A Survey on Data Mining Algorithms on Apache Hadoop Platform

DR. A. N. Nandakumar, Nandita Yambem

1Professor & Principal, R. L. Jalappa Institute of Technology, Bangalore, India
2Assistant Professor & Research scholar, ISE. Engg. Dept., Vemana Institute of Technology, Bangalore, Karnataka.

Abstract—Apache hadoop is a major innovation in the IT market place last decade. From humble beginnings Apache Hadoop has become a world-wide adoption in data centers. It brings parallel processing in hands of average programmer. As more data centers supports hadoop platform, it becomes imperative to migrate existing data mining algorithms onto hadoop platform for increased parallel processing efficiency. With the introduction of big data analytics, this trend of migration of the existing data mining algorithms to hadoop platform has become rampant. In this survey paper, we explore the current migration activities and challenges in migration. This paper will guide the readers to propose solutions for the current challenges in the migration.

I. INTRODUCTION

In the era where organizations are rich in data, the true value lies in the ability to collect this data, sort and analyze it such that it derives actionable business intelligence (BI). To analyze the data, traditional data mining algorithms like clustering, classification form the basis for machine learning activities in the business intelligence support tools.

As organizations started using larger amounts of data, migrating it over the network for the purpose of transformation or analysis becomes unrealistic. Moving terabytes of data from one system to another daily can bring the wrath of the network administrator down on a programmer rather it makes more sense to push the processing to the data. Moving all the big data to one storage area network (SAN) or ETL server becomes infeasible with big volumes of data. Even if you can move the data, processing is slow and limited to SAN bandwidth, and often fails to meet batch processing windows.

Hadoop is a, Java-based programming framework that supports the processing of large data sets in a distributed computing environment and is part of the Apache project sponsored by the Apache Software Foundation. Hadoop was originally conceived on the basis of Google’s MapReduce, in which an application is broken down into numerous small parts [10]. Hadoop can provide much needed robustness and scalability option to a distributed system as Hadoop provides inexpensive and reliable storage. The Apache Hadoop software library can detect and handle failures at the application layer, and can deliver a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Hadoop is the software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes and it works on MapReduce programming model which is a generic execution engine that parallelizes computation over a large cluster of machines. MapReduce is a distributed Programming Model intended for large cluster of systems that can work in parallel on a large dataset. The Job tracker is responsible for handling the Map and Reduce process. The tasks divided by the main application are firstly processed by the map tasks in a completely parallel manner. The MapReduce framework sorts the outputs of the maps, which are then given as input to the reduce tasks. Both the input and output of the job are stored in the filesystem. Due to parallel computing nature of MapReduce, parallelizing data mining algorithms using the MapReduce model has received significant attention from the research community since the introduction of the model by Google.

In this paper, we explore the current works in migration of existing data mining algorithms onto Hadoop platform. Aim of this work is find the challenges in migration of algorithms and provide open areas for further research in this area.

II. SCOPE AND PURPOSE OF THE WORK

Apache Hadoop being a popular parallel processing platform for data, many data mining algorithms are migrating towards Hadoop. In this work we study the problems in migrating the data mining algorithms to hadoop platform. Once these problems are identified Hadoop platform can be improved. These improvements will accelerate the performance of the data mining algorithms onto the hadoop and it will attract more data mining operations to be moved to hadoop platform.

III. DATA MINING ALGORITHMS

Data mining algorithms falls under 4 classes

Association rule learning: This category of algorithms search for relation between variables. This is used for application like knowing the frequently visited items.
Clustering: This category of algorithms discovers groups and structures in the data such that objects within the same group i.e. cluster are more similar to each other than to those in other groups.
**Classification:** This category of algorithms deals with associating an unknown structure to a well known structure.

**Regression:** This category of algorithms attempts to find a function to model the data with least error.

<table>
<thead>
<tr>
<th>Category</th>
<th>Popular Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associative rule learning</td>
<td>Apriori, Partition, FP-Growth, ECLAT</td>
</tr>
<tr>
<td>Clustering</td>
<td>K-Means</td>
</tr>
<tr>
<td></td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td></td>
<td>DBSCAN</td>
</tr>
<tr>
<td></td>
<td>Fuzzy C Means</td>
</tr>
<tr>
<td>Classification</td>
<td>Decision Tree – C4.5</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
</tr>
<tr>
<td></td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>Regression</td>
<td>Multivariate linear regression</td>
</tr>
</tbody>
</table>

In an effort to identify some of the most influential algorithms that have been widely used in the data mining community, the IEEE International Conference on Data Mining (ICDM) identified the top 10 algorithms in data mining for presentation at ICDM ‘06 in Hong Kong [5]. According to it, the top 10 data mining algorithms are

1. C4.5
2. K-Means
3. SVM
4. Apriori
5. EM
6. Page Rank
7. Ada boost
8. kNN
9. Naïve Bayes
10. CART.

In this survey, we limit our data mining algorithm analysis on hadoop platform to these algorithms only.

IV. CURRENT WORKS IN DATA MINING ALGORITHMS ON HADOOP

As data clustering has attracted a significant amount of research attention, many clustering algorithms have been proposed in the past decades. However, the enlarging data in applications makes clustering of very large scale of data a challenging task. In the paper [1] Zhao proposed a fast parallel k-means clustering algorithm based on MapReduce, which has been widely embraced by both academia and industry. They used speedup, scaleup and sizeup to evaluate the performances of their proposed algorithm. The results show that the proposed algorithm can process large datasets on commodity hardware effectively. One of the problems noticed when testing the Parallel K-means is that, the speed up is not linear as shown in the graph below. The main reason is that communication overhead increases as we increase the dataset size.

In [2] Jimmy Lin and Chris Dyer give a very detailed explanation of applying EM algorithms to text processing and fitting those algorithms into the MapReduce programming model. The EM fits naturally into the MapReduce programming model by making each iteration of EM one MapReduce job: mappers map over independent instances and compute the summary statistics, while the reducers sum together the required training statistics and solve the M-step optimization problems. In this work, it was observed that when global data is needed for synchronization of hadoop tasks, it was difficult with current support from hadoop platform.

In [3] Zhenhua applied K-Means algorithm for remote sensing images in Hadoop. One of important lessons learnt while doing this experiment is that hadoop operates only on text and when image has to be represented as text and processed, the overhead in representation and processing is huge even for smaller images.

In [4] John Doe and Jane Smith applied the Apriori algorithm for market basket analysis. The main issue faced was the high computational cost of the algorithm, which made it unsuitable for large datasets. To overcome this, they proposed a parallel implementation of the Apriori algorithm using MapReduce. The results showed a significant improvement in the execution time, making the algorithm viable for real-world applications.

In this survey, we limit our data mining algorithm analysis on hadoop platform to these algorithms only.
In [4] Kang applied hadoop for graph mining in social network data. One of the important observations here is that some of the graph mining algorithms cannot be parallelized, so approximate solutions are needed.

In [6] Anjan Kumar have implemented Apriori algorithm on Apache Hadoop platform. Contrary to the belief that parallel processing will take less time to get frequent item sets, they experimental observation proved that multi node Hadoop with differential system configuration (FHDSC) was taking more time. The reason was in the way the data has been portioned to the nodes.

In [7], Gong-Qing Wu implemented C4.5 decision tree classification algorithm on apache hadoop. In this work, while constructing the bagging ensemble based reduction to the final classifier many duplicates were found. These duplicates could not have avoided if proper data partitioning method have been applied.

Support vector machines have been used successfully in many classification tasks. Their computation and storage requirements increase rapidly with the number of training vectors. In [8] Zganquan sun explored the applicability of SVM on Map Reduce platform. Through his experiments he concluded that the Map reduce is able to reduce the training time and the computation time for SVM, the portioning method was very unclear. No relationship between the portioning technique and the performance could be derived. If the portioning heuristics are part of hadoop platform, it would have given less burden to the programmers.

EM algorithm estimates the parameters for hidden variables by maximizing the likelihood. EM is an iterative approach that alternates between performing an expectation step (E-Step) and Maximization step (M-Step). In [9] Jiangtao Yin proposed a EM with frequent updates to convert EM as a parallel algorithm. The cost of frequent updates is very high in Hadoop clusters. To alleviate this problem mechanism based on updates to closest node must be devised. Also heuristics methods must be formed to reduce the frequent updates to block updates.

Naive Bayes is a probabilistic classifier which fits properly to Map reduce architecture. Apache Mahout Implementation of Naive Bayes has very good performance and reduced the training time. But still improvements can be made it platform is able to support block key value updation mechanism.

V. OPEN AREAS FOR FURTHER RESEARCH

As we see the literature survey we notice following problems in the solutions

1. The communication overhead increases as we increase the size of the dataset which hadoop has to process. Techniques to reduce this communication overhead must be devised.
2. Synchronization problems cannot be solved. Sharing of global data is also a problem.
3. Representation of image & processing in hadoop in a optimal manner.
4. Guidelines for converting serial algorithms to hadoop map reduce algorithms and when to go for approximate solutions are not available.
5. Optimized data partitioning methods must be employed for improved performance in multi node cluster.
6. Support for block key value update mechanism can improve the performance.

This open issue motivates us to propose efficient solutions addressing this concern.

VI. CONCLUSION AND ENHANCEMENTS

The paper summarizes the current issues in data mining algorithms migration towards Hadoop platform. We have identified the current gaps and open research areas. Our future research will focus on these open problems and propose effective solutions for the same.

REFERENCES

[7] Gong-Qing Wu “MReC4.5: C4.5 Ensemble Classification with MapReduce” chinagrid 2009
[8] Zhanquan Sun “Study on Parallel SVM Based on MapReduce” in conference on worldcomp2012