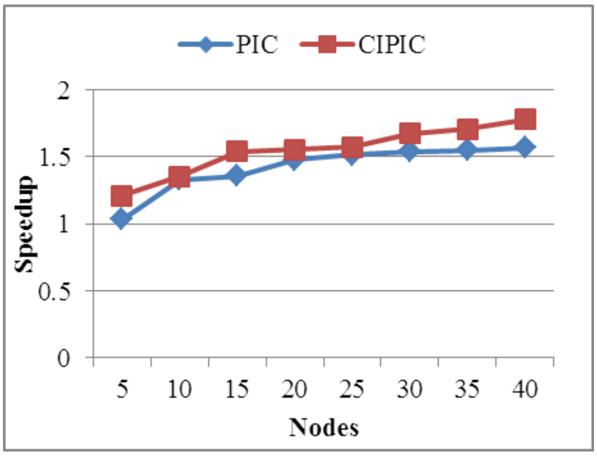

Figure 4 Speedup comparison with respect to the nodes.

The result of speedup ratio performance tests according to the various synthetic datasets are shown in figure 4. It is clearly shown that as the value of the speedup increases the execution time decreases. In the graph the dataset of different size for different nodes are given along the x axis and its execution time in ms is given along the y axis. Since the time taken for execution of the algorithm decreases we can conclude that there is an increase in speed up.

V. CONCLUSION

In clustering algorithm the MapReduce is applied to CIPIC algorithm and it is clustered over a large data sets. To avoid the collision and improve the accuracy, speedup ratio and clustering Purity. The computational complexity is reduced by using MapReduce Technique. From the comparisons the increase in nodes may results in speed and also accuracy. The speedup factor of the proposed design is improved with respect to the number of nodes in clustering. Hence in big data applications with large number of data sets this algorithm is applicable to perform. One possible future direction is to integrate this algorithm with real time processing units and also in medical image capturing techniques.

Reference

1]. Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In New directions in statistical physics (pp. 273-309). Springer Berlin Heidelberg.

2]. Fowlkes, C., Belongie, S., Chung, F., & Malik, J. (2004). Spectral grouping using the Nystrom method. IEEE transactions on pattern analysis and machine intelligence, 26(2), 214-225.

3]. Poteraş, C. M., Mihăescu, M. C., & Mocanu, M. (2014, September). An optimized version of the K-Means clustering algorithm. In Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on (pp. 695-699). IEEE.

4]. Lin, F., & Cohen, W. W. (2010). Power iteration clustering. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 655-662).

5]. Yan, W., Brahmakshatriya, U., Xue, Y., Gilder, M., & Wise, B. (2013). p-PIC: Parallel power iteration clustering for big data. Journal of Parallel and Distributed computing, 73(3), 352-359.

6]. Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8), 888-905.

7]. Tishby, N., & Slonim, N. (2000, November). Data clustering by markovian relaxation and the information bottleneck method. In NIPS (pp. 640-646).

8]. Dhanapal, J., & Perumal, T. (2016). Inflated Power Iteration Clustering Algorithm to Optimize Convergence Using Lagrangian Constraint. In Software Engineering Perspectives and Application in Intelligent Systems (pp. 227-237). Springer International Publishing.

9]. Heller, E. J., Kaplan, L., & Pollmann, F. (2008). Inflationary dynamics for matrix eigenvalue problems. Proceedings of the National Academy of Sciences, 105(22), 7631-7635.

10]. Weizhong Yana et.al (2013), p-PIC: Parallel power iteration clustering for big data, Models and Algorithms for High Performance Distributed Data Mining. Volume 73, Issue 3.

11]. http://en.wikipedia.org/wiki/Cosine_similarity

12]. Ran Jin, Chunhai Kou , Ruijuan Liu and Yefeng Li(2013), Efficient Parallel Spectral Clustering Algorithm design for large datasets under Cloud computing, Journal of Cloud computing: Advances, Systems and Applications 2:18.

13]. Niu XZ, She. K(2012), Study of fast parallel clustering partition algorithm for large dataset. Comput Sci 39; 134-151

14]. Kumar, A., Kiran, M., & Prathap, B. R. (2013, July). Verification and validation of mapreduce program model for parallel k-means algorithm on hadoop cluster. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-8). IEEE.

15]. Dhanapal, J., & Perumal, T. Analysis of Deflated Power Iteration Clustering on Hadoop Cluster.

16]. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.