Survey on Data Mining Algorithm and Its Application in Healthcare Sector Using Hadoop Platform

K. Sharmila¹, Dr. S. A.Vethamanickam²

¹Asst Prof. & Research Scholar, Dept. of Computer Science, Vels University, Chennai, India.  
²Research Advisor, Chennai, India.

Abstract- In this survey paper, we have scrutinized and revealed the benefits of Hadoop in the Healthcare sector using data mining where the data flow was in massive volume. In developing countries like India with huge population, there exists various problems in the field of healthcare with respect to the expenses met by the economically underprivileged people, access to the hospitals and research in the field of medicine for Big Data. The Apache Hadoop has become a world-wide adoption and it has brought parallel processing in the hands of average programmer for Big data. It has become imperative to migrate existing data mining algorithms onto Hadoop platform for increased parallel processing efficiency. In this paper, we have surveyed various progress made in the area of data mining technique, its latest adoption in Hadoop platform and Big data, algorithms used in such platform, and listed out the open challenges in using such algorithm in the Indian medicinal data set.

Keywords: Hadoop, Data mining, Healthcare, Big data.

I. INTRODUCTION

In this era of Big data the organizations and health industry are facing problems of three V's namely Volume, Velocity and Variety in migrating the data over the network for the purpose of transformation or analysis that has become unrealistic. Moving terabytes of data from one system to another often has brought the network administrator infeasible and made the process slow and limited to SAN (Storage Area Network) bandwidth.

The distributed processing of huge data sets across groups of systems is facilitated by using a computing model of the Apache Hadoop Framework. The framework was projected to widen from solitary server to thousands of systems for the computation and storage. This forceful feature of Hadoop framework attracts variety of companies and organizations to use it for both research and production.

Healthcare is one of the most important areas of developing and developed countries to ease the priceless human resource. Commonwealth Governments have identified various health issues like diabetes as a significant and growing global public health problem. Estimation shows 40 million Indians suffer from diabetes, and the crisis seems to be growing at a shocking rate. By 2020, the number is expected to twice, even though half the numbers of diabetics in India remain undiagnosed due to the massive volume of data.

Since healthcare industry nowadays has flooded with massive amount of data, need validation and accurate analysis. Even though Big Data Analytics and Hadoop can contribute a major role in processing and analyzing the healthcare data in variety of forms to deliver suitable applications and in turn reduces the cost of services to a common man in the country, there are some open challenges to conquer which are explicitly stated in this survey paper.

II. DATA MINING

Data mining is the core step, which has resulted in the discovery of hidden but useful knowledge from massive databases. “It is the non-trivial extraction of previously unknown and useful information about data”. It can also be defined as “the science of extracting useful information from large databases”. The two primary goals of data mining are prediction and description.

- **Prediction** involves some variables or fields in the data set to predict unknown or future values of other variables of interest.
- **Description** focuses on finding patterns describing the data that can be interpreted by humans.

Categorization of Data Mining Techniques:

2.1 Data mining algorithms falls under 4 classes

a. Association rule learning:

This category of algorithms search is for relation between variables. This is used for application like knowing the frequently visited items.
b. 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.

c. Classification:

This category of algorithms deals with associating an unknown structure to a well known structure.

d. Regression:

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

III. CURRENT WORKS IN DATA MINING ALGORITHMS ON HADOOP

<table>
<thead>
<tr>
<th>S.No</th>
<th>Author</th>
<th>Name of the Algorithm used</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fayyad et.al (1996)</td>
<td>Data Mining techniques</td>
</tr>
<tr>
<td>2</td>
<td>Zhao et.al (2009)</td>
<td>k-means clustering</td>
</tr>
<tr>
<td>3</td>
<td>Gong-Qing Wu (2009)</td>
<td>C4.5 decision tree</td>
</tr>
<tr>
<td>6</td>
<td>Kang and Christos Faloutsos (2012)</td>
<td>Graph mining algorithms</td>
</tr>
<tr>
<td>7</td>
<td>Anjan K Koundinya etal (2012)</td>
<td>Apriori algorithm</td>
</tr>
<tr>
<td>9</td>
<td>Jiangtao Yin (2012)</td>
<td>EM</td>
</tr>
<tr>
<td>10</td>
<td>Nandakumar and Nandita Yambem (2014)</td>
<td>Existing Algorithm</td>
</tr>
<tr>
<td>11</td>
<td>Ranshul Chaudhary et.al (2014)</td>
<td>k-means, Hierarchical clustering, COBWEB and DBSCAN</td>
</tr>
</tbody>
</table>

1. Fayyad et.al(1996) describes the various data mining techniques that allow extracting unknown relationships among the data items from large data collection that are useful for decision making.

2. A fast parallel k-means clustering algorithm was proposed by Zhao etal(2009) based on MapReduce, which has been widely embraced by both academia and industry. 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 speedup is not linear. The main reason is that communication overhead increases as we increase the dataset size.

3. C4.5 decision tree classification algorithm was implemented on Apache hadoop by Gong-Qing Wu(2009). In their work, while constructing the bagging ensemble based reduction to construct the final classifier many duplicates were found. These duplicates could not have avoided if proper data partitioning method have been applied.

4. A very detailed explanation of applying EM algorithms to text processing and fitting those algorithms into the MapReduce programming model was proposed by Jimmy Lin and Chris Dyer (2010). The EM fits naturally into the MapReduce programming model by making each iteration of EM one Map Reduce job. 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.

5. K-Means algorithm for remote sensing images in Hadoop was applied by Zhenhua(2010). One of the important issues learnt while doing their 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.

6. Kang and Christos Faloutsos(2012) 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.

7. Apriori algorithm has been implemented on Apache Hadoop platform by Anjan K Koundinya etal(2012). 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 partitioned to the nodes.

8. Zganquan sun (2012) explored the applicability of SVM (Support Vector Machine) 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.

9. EM is an iterative approach that alternates between performing a expectation step (E-Step) and Maximization step (M-Step). An EM with frequent updates to convert EM as a parallel algorithm which has been proposed by Jiangtao Yin (2012) shows that the cost of frequent updates was very high in Hadoop clusters. To alleviate this problem, new mechanism must be explored to reduce the frequent updates to block updates. Apache Mahout Implementation of Naive Bayes has very good performance and reduced the training time but still improvements can be made the platform to support block key value updating mechanism.

10. Nandakumar and Nandita Yambem (2014) had described that Apache Hadoop has become a worldwide adoption in data centers and hence 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 widespread. In this paper, they explored the current migration activities and challenges in migration.

11. Ranshul Chaudhary et al. (2014) surveyed that some well known algorithms concerned with data mining under the clustering techniques namely: k-means, Hierarchical clustering, COBWEB and DBSCAN algorithms are studied. The results are compared and analyzed in accordance to their efficiencies.

IV. HADOOP

Hadoop is a Java-based programming framework which 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 Map Reduce, in which an application is broken down into numerous small parts. The Apache Hadoop software library can detect and handle failures at the application layer. The Hadoop mainly includes:

1) Hadoop Distributed File System (HDFS)
2) Hadoop Map Reduce.

4.1 HDFS:

The HDFS has some desired options for enormous information parallel processing, such as: (1) work in commodity clusters in case of hardware failures, (2) access with streamed information, (3) dealing with big data set (4) uses an easy coherency model, and (5) moveable across various hardware and software platforms. The HDFS is designed as master/slave architecture.

As shown in figure HDFS cluster consist of a single node known as a Name Node, which manages the file system namespace and regulates client access to files. Data Node store data as blocks within files. Name Node is responsible for mapping of data blocks to Data Node. Also Name Node manages file system operations like opening, closing, renaming files and directories.

The Name Node information must be preserved even after the Name Node machine fails. There are multiple copies of data on Name Node is maintained on number of machines. So that in case of crashes of Name Node these nodes can be used by other nodes in cluster. These nodes are called as secondary Name Node.

4.2 Hadoop Map Reduce:

Map Reduce was originally proposed by Google to handle large scale web search applications. This approach has been proved to be an effective programming approach for developing machine learning, data mining and search applications in data centers.

The Hadoop Ecosystem consists of the following components namely HBase, Hive, Pig, Zookeeper, Sqoop, and Flume.

HBase: Distributed column-based database

Hive: Distributed Data Warehouse, provides SQL-based Query language

Pig: Data Flow Language and execution environment

Zookeeper: Allows distributed processing of data in large systems to synchronize with each other in order to provide consistent data to client requests.

Sqoop: Tool designed for efficiently transferring bulk data between Hadoop and structured data stores such as relational databases.

Flume: A distributed service for collecting, aggregating, and moving large amount of log data.
V. CURRENT APPLICATIONS OF DIABETIC DATA SET USING DATA MINING

1. Tamilarasi and Sapna (2008) has described research on complex diseases only seems to be approaching the final goal, the prevention and cure of the diseases, very slowly.

2. Huy Nguyen Anh Pham and Evangelos Triantaphyllou predicts whether a new patient would test positive for diabetes using a new approach, called the Homogeneity-Based Algorithm (or HBA) on the dataset (Pima Indian diabetes data set) and the author concluded that it is very important both for accurately predicting diabetes and also for the data mining community, in general.

3. Parthiban et.al (2011) tried to predict the chances of getting a heart disease using attributes from diabetic’s diagnosis. This can be extended to predict other type of ailments which arise from diabetes, such as visual impairment in future. Further, the data analysis results can be used for further research in enhancing the accuracy of the prediction system in future.

4. Karthikeyani et.al (2012) concluded that there are different data mining classification techniques can be used for the identification and prevention of diabetes disease among patients. Among ten classification techniques in data mining to predict diabetes disease in patients the CS-CRT algorithm is best.

5. Several intelligent classifiers such as Bayesian, Functional, Rule-base, Decision Trees and Ensemble for diabetes mellitus diagnosis on PID dataset has been applied by Najmeh Hosseinpour et.al (2012) and therefore, they concluded that, Bagging with logistic core had the best performance.

6. Parthiban and Srivatsa (2012) have tried to predict the chances of heart disease using attributes from diabetic’s diagnosis and we have shown that it is possible to diagnose heart disease vulnerability in diabetic patients with reasonable accuracy. Hence SVM model can be recommended for the classification of the diabetic dataset.

VI. CURRENT APPLICATIONS OF DIABETIC DATA SET USING HADOOP

<table>
<thead>
<tr>
<th>S.No</th>
<th>Author</th>
<th>Feature of Algorithm Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sadhana and Savitha Shetty (2014)</td>
<td>Implementation of hive and R.</td>
</tr>
<tr>
<td>2</td>
<td>Peter Augustine(2014)</td>
<td>Consideration of clinical facts</td>
</tr>
</tbody>
</table>

1. Sadhana and Savitha Shetty (2014) have proposed a detailed analysis of the diabetic data set which was carried out efficiently with the help of hive and R. The facts which were revealed during the process can be used for developing some prediction models. The information which was revealed can be further used to develop efficient prediction models.

2. Achieving better outcomes at lower costs has become very important for healthcare, and Big Data Analytics and Hadoop’s presence are positively part of the solution in reaching the goal has been concluded by Peter Augustine (2014). Although we are in the early days of healthcare big data, it is clear those strategies for big data in the integration of R&D data, efficient clinical trials, and finally in clinical outcomes is foundational to building that solution, lowering costs, and enhances the accessibility and availability of healthcare to all in 1.2 billions of Indians.

3. Wullianallur Raghupathi and Viju Raghupathi (2014) have described that big data analytics and applications in healthcare are at a promising stage of development, but rapid advances in platforms and tools can accelerate their maturing process.

VII. OPEN CHALLENGES

During the survey of all works in respective areas, we had come across many open challenges and problems in their respective work. We are listing down some of the significant problem, which are pertinent to our work.

1. Migrating existing data mining algorithm onto Hadoop platform to increase the parallel processing efficiency is a direct challenge.[1]

2. Synchronization problems cannot be solved. Sharing of global data is also a problem.[7]
3. 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. [8].

4. Though Zganquan sun 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 fewer burdens to the programmers. [12]

5. 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 [13].

The above challenges have to be overcome by applying assorted techniques and algorithm in data mining the above challenges can be overcome.

VIII. CONCLUSION

This paper presents an overview of various data mining techniques and application of diabetic dataset using Hadoop platform. This helps us to acquire knowledge about how Hadoop can be implemented to predict the diabetics and related diseases to it for huge dataset using data mining techniques. As per the references considered above, we have reached to the conclusion that, the existing platform has implemented the data set with immature development. Thereby we are planning to carry out the further research in Hadoop platform with own clinical data set.

REFERENCES


[12] Zhanquan Sun —Study on Parallel SVM Based on MapReducel in conference on worldcomp2012


