International Journal of Engineering & Technology, 7 (1.9) (2018) 186-188

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Review Paper

Cloud based computational intelligence approaches to machine learning and big data analytics: literature survey

D Venkata Siva Reddy [1] *, R. Vasanth Kumar Mehta [2]

[1] Research Scholar, Dept. of CSE, SCSVMV University, Kanchipuram- 631561Tamil Nadu
[2] Associate Professor and Head, Dept. of CSE, SCSVMV University, Kanchipuram- 631561Tamil Nadu
**Corresponding author E-mail: lionshivareddy@gmail.com*

Abstract

Today there are many sources through which we can access information from internet and based on the dependency now there is an over flow of data either in refined form or unrefined form. Handling large information is a complicated task. It has to overcome many challenges. There are some challenges like drawing useful information from undefined patterns which we can overcome by using data mining techniques but certain challenges like scalability, easy accessing of large data, time, or cost are to be handled in better sense.

Machine learning helps in learning patterns from data automatically and can be leverage this data in further predictions. Cloud computing has now turned out to be a big alternative while handling big data because cloud itself carry certain features which help in analyzing and accessing big data in proper manner. Before switching to Cloud based approaches it provides an ease of set up or testing and is economical. Thus there is a demand for cloud computing and machine learning techniques with Hadoop or Spark.

Mainly we are focusing on various works that have been done in handling big data. Here the analysis of various algorithms that are used by various researches in handling big data as well as outcome that they obtained in overcoming the challenges in handling big data.

Keywords: *Machine Learning; Cloud Computing; Big Data Analytics; Hadoop and Spark.*

1. Introduction

Machine learning is ultimate for utilizing the prospects covered in big data. It assures in digging out significant material from big as well as unrelated data resources with extremely less dependence upon individual control. It is the procedure that balance the data determined. It is appropriate in dealing the complications that arise during the process of managing unrelated data supply. It is also found to be useful in handling the massive array of inconsistent and sum of information concerned.

Machine learning does well on massive datasets. This procedure is capable of learning as well as capable of relating to the consequences in better way when extra data is added. Machine learning approaches are capable to find out and demonstrate the different prototypes hidden in the data. Labeled data for machine learning is often very difficult and expensive to obtain. Thus, when we speak about enormously growing uses of intelligence, its capability to exercise unlabeled data marks to be the major idea.

Ample data come from innumerable diverse resources such as business sales records, the collected results of scientific experiments or real time sensors. Information collected may be rare or developed using special software tools before analytics are applied. Existing data is available as structured data or unstructured data or semi-structured data or streaming data. This Voluminous data may use numerous, real-time data sources which may not be incorporated.

Here we study how to make use of available data for improving presentation on administering knowledge tasks. The primary test for big data relevance is to discover the huge quantity of data and mine useful information or knowledge for prospects in order to complete the task with large scale data in an individual background.

A practical approach is to propose parallel distributed algorithm to depend on cloud computing cluster. Thus the pattern mining methods are redesigned in map/reduce frame work. In such process, the item sets counting is mapped into the sub tasks for distributed mappers and then the semi results of the mappers are aggregated into the final counting result. As the file operations and algorithms work rely on a cluster that is outside the process control environment, these services are released above the cloud computing resources for remote access. Corresponding service operators manage to use the service in task processes. Usually data flown between these operators are large. So instead of transferring directly in the service process, temp file in the distributed file is used. The service operators who control the algorithm details always appear along with that work on the data upload and data view. The same data view is also applicable to get other kind of distributed resources.

Each Big data analytics task will compare and evaluate the data sources then submit an answer or outcome based on overloaded twisted uncertainties. Thus velocity is principally regarded among the 3 important V’s in big data for the reason that it refers to the speed at which big data should be scrutinized. Velocity is also significant as big data analysis broadened into the fields like machine learning and artificial intelligence where the systematic practice copy observations by finding and using patterns in the collected data. Attaining velocity with minimized cost is an extremely difficult task.

The unstructured form of the big data must be converted into structured form of data. The length of the characteristic space is identical to the quantity of aspects that can emerge in the data set. The length of big data analytics difficulty is higher to a large extent than the conventional problems. The big data analytics is essential to deal with massive quantity of data rapidly. The sum of data is drawing further attentions, however the element of the data and the amount of the problem also raising the problem rigidity. The benefits which we can achieve by applying the combination of Data mining techniques with Machine learning methods is more beneficial than compared to result obtained by any one technique. There are many researches that are done in to overcome the difficulties to solve big data problems.

Our purpose is to study the various previous works that have been done in this area and gather information regarding various algorithms so as to handle problems of any kind in big data. Here our concept is to study algorithms which are applicable for different type of problems which may be static or multi dimensional or dynamic or single. Big data analytics or the proposed algorithm is capable to evaluate the task of any kind.

2. Machine learning in cloud

Conventional systematic approaches are unsatisfactory to examine big data. The importance is on wide-range study of extremely scalable unstructured data confined in real time. Machine learning helps to resolve the challenges, by enabling the system to robotically study patterns from data that can be influenced in further calculations. Acquiring information can be attained by means of administered, unverified or unbreakable cleaning. Understanding big data with standard machine learning algorithms happen to be costly and at times difficult, for which cloud computing offer a convenient replacement. Given that data and software previously exist as a feature in a cloud bringing over the computation logic in a natural sequence thus turning out to be cost-effective.

3. Literature survey

We have referred a number of papers which are based on cloud Computing, machine learning and Big data. Here we are referring only few of them. There are a lot of interesting works done in handling big data analytics & machine learning. There are few papers which we found to be informative in understanding algorithms which can be applied in cloud computing.

1) Ren Gao and Juebo Wu mainly emphasizes on the significance of cloud computing. It principally stress over how load balancing facilitate in resolving a variety of troubles faced in infra structure of cloud computing. The authors explained about the advancement that has been achieved in load balancing by applying swarm intelligence algorithm which is based on local information to formulate conclusion. At this point the structure does not need a comprehensive control center. The motivation of paper is to create a load balancing mechanism which utilizes Ant Colony Optimization to balance the tasks with nodes in cloud computing. Here they have applied two dynamic balancing strategies with forward and backward ant mechanics and max-min rules [1].
2) Zaheer Khan, Ashiq Anjum, Kanran Soomro and Muhammad Atif Tahir focused on smart city administration using latest technology. Smart cities supply a chance to bond people and places by means of novel technologies that facilitate in improved city scheduling and administration. Smart cities data can be collected straight from array of sensors, smart phones, and general public. This information is incorporated with city data warehouse to execute systematic interpretation and produce information or new data for assessing better metropolitan control. Novelty in ICT presents the chance to supervise crucial information to appropriate owners for decision making [9]. This paper discussed the application of cloud based big data analytics for smart future cities. A number of considerations like data collection, preparation, semantic linking and use of appropriate data mining, machine learning or statistical analytical techniques are to be observed. This paper proposed architecture which provides basic components to build necessary functionalities for a cloud based big data analytical service for smart cities data. They developed a prototype using Map Reduce that demonstrates how cloud infra structure will analyze a sample set of Bristol open data. The model has been applied with Hadoop and Spark. Then finally they have compared the results and found that Hadoop sustain considerably over head when job are submitted to the cluster due to expensive data access operation while Spark is much faster and incurs significantly less over head. Therefore they concluded that Spark is more appropriate for chosen Bristol open data set [2].
3) Laouratou Diallo, Aisha Hassan Etal mainly focuses on Iaas clouds where users are given a virtual pool of unlimited resources to develop their applications. So users can lend resources exactly for the applications in need. This paper strongly proposes meta-heuristic scheduling algorithm which support in meeting target and lessen cost. Meta heuristic scheduling algorithm is developed for large applications in a cloud based environment and also this approach uses all features of cloud computing. Evaluated Heuristic algorithm with maximum accuracy in terms of meeting deadlines at smaller cost and also its performance on cloud simulator cloud sim is done. This paper concluded that Heuristic proposed algorithm performed well and gave better result in both traditional rule based scheduling algorithm and also heuristic scheduling algorithm in finding solutions for the work schedule as well as Hadoop map task scheduling cloud computing environments [3].
4) R. Priyanka And M. Nakkeeran explained the tasks in scheduling process and also focused their work on heuristic algorithm so as to overcome scheduling difficulties.Scheduling process in cloud is divided into 3 stages – a) Resources discovering and filtering b) Resource selection c) task allocation. Hyper heuristic aims to discover some algorithms that solve a whole range of problems robotically. Heuristic or search algorithm uses two revealing operators on impulse to define when to variant the low level heuristic algorithm and perturbation operator to finely tune the answer gained by each low level algorithm to expand the scheduling effects in expression of make span. The proposed algorithm can not only deliver better results than the traditional rule based scheduling algorithms in resolving the work flow scheduling and map task scheduling difficulties on cloud computing environments. Enhanced hyper heuristic algorithm is to influence the possessions of all the low level algorithms [4].
5) Deepa. K, Prabhu. S, Dr. N. Sengottaiyan expressed that Hyper heuristic are trouble solving methods which can be exploited to patch up the tough and non-routine issues. The three main key operations are Transition (T), Evaluation (E) and Determination (D) [TED] of heuristics has been utilized to search for the feasible solutions on the convergence system. Transition (T) makes collection of S in the solution space. Evaluation (E) measures the appropriateness of S. Determination (D) decides the subsequent search directions based on S. In cloud computing system to deal the duration of task, a high level performance Improved Hyper Heuristic Scheduling Algorithm (IHHSA) is proposed in this paper. The key idea of Hyper Heuristic Algorithm is to further boost duration time [5].
6) Shi Cheng, Yuhui Shi, Quande Qin And Ruibin Bai articulated that Big data has drawn more and more attentions. Currently most of the Big data researchers are focusing on the enormous amount of data, however managing the high dimensional data and the numerous objectives are also vital in resolving Big data problems [5].

Here the difficulty of big data analytics problem is scrutinized. Big data analytics include handling large amount of data, high dimensional data, dynamic data and multi-objective optimization. The majority of the big data problems can be represented as large scale dynamical multi-objective problems. This paper mainly focused on signifying the prospective of swarm intelligence in big data analytics. Big data involve high dimensional problems and a large amount of data. Swarm intelligence study the combined performance. It has made major success in resolving huge range, vibrant and multi -purpose tasks. In the course of the applying the swarm intelligence, further quick and efficient system can be deliberated to resolve big data analytics challenges [8].

7) P. Kowsik, K. Rajakumari have reviewed a variety of existing workflow scheduling algorithms and tabulate them on the base of character of scheduling algorithm, type of algorithm, objective conditions to which the workflow scheduling algorithm can be useful. It is obvious that set of effort has been done in the area of workflow scheduling but still there are many areas which require further attention. In this paper, a high- performance hyper-heuristic algorithm is presented to discover improved scheduling solutions for cloud computing systems. The proposed algorithm uses two detection operators to mechanically establish when to change the low level heuristic algorithm and a perturbation operator to finely tune the result obtained by each low level algorithm to further advance the scheduling consequences in terms of duration [7].

The summary of the above discussed papers is tabulating below with problem identified, algorithms used. It allows for quick and easy comparison between each result.

Paper Title	Problem Identified	Algorithm used	Brief discussion on Result
Dynamic Load Balancing Strategy for Cloud Computing With Ant Colony Optimization	Trouble faced in infrastructure of cloud computing due to dynamic load.	Ant Colony Optimization(forward &back ward) & max-min rules	Feasible and effective on load balancing in cloud computing & has better performance than a random algorithm.
Towards Cloud Based Big Data Analytics For Smart Future Cities	Basic components to build cloud based big data analytical service for smart cities data	Map Reduce (Hadoop & Spark)	Discussed the suitability of elastic nature of cloud resources to fulfill the demand of smart cities data analytic needs.
Two Objective Big Data Task Scheduling Using Swarm Intelligence In Cloud Computing	Support meeting target and reducing cost	Swarm Intelligence (Particle Swarm Optimization)	PSO works better for big data applications and reduce cost to half than ordinary scheduling algorithm.
An Enhanced Hyper Heuristics Task Scheduling In Cloud Computing	Scheduling algorithm to perform multi tasking	Enhanced Hyper Scheduling Algorithm	Can deliver better results. Resolves work flow and also map task scheduling difficulties on cloud computing.
A Hyper Heuristic Method for Scheduling The Job In Cloud Computing Environment	Job scheduling problem	Hyper Heuristic Scheduling Algorithm	Operators are used to select the low level heuristics automatically. Conditional revealing algorithm helps in finding job failures while allocating the resources.
Swarm Intelligence in Big Data Analytics	Big data analytics problem is scrutinized	Swarm Intelligence	Solve many large scale, dynamical, multi-objective problems
A Comparative Study on Various Scheduling Algorithm in Cloud Computing	Find better scheduling solution for cloud computing systems	Hyper- Heuristic Algorithm (HHSA)	HHSA can notably reduce the duration of task scheduling compared with other scheduling algorithms.

Fig. 1: Survey Report of Different Works that have been Done.

4. Further work

We are basically interested in studying the Cloud based computational intelligence approach to machine learning and big data analytics. In order to make this concept elaborated properly we are focusing on Smart Transportation System so that it will be clear in defining the problem. Based on the literature reviews we have referred, we found two algorithms as Swarm intelligence and Hyper Heuristic methods apt to our problem approach.

5. Conclusion

We are planning to propose a single perfect different system target application which works on distributed heterogeneous environment, yet to be developed. Furthermore, this system will provide fully transformative solutions and will also address naturally for the upcoming generation of applications.

References

[1] A. Abouzeid, K. B. Pawlikowski, D. J. Abadi, A. Rasin, A. Silberschatz, "HadoopDB: An ArchitecturalHybrid of MapReduce and DBMS Technologies for Analytical Workloads", PVLDB, vol. 2, no. 1, pp. 922-933, 2009.

[2] D. Agrawal, S. Das, A. E. Abbadi, "Big data and cloud computing: New wine or just new bottles?", *PVLDB*, vol. 3, no. 2, pp. 1647-1648.

[3] D. Agrawal, A. El Abbadi, S. Antony, S. Das, "Data Management Challenges in Cloud Computing Infrastructures", *DNIS*, pp. 1-10, 2010.

[4] Rajesh, M., and J. M. Gnanasekar. "Path Observation Based Physical Routing Protocol for Wireless Ad Hoc Networks." Wireless Personal Communications 97.1 (2017): 1267-1289.

[5] Chen Hsinchun, H. Roger, L. Chiang et al., "Business intelligence and analytics: from big data to big impact", *MIS Quarterly*, vol. 36, no. 4, pp. 1165-1188, December 2012.

[6] J. B. Rothnie, P. A. Bernstein, S. Fox, N. Goodman, M. Hammer, T. A. Landers, C. L. Reeve, D. W. Shipman, E. Wong, "Introduction to a System for Distributed Databases (SDD-1) ACM Trans", *Database Syst.*, vol. 5, no. 1, pp. 1-17, 1980.

[7] Venkata Narasimha Inukollu et al., "Security Issues Associated With Big Data Incloud Computing", *International Journal of Network Security & Its Applications (IJNSA)*, vol. 6, no. 3, pp. 45-56, May 2014.

[8] AiLing Duan et al., "Research and Practice of Distributed Parallel Search Algorithm on Hadoop_MapReduce", *International Conference on Control Engineering and Communication Technology 2012 IEEE*, 2012.

[9] Xiaofei Hou, Kumar T K Ashwin et al., "Dynamic Workload Balancing for Hadoop MapReduce", *IEEE Fourth International Conference on Big Data and Cloud Computing 2014 IEEE*, pp. 56-62, 2014.