Optimization of Data using Artificial Bee Colony Optimization with Map Reduce

Parikshit Patil[1]*, Hrishikesh Mhatre*[1]*, Ruchita Patil1, Apurva Shinde*[1]*, Bharti Joshi*[1]
[1]Department of Computer Engineering , Ramrao Adik Institute of Technology, Nerul.

Abstract - We are developing an idea through this paper which would give any question the perfect and best answer. Existing system is not capable of classifying according to different patterns. This compromises with efficiency of the system and quality of final result. In this project Parallel Clustering optimization method is formed by amalgamation of Map-Reduce with Ant Bee Colony Optimization Technique for improving the efficiency and success of the data science method. In addition, related running services on a Hadoop network are predicted with the help of Map Reduce algorithm.

Keywords: Map Reduce, Hadoop, Ant Bee Colony Optimization.

1. INTRODUCTION

The scope of data science and analytics is getting stronger day by day. Knowledge created by analyzing and processing data obtained fromdifferent sources can become useful in different ways. Instead of developing new and different algorithms to reach the efficient and satisfactory results, optimization of available techniques can be an option. Modification of core method used to implement those algorithms can be a better option [1]. The K-Means range algorithm is suited to e-commerce application development. The clustering algorithm works efficiently for numerical and categorical data as well as on large data sets clustering [2]. A novel hybrid optimization algorithm called BBO-PSO is proposed, which combines biogeography-based optimization with particle swarm optimization In this optimization algorithm, BBO will be used for local search while PSO will be used for mapping global data, which will allow the algorithm to have powerful search capabilities in solution space [3]. Insists strongly on increasing the convolution cycle speed but to reduce the difficulty of the convolution measurement [4]. The Particle Swarm optimization techniques instead of Ant Bee Colony Optimization with map reduce for maintaining clustering quality. New techniques need to be applied to parallel computing principles in order to be able to scale with that data set sizes [5]. Some algorithms that are capable of handling large and semi-structured data, are K-Means and ISODATA. The working analysis is carried out based on different factors, namely execution time, squared error sum, and R-square. Essentially, clustering is an unsupervised learning technique splitting into consequential groups of different similar data items. -category that is named as a cluster has similar object categories and divergent objects in other groups [6]. Sparse and compact zones should be consistent by clustering. Absolute dispersal trends and excellent collaboration lead to other areas of research such as text analysis, spatial database technology, and data extraction. Clustering was a vigorous research due to the enormous amount of data storage in databases and repositories. The experimental effects of various clustering approaches to perform reprovision of required information from massive repositories of data to make successful decisions on multiple applications. Analysis of the clusters is a difficult area. There are some distinct clustering requirements to handle high-dimensional data, user-defined limitations, noistic data, and to establish few input details by information of prior domain by using the type of attributes and the handling outliers. Our program is aimed at improving the current system while making the clustering techniques and the existing system more efficient.

2. LITERATURE SURVEY

2.1 In [4] proposed MR-CPSO algorithm increases the effectiveness of cluster quality. With growing data set sizes, it scales very well and achieves a very similar linear speed while retaining the cluster quality. The parameters focused while analyzing result of proposed system were scalability, speedup and clustering quality. In this system, on increasing number of nodes, the dataset sizes with the same ratio. But the PSO algorithm (which is the core algorithm used) is not so efficient as compared to other advance optimization algorithms like Artificial Bee Colony Optimization.

2.2 In [6] proposed the detailed analysis of performance and also the characterization of Hadoop k-means iterations by scaling different processor micro-architecture parameters. This system is based on the performance of k-means algorithm based on the hardware being used. Comparison regarding Intel and AMD processor is also discussed in this paper. The focused factors for analysis used in this paper are Total Run Time, Number of Sockets, and Number of Cores.

2.3 Authors in paper [3] proposed a winogard algorithm to improve the normal Convolution Neural Network. More precisely, it is related to the acceleration of widely used small convolution sizes. This improved algorithm is capable of reducing model sizes and increasing inference speed which leads to the increasing the ability of hardware accelerators and kernels to work with integer data. But this algorithm fails to maintain the efficiency of result with increasing data (or big dataset). Also, the quality of cluster is not guaranteed.

2.4 In paper [5] authors have derived evolutionary based clustering methods from the existing hard clustering methods. The main objective of the model was to find the optimum results. The different clustering algorithms used in this paper are ISODATA, K-means, Leader. The performance analysis of different clustering methods reveal that the rate of convergence is greater than the methods of partitioning. The different parameters used execution time, sum of squared errors, rms standard deviation. But, most important parameter like number of iterations is ignored in this evaluation. This algorithm is capable of giving better results irrespective of fitness value of different clustering algorithms Leader. The performance analysis of different clustering methods reveals that the rate of convergence is greater than the methods of partitioning. The different parameters used execution time, sum of squared errors, rms standard deviation. But, most important parameter like number of iterations is ignored in this evaluation. This algorithm is capable of giving better results irrespective of fitness value of different clustering algorithms.

2.5 Authors in paper [6] focused on inventing an algorithm which consists of amalgamation of Map Reduce and Artificial Bee Colony Optimization. In together the final system is called as Adaptive Artificial Bee Colony optimization. The algorithm is then compared with existing optimization algorithms like Artificial Bee Colony Optimization and particle swarm optimization. After analyzing the dataset, it is observed that new technique is very much efficient in all the fields like number of iterations required to get results, execution time. It also enhances the effectiveness of existing system. It reduces the search space thereby enhancing the time factor. The only disadvantage of this system is that there are no patterns introduced according to results we want. Thus, this leads to formation of unnecessary clusters which in turn reduces the efficiency of the system (increasing execution time, number of iterations unnecessarily). Table 1 represents the survey of Existing System done in this section.

Table 1. Survey of Existing System.

Sr. no.	Title of the paper	Algorithm	Advantages	Disadvantages	Result
1.	Parallel Particle swarm optimization algorithm [2]	PSO algorithm with Map Reduce	1. With Rising data set sizes it scales very well. 2.It achieves clustering quality.	Clustering algorithm is not so efficient and fast as compared to Artificial Bee Colony Optimization	The dataset scales with the increasing number of nodes.
2.	Performance characterization and analysis for Hadoop K-means iteration.[4]	K-means	Data efficiency is high	Efficiency of the result is mostly affected due to the number of cores.	Any changes to the micro-architecture processor parameters and to estimate performance or runtime can be accepted by the model. All the test cases generate error margin of five percent
3.	A parallel optimization of fast algorithm of Convolution Neural network on CPU.[5]	Convolution Neural Network based on winogard algorithm	1.It is more useful to speed up the algorithm. 2.Reduces the complexity of CNN.	1. Data efficiency decreases with increasing data set sizes. 2.Quality of Cluster is not guaranteed.	Used GLOFS to increase the amount by about 30 times.
4.	Performance analysis of partition and evolutionary Clustering Methods on various Cluster Validation Criteria.[6]	K-means, Leader, ISODATA	Different performance measures like execution time, sum of squared errors, rms standard deviation are discussed.	Important measure i.e., no of iterations is ignored.	In the simulation process, the algorithm gives the better outcome in the aspect of quantization error better outcome, though the fitness value for k-means falls below the local minima.
5.	Performance measures analysis based on parallel clustered optimization techniques for I-Commerce.[7]	Map Reduce Artificial Bee Colony Optimization Particle Swarm Optimization	Enhances the effectiveness and performance of me- Commerce.	In case of large datasets, many clusters are formed thereby reducing the efficiency of algorithm.	Proposed algorithm has shorter runtime than PSO. Enhanced searching ability in solution space.

3. PROBLEM DEFINITION

The data should be used at maximum efficiency and less processing time so the map reduce technique is used with simplified datasets that is obtain by applying pattern on datasets then Artificial Bee Colony optimization is applied. To compare the performance of Artificial Bee Colony Optimization on the different datasets (Map Reduce and Non-Map Reduce datasets). Compare the different datasets according to their execution time, no of iteration, size etc. To simplify datasets according to the type of database.

4. PROPOSED METHODOLOGY

The proposed algorithm for this paper is Artificial Bee Colony Optimization using Map Reduce.

Map Reduce: Map Reduce algorithm helps for processing data in parallel manner. This reduces time by distributing data in different nodes. As its name suggests, it contains two phases. They are as follows:

Map: In this section, keys and their respective values are mapped with each other. The tasks performed in Map phase are as follows:

The data is splitted into multiple chunks and given as the input to different node for parallel processing.

In next step, each key is mapped to the respective value.

The data is then passed to the reducer phase.

4.1.2 Reduce: This phase processes the data received from Map stage and produces the final output. The tasks performed in Reduce phase are as follows:

The data received from Map phase is shuffled and then arranged in organized way.

In this step, values of same keys are added and the data is reduced.

This is the final step of Reduce phase as well as Map Reduce. In this step, final output is saved into the data block.

In this way Map Reduce helps to improve the efficiency of data by storing and representing data in the form of key-value pair.

In this paper, we have incorporated Map Reduce with patterns which will help to store and use database according to the user needs. This avoids the unnecessary cluster formation.

The dataset processed with Map Reduce is Further processed with the help of Artificial Bee Colony Optimization and the performance is measured on the basis of parameters like execution time, iterations.

Artificial honey bees is made up of three types of individuals: employees, on lookers and scouts.

The Employee Bees and Onlooker Bees exploit the nectar sources around the bee hive this corresponds to the exploitation phase while the Scout Bees explore the solution domain this is the exploration phase.

Employed bees go back to their nectar source to recruit on looker bees through the honey bee waggle dance. The number of dance repetitions performance by the worker depends on the quality of the food source.

A scout bee start searching randomly around the hive to find a new necter source around it.

The algorithm for Artificial Bee colony Optimization is as Follows:

Initialization: - Use the random distribution method to initialize all possible solutions in a search space.

Employed phase search: -Create a new solution and change the fitness number as modified solution with the greatest possible fitness for any solution.

If possibility is improved modify the solution (If fre==0: tag 2) End loop (Until exit condition) and search optimized Result.

1+(1+fre) if fre > 0 (1)

1+absolute(fre) if fre < 0 (2)

Where fitness= fi is the cost value of solution for neighboring food source

jk = yjk+r(yjk-ylk) (3)

r=any real random no [0,1] (4)

Where, yjk = yk min + rand (0,1) (yk max – yk min) j= no of ways to find solution

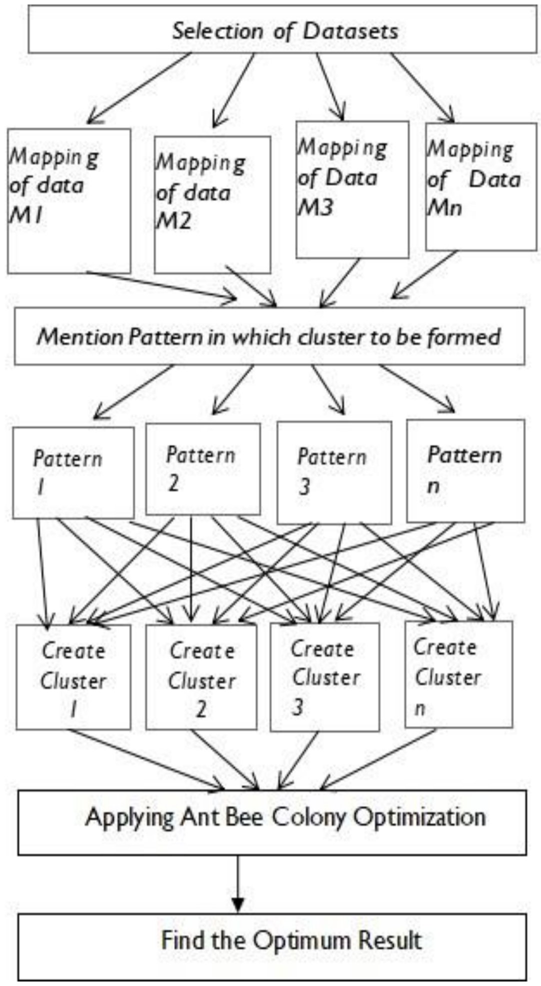
j= no of trails (5)

Distributions of probability: -Find potential candidate likelihood. If used bees are searched, the info of the highest hierarchy is passed. The probability of which segment depends is to find the best bees for honey.

on fitness value.

j= fitness j/j=1 n fitness (6)

Food collection by top-line bee Food from different places is scanned. Apply map reduce the cycle and produces the next nearest cluster by K-means. If quality is determined based on fitness interest, the bees test the minimum distance & quality of food



5. IMPLEMENTATION AND RESULTS

The Map Reduce algorithm with pattern has done its work flawlessly. The data sets used is the record of flights (Previous three months). The patterns apply for this data are able to find the best utilized flights where the passenger count is less than number of trips. (Patterns can be changed according to the problem statement or requirements).

The results of the map Reduce are further passed to the Artificial Bee Colony Optimization and the database is processed. The Best cost obtain in each iteration is 0 and it is consistent up to the last iteration. Also, the map reduce data is able to give results with a smaller number of iterations as compared to non-Map Reduce data. This also leads to faster processing of the data. Results are:

5.1. The data is passed through map reduce algorithm with pattern which gave value in key value pairs.

5.2. Map Reduce algorithm with pattern has done its work flawlessly. .

5.3. In Map Reduce, we got specific results according to the patterns introduced. The results of the map Reduce are further passed to the Artificial Bee Colony Optimization and the data obtained is processed.

5.4. The Best cost obtain in each iteration is 0 and it is consistent up to the last iteration. Also, the no. of iteration required to process Map Reduce data are very much less than that of required to process Non-Map Reduce data. Also, the different time taken by Artificial Bee Colony to process Map Reduce data as compared to that of non-map reduce data is significant. This shows that introduced map reducing along with pattern improves time efficiency of ABC. Following table shows the effect of map reduce on various factors of ABC. Table 2 explains results obtained in this section.

Table 2 : Results obtained using ABC Algorithms.

ABC Result		
Parameters	With map reduce	Without map reduce
Time	2.15565 s	79.8255 s
Iterations	147	1000

6.CONCLUSION

In this paper, Map Reduce and other parallel cluster optimization techniques are mixed up to set up the hybrid optimization algorithm. The different techniques integrate and extract the characteristics of both the Map Reduce and a particular technique Improve the quality of the search and high local search time in global search. Also the system can be incorporated with different patterns according to the user needs which in turn increase the efficiency and quality of cluster produced. The new optimization algorithm therefore has a good solution space searching capability and pattern is used for getting optimal results with less iterations.

The Future scope can be this Paper is Compare the performance of different parallel cluster Optimization techniques using different measures like 'iterations' , 'time'. To build the system which will cluster the data set according to the patterns. To choose perfect cluster Optimization techniques for a specific data based on their performance.

REFERENCES

1. Georgios P. Papamichail, Dimitrios P. Papamichail , “The k-means range algorithm for personalized data clustering in e-commerce”, European Journal of Operational Research, ELSEVIER,2007.

2. Ibrahim Aljarah and Simone A. Ludwig ,"Parallel Particle Swarm optimization Clustering Algorithm". Fourth World Congress on Nature and Biologically Inspired Computing,IEEE, 2012

3. Gang Cheng, Chao Lv, Shi Yan, Li Xu , "A Novel Hybrid Optimization Algorithm Combined with BBO and PSO",in Chinese Control and Decision Conference ,IEEE, 2016.

4. Joseph Issa, “Performance characteristics and analysis for Hadoop k-Means iteration,’Issa Journal of cloud computing:Advances, system, SpringerOpen,2016.

5. JiaHao Hunag, Tiejun wang, , Xuhui Zhu, Min Wei, Tao Wu, Xi Wu, Min Huang,"A Parallel Optimization of Fast Algorithm of convolution Neural network on CPU”, 10’thInternational Conference on Measuring Technology and Mechatronics Automation,IEEE, 2018.

6. R. S. M. Lakshmi Patibandle, N. Veeranjaneyulu, ‘Performance Analysis ofPartition and Evolution Clusterin

Methos on Various Cluster Validation Criteria'er Open, 2018.

7. Mr.Likhesh Kolhe, Dr. Vaishali Khairnar, Dr. Ashok kumar jetwa;' Performance Measures Analysis Using Parallel Clustered Optimization techniques For I-Commerce,"8th International Conference on cloud computing, data science Enginnering,IEEE,2018.

8. https;//drive.google.com/open?id=1rZsvBXPhl NeBMZRqKINhpGUJrYTqNqc

9. https://drive.google.com/open?id=1pkVTQCd94EA6NGuw_WRQrOjt52b37PVu.

10. https://drive.google.com/open?id=1jBydlmonXOmr1SrhJWv c6CZgsqrnR