LARGE BIOLOGICAL DATASET ANALYSIS USING ENHANCED MAP REDUCING METHOD WITH MODIFIED ARTIFICIAL BEE COLONY OPTIMIZATION (MABC)

S.A.Gowri Manohari,Research Scholar - Department of Computer Science,Kovai Kalaimagal College of Arts & Science,Coimbatore,Tamil Nadu

Mr.S.Jawahar,Assistant Professor - Department of Computer Science,Kovai Kalaimagal College of Arts & Science,Coimbatore,Tamil Nadu

ABSTRACTs

Due to advances in high-throughput biotechnologies biological information is being collected in databases at an amazing rate, requiring novel computational approaches that process collected data into new knowledge in a timely manner. In this work proposed an effective technique for reducing the resource problems in clouds using the enhanced map reducing algorithm. The map reduce algorithm generates a solution which is further optimized with the help of optimization algorithm. It shows an effectual topology for solving the resource model for resource problem solution. Deployment of optimization algorithm opted for multi objective problem is Modified Artificial Bee Colony Algorithm (MABC) which delivered best optimized result and less computation time is utilized to unfurl the determined objective, upshots of the proposition topology has depicted promising and effectual results and minimized computational exertion. In this proposed method, the execution time is reduced largely when compared to the existing method.

Keywords: Modified Artificial Bee Colony Algorithm (MABC), enhanced map reducing algorithm

1. INTRODUCTION

In many branches of scientific information is collected in tables, forms or questionnaires. Most biological databases, for example, accumulate knowledge by annotating or curating different biological objects or their relationships [1]. This information includes, but is not limited to, characteristics of sequenced genomes [2], genes [3–5], chemicals [6 , 7] and enzymes/metabolic pathways [8–10]. With advances in high-throughput sequencing and omics technologies, the number of such resources is growing at an unprecedented rate [11–13]. To facilitate their usage, a dedicated academic journal that introduces their description [14] and even a new resource, BioDBCore, to collect attributes of the databases, has emerged [15]. While databases help scientists to gather and integrate massive amounts of information by downloading various types of data, the task of identifying hidden regularities in the data is left open [16]. For this reason, computational approaches that sift non-spurious associations hidden in large and complex data and discover clusters of these annotations are needed.

One known approach to mining associations in large data sets is association rule (Arule) learning [17]. This algorithm was initially designed to find frequently associated products in supermarket-sale data to understand consumer purchasing behaviors. Recently, the technique was applied to mine biological associations: to identify a predictive combinations of genes in the genotype–phenotype relationships [18], to discover adjacent amino acids on a binding site of a protein complex [19], to analyze disordered proteins in prokaryotes [20] and to extract combinations of gene annotations from a list of over-expressed genes [20 , 21]. Association rule learning, however, has serious drawbacks for extracting hidden regularities among biological annotations.

The Modified artificial bee colony (MABC) algorithm, which is a biologically inspired population-based metaheuristic algorithm, was recently introduced for continuous function optimization by Karaboga [22]. Due to its simplicity and ease of implementation, the MABC algorithm has been extensively applied to both continuous and discrete optimization problems since its invention. Various comparison studies, in which the MABC algorithm was compared to novel metaheuristic algorithms, such as particle swarm optimization (PSO), differential evolution (DE), and genetic algorithm (GA), have been performed to show its effectiveness [23, 24]. These studies show that the MABC algorithm outperforms other novel algorithms on several instance problems.

In this work section 2 discussed about the literature review, section 3 discussed about the proposed method about the parallel preprocessing using Enchanced map reduce method and optimization process using with the Modified ABC method.

2. LITERATURE REVIEW

In this introduces the basic concepts of Parallel processing, hybrid Map reducing and Multivariable selection using optimization techniques.

In [25] proposed the adoption of a community-defined, uniform, generic description of the core attributes of biological databases, BioDBCore. The goals of these attributes are to provide a general overview of the database landscape, to encourage consistency and interoperability between resources; and to promote the use of semantic and syntactic standards. BioDBCore will make it easier for users to evaluate the scope and relevance of available resources. This new resource will increase the collective impact of the information present in biological databases.

In [26] the astonishing rate of data generation by these low-cost, high-throughput technologies in genomics is being matched by that of other technologies, such as real-time imaging and mass spectrometry-based flow cytometry. Success in the life sciences will depend on our ability to properly interpret the large-scale, high-dimensional data sets that are generated by these technologies, which in turn requires us to adopt advances in informatics. Here it discuss how it can master the different types of computational environments that exist — such as cloud and heterogeneous computing — to successfully tackle our big data problems.

In [27] proposed to make use of the MapReduce programming model which achieves multifold scalability on a set of labeled graphs. It tested this method on both real and synthetic datasets. To the best of this knowledge, this is the first attempt to implement transaction graphs using the MapReduce model. There are many promising main memory-based techniques available in this area, but they lack scalability as the main memory is a bottleneck. Taking the massive data into consideration, traditional database systems like relational databases and object databases fail miserably with respect to efficiency as frequent subgraph mining is computationally intensive. With the advent of the MapReduce framework by Google, a few researchers have applied the MapReduce model on a single graph for mining frequent substructures.

In [28] proposed an efficient approach for mining maximal contiguous frequent patterns in large DNA sequence data using MapReduce framework which can handle a massive DNA sequence datasets with a large number of nodes on a Hadoop platform. Current DNA sequence datasets have become extremely large, making it a great challenge for single-processor and main-memory-based computing systems to mine interesting patterns. Such limited hardware resources make the performance of most Apriori-like algorithms inefficient. However, recent implementation of a MapReduce framework has overcome these limitations. Furthermore, mining with maximal contiguous frequent patterns to express the function and structure of DNA sequences is a useful technique, capable of capturing the common data characteristics among related sequences.

In [29] presents here msABC, a coalescent-based software that facilitates the simulation of multi-locus data, suitable for an ABC analysis. msABC is based on Hudson’s ms algorithm, which is used extensively for simulating neutral demographic histories of populations. The flexibility of the original algorithm has been extended so that sample size may vary among loci, missing data can be incorporated in simulations and calculations, and a multitude of summary statistics for single or multiple populations is generated.

In [30] proposed a new hybrid gene selection method, namely Genetic Bee Colony (GBC) algorithm. The proposed algorithm combines the used of a Genetic Algorithm (GA) along with Artificial

Bee Colony (ABC) algorithm. The goal is to integrate the advantages of both algorithms. The proposed algorithm is applied to a microarray gene expression profile in order to select the most predictive and informative genes for cancer classification. In order to test the accuracy performance of the proposed algorithm, extensive experiments were conducted. Three binary microarray datasets are use, which include: colon, leukemia, and lung. In addition, another three multi-class microarray datasets are used, which are: SRBCT, lymphoma, and leukemia.

In [31] proposed an innovative feature selection algorithm, minimum redundancy maximum relevance (mRMR), and combine it with an ABC algorithm, mRMR-ABC, to select informative genes from microarray profile. The new approach is based on a support vector machine (SVM) algorithm to measure the classification accuracy for selected genes. It evaluate the performance of the proposed mRMR-ABC algorithm by conducting extensive experiments on six binary and multiclass gene expression microarray datasets. Furthermore, it compare this proposed mRMR-ABC algorithm with previously known techniques. It reimplemented two of these techniques for the sake of a fair comparison using the same parameters. These two techniques are mRMR when combined with a genetic algorithm (mRMR-GA) and mRMR when combined with a particle swarm optimization algorithm (mRMR-PSO).

In [32] introduced the Maximal Information Coefficient (MIC) has been proposed to discover relationships and associations between pairs of variables. It poses significant challenges for bioinformatics scientists to accelerate the MIC calculation, especially in genome sequencing and biological annotations. It explored a parallel approach which uses MapReduce framework to improve the computing efficiency and throughput of the MIC computation. The acceleration system includes biological data storage on Hadoop Distributed File System (HDFS), preprocessing algorithms, distributed memory cache mechanism, and the partition of MapReduce jobs. Based on the acceleration approach, is extend the traditional two-variable algorithm to multiple variables algorithm.

In [33] describes a modified ABC algorithm for constrained optimization problems and compares the performance of the modified ABC algorithm against the algorithms for a set of constrained test problems. For constraint handling, ABC algorithm uses Deb’s rules consisting of three simple heuristic rules and a probabilistic selection scheme for feasible solutions based on their fitness values and infeasible solutions based on their violation values. Moreover, a statistical parameter analysis of the modified ABC algorithm is conducted and appropriate values for each control parameter are obtained using analysis of the variance (ANOVA) and analysis of mean (ANOM) statistics.

In [34] proposed an improved ABC algorithm called gbest-guided ABC (GABC) algorithm by incorporating the information of global best (gbest) solution into the solution search equation to improve the exploitation. Artificial bee colony (ABC) algorithm invented recently by Karaboga is a biological-inspired optimization algorithm, which has been shown to be competitive with some conventional biological-inspired algorithms, such as genetic algorithm (GA), differential evolution (DE) and particle swarm optimization (PSO). However, there is still an insufficiency in ABC algorithm regarding its solution search equation, which is good at exploration but poor at exploitation.

3. PROPOSED METHODOLOGY

Parallel processing approach is proposed which uses the enhanced Map Reduce removes the barriers of the traditional Map Reduce. The algorithm for multi-variable situation will be investigated by using Modified Artificial Bee Colony Algorithm (MABC) algorithm.

3.1.ENCHANCED MAP REDUCE FRAMEWORK

MapReduce program is divided into two stages Map and Reduce. Map stage writes the output locally and Reducers aggregates the output by remotely reading from the Mappers. This process of transferring the data is called as Shuffling. In this Non-conventional MapReduce we device a technique by which we bypass the sorting mechanism and modify the invocation of reduce function so that it can be called with a small set of records. Reducers no longer needs to wait for remotely read from Mappers and then to be grouped. Due to this performance gets improved as now Reducers need not to wait till the Mappers complete their whole work and shuffling gets completed.

In Non-conventional Map Reduce we overcome the following overheads:

1. Waiting time between Remote reading of the first and the last records.

2. Time taken for sorting of whole records.

3. Maintaining a local buffer for each Mapper.

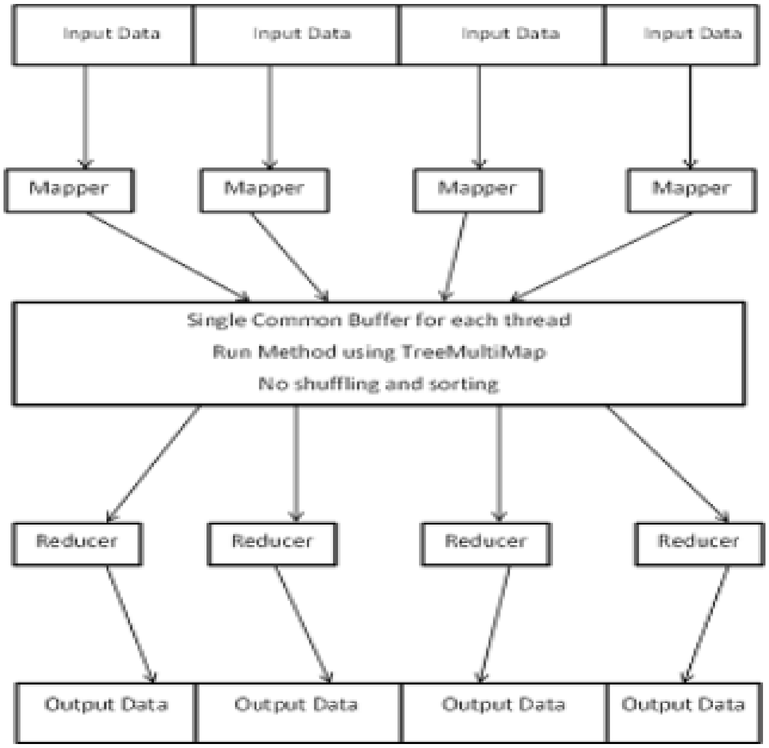

Figure 1: Enhanced Map Reduce Framework

In this Non-conventional Map Reduce the intermediate results are not stored as a whole for each and every key. And Reduce function works at one record at a time. In traditional Map Reduce shuffle stage is designated to work in an efficient and asynchronous manner by providing a thread for each and every Mapper which reads the data from Mappers and this data is stored in their local Buffers which is then merge sorted. Then the key and corresponding values are passed on to the Reduce function.

Map Reduce programming model is used to develop a parallel applications that process and generate a large amount of data. Map Reduce provides high scalability and reliability because of the division of the work into smaller units. The data are mainly given to master key pair value, which is responsible for managing the execution of applications in the cluster. Map Reduce is extremely appropriate for huge data searching and processing operations. For traditional clusters, the model has shown excellent I/O features, which is apparent from its successful application in large scale search applications by Google. Data in the MapReduce frame work are usually delineated in the form of key-value pairs . In the Primary step of the computation is the map function where the frame work reads input data and optionally changes it in to proper key-value pairs. In the Second step which is the map phase, on each pair < key, value>. A function F which returns a multi set of new keyvalue pairs is applied. The function is expressed as below. F({key,value}){<key1, value1>,<key2,value2> … ,<keyn,valuen> }
(1)

In the reduce phase, all pairs that are generated in the preceding step are grouped according to their keys and the their values are reduced using a function which is given by,

R({<key1, value1>,<key2,value2> … ,<keyn,valuen> })= <keyn,valuen>} (2)

The Input data set which is a document is first applied to the mapper, which generates the key value pair. Here key is the document name and value is the document content. The same words and the number are grouped as key-value pair and this is given to the reduce function. The reduce function then obtains all the pairs of word with the same key and value and counts the number of pairs in the document and then reduces the count by considering the same key-value pair as a single one. Once the map Reduce step is over next the data's are stored in HDFS.

When the program call MapReduce function, the following sequence of actions occurs, The Map Reduce function first splits the input files into number of Sections. Then it starts producing number of copies of the program on cluster of key pair values. Among this, one is the Master, which assigns the process to the remaining key pair value which is the function.

The Intermediate key/value pair produced by the Map function is then given to the memory for storage. The Location is then passed to the Master Key pair value which then forwards this location to reduce function. When the reduce function get the location of the intermediate data output, it reads the entire intermediate data and based on the keys they are stored such that data with same key are collected into one group. This process continues until the entire Map and Reduce tasks are completed.

3.2. ARTIFICIAL BEE COLONY ALGORITHM

The ABC algorithm is a population-based metaheuristics algorithm that mimics the foraging behavior of honey bee swarms. The ABC algorithm classifies bees in a colony into three main groups: employed bees, onlooker bees, and scout bees. Employed bees are responsible for exploiting the food sources and sharing the information about these food sources. Onlooker bees wait in the hive and take the food source information from employed bees to make a decision on further exploiting the food source. Scout bees randomly search the environment to find a new food source.

In the ABC algorithm, each candidate solution to the problem is associated with a food source and is represented by an -dimensional real-coded vector. The quality of a solution corresponds to the nectar amount on that food source, and one employed bee explores each food source. In other words, the number of the employed bees is equal to the number of food sources. The colony is equally divided into employed and onlooker bees. A food source, which cannot be improved for a predetermined number of tries, is abandoned and the employed bee associated with that food source becomes a scout. In the ABC algorithm, the employed and onlooker bees are responsible for exploiting, whereas the scout bees handle exploring.

The main steps of the ABC algorithm are as follows [22]:

(1) initialization,

(2) evaluating the population,

(3) repeat,

(4) employed bee phase,

(5) onlooker bee phase,

(6) scout bee phase,

(7) until (termination criteria are satisfied).

a. Initialization

In the initialization step, the ABC algorithm generates a randomly distributed population of SN solutions (food sources), where SN denotes the number of employed or onlooker bees. Let $x_i = \{x_{i,1}, x_{i,2}, \ldots\ldots, x_{i,D}\}$ represent the i^th food source, where D is the problem size. Each food source is generated within the range of the boundaries of the parameters by

$$x_{i,j} = x_j^{min} + rand\ (0,1)(x_j^{max} - x_j^{min}) \qquad (3)$$

Where i=1,…SN, j=1,….,D. x_j^{min}, and x_j^{max} are the lower and upper bound for the dimension j, respectively.

b. Employed Bee Phase

In the employed bee phase, employed bees generate a neighboring food source $v_i$by performing a local search around each food source $i \in \{1,2, \ldots.., SN\}$ as follows:

$$v_{i,j} = x_{i,j} + \phi_{i,j}(x_{i,j} - x_{k,j}) \qquad (4)$$

where j is a random integer in the range [1,D] and $k\epsilon\{1,2, \ldots.., SN\}$ is randomly chosen food source that is not equal to i . $\phi_{i,j}$ is a random number in the range [-1,1] . A greedy selection is applied between x_i and v_i in which the better solution will be retained. Then, employed bees will return to their hive and share the information on new solutions with onlooker bees.

c. Onlooker Bee Phase

Onlooker bees select a food source depending on the probability value prob associated with that food source. The value p is calculated as follows:

$$prob_i = \frac{f_i}{\sum_{j=1}^{SN} f_j} \quad (5)$$

where f_i is the objective function value of solution . By using this mechanism, food sources having better fitness values will be more likely to be selected. Once the onlooker bee has chosen the food source, she generates a new solution using (3). As in the employed bee phase, a greedy selection is carried out between x_i and v_i .

In the employed bee phase, a local search is applied to every food source, whereas only the selected food sources will be updated in the onlooker bee phase.

d. Scout Bee Phase

If a food source cannot be improved for a predetermined number of tries, then the employed bee associated with that food source becomes a scout bee. Then, the scout bee finds a new food source using (4). After the scout bee finds a new source, she becomes an employed bee again.

3.2.1. MODIFIED ARTIFICIAL BEE COLONY ALGORITHM (MABC)

All stochastic optimization methods are based on two principles: exploitation and exploration. Exploitation is the ability to search in neighborhood of a good solution in order to find better candidate solutions if not the global optima while exploration refers to the ability to breadth search the solution space in order to find new candidate solutions. The success of any such optimization algorithm depends on an appropriate balance between exploitation and exploration processes. From the above description of ABC algorithm, we can consider that the process of exploitation is performed by employed and onlooker bees, while scout bees are responsible for the process of exploration. According to the solution search equation of ABC algorithm described by (5), in each cycle, a new candidate solution is generated by moving the old solution towards (or away from) another solution selected randomly from the population. However, there is no guaranty that the new candidate solution is better than the previous one. Therefore, the exploitation abilities of the algorithm need to be fostered through this equation. This requires modifying this equation in a way to guide the search towards promising regions. In addition, in the classical ABC, a new solution is produced by changing only one parameter of the memorized solution, which results in a slow convergence rate as stated in [35].

In our Modified ABC has been proposed to increase the exploitation of classical ABC. It modify the search solution described,

$$v_{i,j}(G+1) = z_{gbest,j}(G) + F * \left(z_{i,j}(G) - z_{r1,j}(G)\right) + F * \left(z_{r2,j}(G) - z_{r2,j}(G)\right) \qquad (6)$$

The use of the global best (gbest) solution in employed and onlooker bees' phase can drive the new candidate solution towards the global best solution; therefore, the exploitation of ABC algorithm can be increased.

4. EXPERIMENTAL RESULTS

In this section, it describe the experiments on Modified ABC (MABC) solution, including the execution time, accuracy and error evaluation to be consider. Compared with the existing work based Maximal Information coefficient (MIC), Mapreduce(MR) and Hybrid Mapreduce (HMR). Moreover, the accuracy and efficiency of the multi-variable algorithm is also presented. Due to that the complexity of original algorithm increases notably fast when data set grows, it would be impossible to compute a large enough data set in one server. So in the speed experiments, we mainly measure the speedup in line with the increase of the cluster size to show the speed advantage of our parallel algorithm. Also set up a small cluster of 7 nodes, each of which runs Fedora 17 on Xeon dual-core 2.53 GHz CPU and 6GB memory. All machines are connected with a single gigabit Ethernet link. All nodes are in the same hosting facility and therefore the round-trip time between any pair of machines was less than a millisecond. In our experiment, we use Hadoop 1.0.3 as basic storage and processing framework, which supports HDFS federation and Private Cloud was formed with three machines and Hadoop was configured on this Private Cloud. It has been seen that our proposed method of map reduce using the MABC has shown are markable reduction in the execution time when compared to other existing method.

Accuracy: The Accuracy (Acc) is defined as (TP+TN)/Total Population

Where TP is the number of true Positives in the dataset, FN is the number of false negative and FP is the number of false positives. In this accuracy terms the proposed method is compare with the other two methods. That is known as the MIC method accuracy is very low compare than the Mapreduce method. Then the Mapreduce method is high then the MIC. Then the Hybird Mapreduce method is high then the Mapreduce method. The accuracy of the Modified ABC (MABC) method is high than the Hybrid Mapreduce is illustrated in figure 2.

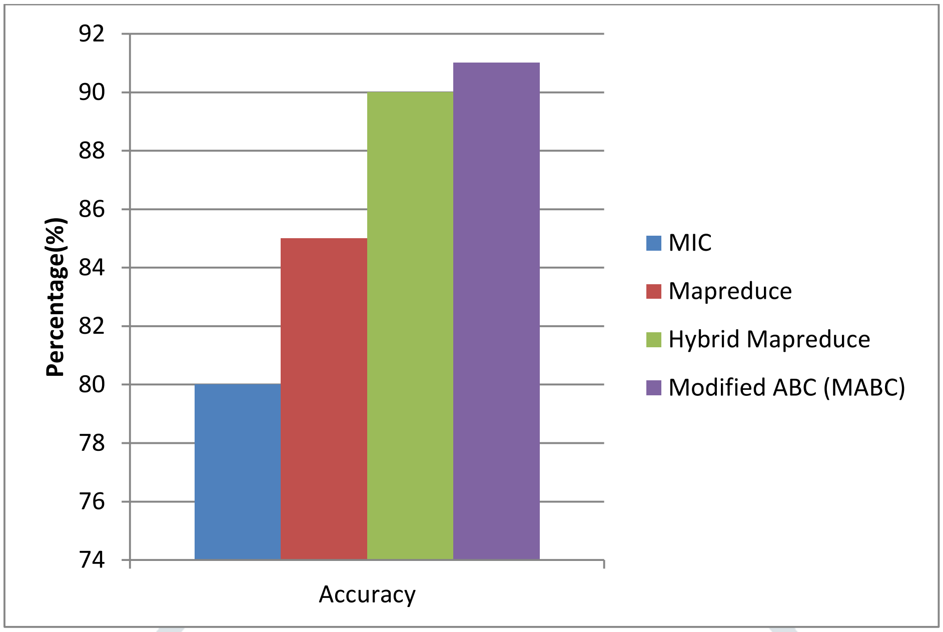

Figure 2. Comparison of the Accuracy

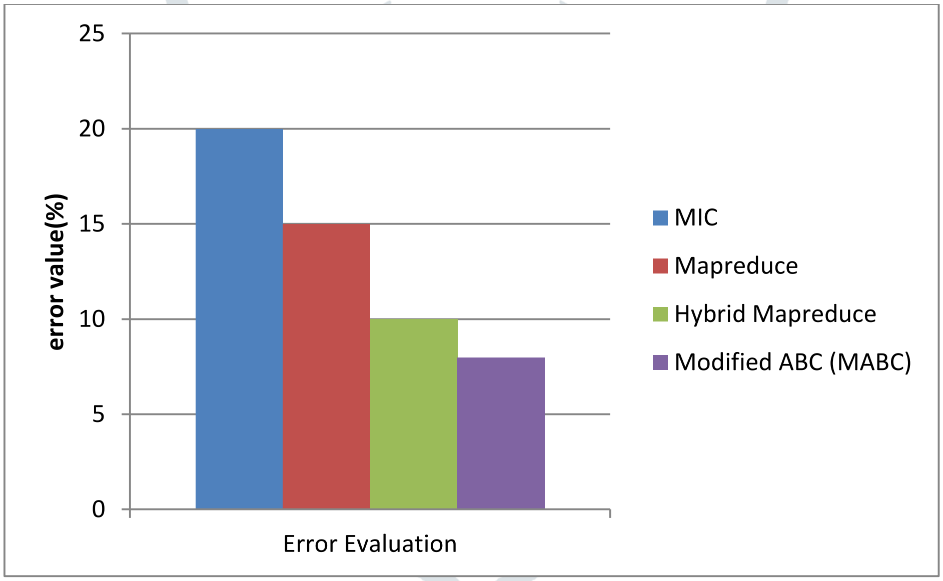

Figure 3. Comparison of the error evaluation

The total error evaluation is the systematically error or the random error. It is compare between the MIC, Mapreduce and Hybrid Mapreduce with the proposed method that is known as Modified ABC (MABC). In this proposed method is very less error compare than the other three methods are shown in figure 4.

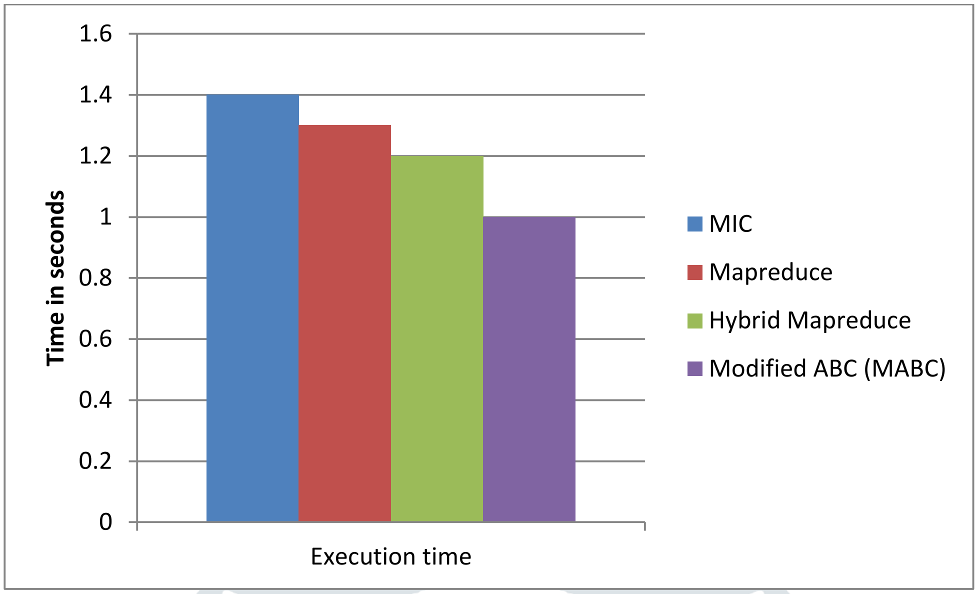

Figure 4. Comparison of Execution time

The execution time of the overall Modified ABC (MABC) computation procedure with different servers and different input dataset size. The execution time does not reduce while the server number is increasing is shown in figure 6. The reason is that MapReduce jobs need lots of network communications and disk I/O, so if the data size is small, the communication delay and I/O cost will be more obvious. This situation will disappear when input data size increases, the execution time decreases dramatically when adding more servers. For example, it needs 2087 seconds to finish the computation in one server but it only cost nearly 653 seconds (31.3%) to finish the same computation using three servers. The Modified ABC (MABC) solution can achieve linear speedup according to the experiments. Experimental results demonstrate that the accuracy of the multi-variable algorithm can achieve high accuracy when the number of reducer tasks is larger.

CONCLUSION

In this paper, it has proposed an effective technique for reducing the resource problems in clouds using the map reducing algorithm. The map reduce algorithm generates a solution which is further optimized with the help of optimization algorithm. The optimization algorithm we utilized is the Modified artificial bee colony algorithm (MABC). The proposed method of resource problem reduction proves to be more effective as it reduces the requirements needed for storage of data to a large extend. As this proposed method, the graph shows the execution time is reduced to a large instant when compared to the existing method. Thus it proved to be an efficient method in reducing the resource problems.

REFERENCES

1. Howe,D., Costanzo,M., Fey,P., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., St Pierre,S. et al. (2008) Big data: the future of biocuration. Nature, 455, 47–50.
2. Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res., 29, 126–127.
3. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. et al. (2010) The Pfam protein families database. Nucleic Acids Res., 38, D211–D222.
4. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford), 2011, bar009.
5. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res., 39, D52–D57.
6. Wiegers,T.C., Davis,A.P., Cohen,K.B., Hirschman,L. and Mattingly,C.J. (2009) Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). BMC Bioinformatics, 10, 326.
7. Davis,A.P., King,B.L., Mockus,S., Murphy,C.G., SaraceniRichards,C., Rosenstein,M., Wiegers,T. and Mattingly,C.J. (2011) The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Res., 39, D1067–D1072.
8. Kanehisa,M. (2002) The KEGG database. Novartis Found Symp., 247, 91–101, discussion 101–103, 119–128, 244–152.
9. Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res., 32, D431–D433.
10. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M. et al. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res., 38, D473–D479.
11. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. BMC Med. Genet., 10, 6.
12. Forbes,S.A., Tang,G., Bindal,N., Bamford,S., Dawson,E., Cole,C., Kok,C.Y., Jia,M., Ewing,R., Menzies,A. et al. (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res., 38, D652–D657.
13. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. et al. (2011) NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Res., 39, D1005–D1010.

14. Landsman,D., Gentleman,R., Kelso,J. and Francis Ouellette,B.F. (2009) DATABASE: a new forum for biological databases and curation. Database (Oxford), 2009, bap002.
15. Gaudet,P., Bairoch,A., Field,D., Sansone,S.A., Taylor,C., Attwood,T.K., Bateman,A., Blake,J.A., Bult,C.J., Cherry,J.M. et al. (2011) Towards BioDBcore: a community-defined information specification for biological databases. Database (Oxford), 2011, baq027.
16. Attwood,T.K., Kell,D.B., McDermott,P., Marsh,J., Pettifer,S.R. and Thorne,D. (2009) Calling International Rescue: knowledge lost in literature and data landslide! Biochem. J., 424, 317–333.
17. Agrawal,R., Imielin´ski,T. and Swami,A. (1993) Mining association rules between sets of items in large databases. In: ACM SIGMOD Proceedings of the 1993 ACM SIGMOD international conference on Management of data, Vol. 22. ACM Press, New York, NY, USA, pp. 207–216.
18. Tamura,M. and D'Haeseleer,P. (2008) Microbial genotype-phenotype mapping by class association rule mining. Bioinformatics, 24, 1523–1529.
19. Kuo,H.C., Ong,P.L., Lin,J.C. and Huang,J.P. (2011) Discovering amino acid patterns on binding sites in protein complexes. Bioinformation, 6, 10–14.
20. Pavlovic-Lazetic,G.M., Mitic,N.S., Kovacevic,J.J., Obradovic,Z., Malkov,S.N. and Beljanski,M.V. (2011) Bioinformatics analysis of disordered proteins in prokaryotes. BMC Bioinformatics, 12, 66.
21. Hackenberg,M. and Matthiesen,R. (2008) Annotation-Modules: a tool for finding significant combinations of multisource annotations for gene lists. Bioinformatics, 24, 1386–1393.
22. D. Karaboga, “An idea based on honey bee swarm for numerical optimization,” Tech. Rep. TR06, Erciyes University Press, Erciyes, Turkey, 2005.
23. D. Karaboga and B. Basturk, “On the performance of artificial bee colony (ABC) algorithm,” Applied Soft Computing Journal, vol. 8, no. 1, pp. 687–697, 2008.
24. D. Karaboga and B. Basturk, “A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm,” Journal of Global Optimization, vol. 39, no. 3, pp. 459–471, 2007.
25. Gaudet, P., Bairoch, A., Field, D., Sansone, S.A., Taylor, C., Attwood, T.K., Bateman, A., Blake, J.A., Bult, C.J., Cherry, J.M. and Chisholm, R.L., 2011. Towards BioDBcore: a community-defined information specification for biological databases. *Database*, *2011*.
26. Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. and Nolan, G.P., 2010. Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, *11*(9), p.647.
27. Hill, S., Srichandan, B. and Sunderraman, R., 2012, October. An iterative mapreduce approach to frequent subgraph mining in biological datasets. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*(pp. 661-666). ACM.

28. Karim, M.R., Hossain, M.A., Rashid, M.M., Jeong, B.S. and Choi1, H.J., 2012. A MapReduce framework for mining maximal contiguous frequent patterns in large DNA sequence datasets. *IETE Technical Review*, *29*(2), pp.162-168.
29. Pavlidis, P., Laurent, S. and Stephan, W., 2010. msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources*, *10*(4), pp.723-727.
30. Alshamlan, H.M., Badr, G.H. and Alohali, Y.A., 2015. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. *Computational biology and chemistry*, *56*, pp.49-60.
31. Alshamlan, H., Badr, G. and Alohali, Y., 2015. mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. *BioMed research international*, *2015*.
32. C. Wang, D. Dai, X. Li, A. Wang and X. Zhou, "SuperMIC: Analyzing Large Biological Datasets in Bioinformatics with Maximal Information Coefficient," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 4, pp. 783-795, 2017.
33. Karaboga, D. and Akay, B., 2011. A modified artificial bee colony (ABC) algorithm for constrained optimization problems. *Applied soft computing*, *11*(3), pp.3021-3031.
34. Zhu, G. and Kwong, S., 2010. Gbest-guided artificial bee colony algorithm for numerical function optimization. *Applied mathematics and computation*, *217*(7), pp.3166-3173.
35. B. Akay, and D. Karaboga, A modified artificial bee colony algorithm for real-parameter optimization, Information Sciences, 2011, doi:10.1016/j.ins.2010.07.015.