Machine Learning and Statistical Approaches for Big Data: Issues and Challenges

Omkar Shelte

Department of Information Technology (Msc.IT Part 2)

MSC IT Department, Model College (Dombivli), University Of Mumbai, India

Abstract

Today, as we are observing, massive sized and complex structured data is becoming available from variety of diverse sources, organizations are making attempt to utilize these plentiful resources for the purpose of enhance innovation, increase decisional and operational efficiency. Machine learning is a kind of artificial intelligence method to discover knowledge for making intelligent decisions. Big Data has vast impacts on scientific discoveries and value creation. This paper presents an extensive literature study and review of latest advances, developments and new methodologies in researches on machine learning for processing big data. We have discussed various types of data types, learning methods, vital issues in big data processing and application of machine learning approaches in big data. Finally, we have outlined some open problems in this domain and our further research aims and directions. Keywords: Machine learning, Data mining, Big data, Data analysis, Distributed computing, Knowledge discovery.

Keywords: Machine learning, Data mining, Big data, Data analysis, Distributed computing, Knowledge discovery.

INTRODUCTION

Big data are expanding in a rapid manner in all engineering disciplines and science domains. Volume of data explodes at high rate today as a result in advancements of ”Web technologies, social media, and mobile devices”. For eg., Twitter use to process over 70 million tweets daily, through producing over 8TB in daily manne. According to one research estimation, there will around 30 billion computing machines, connecting each other, by 2020. Big Data employs amazing potential for trade value in diverse fields like – “health sector, biology, medicine transportation, online advertising and financial services”. Though, traditional strategies struggles when deal with this large data. Learning from massively large data brings significant opportunities for numerous sectors. Still, most of these routines are not much practical or scalable enough. Therefore, ML demands to deeply discover itself for processing big data. According to a study by Oracle Company, around 90% of the world’s knowledge data is held in unstructured form. Big data may be explained in terms of three traits - velocity, volume and variety. Variety meant for heterogeneous nature, Velocity meant for the frequency at which data is being captured, and Volume meant for size of data (PB, EB and TB). Machine learning algorithms categorize the learning task in two types i.e. Supervised learning and Unsupervised learning. Mining of big data and knowledge discovery is the process of an efficient extraction of implicit, relevant, previously unknown, potentially useful (rules, regularities, patterns, constraints) from incomplete, noisy, random and unstructured data in large web databases.

TYPES OF LEARNING METHODS

This subsection presents some recent learning methods that may play vital role in solving the big data problems.

1) Kernel-based learning:
   Kernel-based learning is proven to be very dominant methodology to efficiently enhance the computational capacity. The notable advantage of this method is that both linear as well as non-linear vector kernel functional methods are present to deal with the non-linearity of data in N-dimensional feature space.
2) Depiction based learning:
   This kind of learning, is a solution to study valuable representations of the raw data. It is comparatively simpler to get knowledge information while processing through classifiers . Some variants of representational learning are evolved in past years.

3) Active learning:
   This learning chooses a subset of an unstructured and critical occurrence for purpose of labeling . The active learner obtains larger accuracy using reduced number of

occurrences.

4) Deep learning:
These designs take more complicated, compartmented statistical patterns of inputs and manages to be robust for new fields as compare to traditional learning systems. “Deep belief networks (DBNs)” and” convolutional neural networks (CNNs)” are two deep learning methodologies.
5) Transfer learning:
The prime intention of transfer learning is to derive knowledge features from input source and later implement the knowledge to the target task. The main benefit is that it can efficiently apply knowledge, which has been learned previously in order to find solution for new problems in fast manner.
6) Parallel & Distributed learning:
The data which is avail-able in incomplete, inconsistent and unstructured format, is first pre-processed, then cluster forming is done. Count of such distributed clusters is performed. Further one processing thread is assigned to each cluster in order to perform multi-threading in parallel and distributed manner.
Vital issues of machine learning for Big data This section presents a review about the critical concerns of machine learning procedures for big data from diverse viewpoints, an overall scenario is presented. It includes

Vital issues of machine learning for Big data This is presented. It includes
(i) learning for massive scaled data, (ii) learning for diverse structured data, (iii) learning for high frequency streamed data, (iv) learning for imprecise and incomplete data, (v) learning for deriving valuable knowledge from massive sized volumes of data.

1. Learning for massive sized data:
Considering only digital information, every day, Google processes approx. 24 PB data. Under modern development courses, data analyzed by big companies will unquestionably cross this petabyte magnitude. We are presently swimming in a deep and expanding ocean of data which is too bulky to train ML algorithm. Though, distributed and parallel frameworks are preferred. Cloud computing and MapReduce-assisted learning methods are another progress aspects which deal with core challenges of big data. It can improve computing and storage capacity through cloud infrastructure.

2. Learning for different structures of data: Immense variety of data is another aspect, that addresses big data interesting as well as challenging. It resulted of the aspect that data usually collected from diverse sources and are of varying types. Structured, semi-structured or fully unstructured data sources stimulate formation of heterogeneous, high-dimensionality, and nonlinear data e.g. global environment patterns, astronomical spectra, and human gene patterns with varying representation patterns.

3. Learning for high speed of streaming data: Speed or velocity really matters in big data scenario. In the timesensitive cases like earthquake prediction, stock market prediction etc., the inherent value of data is depending upon factor of data freshness which requires to be treated in a real-time fashion. Other challenging problem connected with high velocity is that data usually are not stationary, which requires learning procedures to determine the data as stream.

MOTIVATION TO THE PROBLEM
As more scenarios e.g. global economy, society administration, national security involves Big Data problems, traditional strategies struggles when deal with this large data. Learning from massively large data brings significant opportunities for numerous sectors. Still, most of these routines are not much practical or scalable enough . From massive amount of available data, fetching (deriving) structured, useful and learning(ML) techniques are lacking computational efficiency, practicality or scalability to handle the data with traits of massive volume, varying types, great speed, uncertainty, inconsistency and incompleteness. So, to discover more optimal techniques which can process huge sized unstructured data efficiently are much desired.

Data Mining and Machine Learning Techniques
This section summarizes the mathematical, statistical techniques that are very useful while performing data mining or machine learning.

Hadoop HDFS
HDFS is Hadoop’s storage layer which provides the high availability, fault tolerance and reliability. It is probable that worlds 75% of data will be stored in Hadoop HDFS by the end of 2017. Apache Hadoop HDFS is a kind of distributed file system(DFS) which affords redundant storage space for caching files which are enormous in sizes; files which are in the range of TB and PB. Files are split into blocks and diffused across junctions in a cluster. After that each block is replicated. Hence suppose a machine goes down or goes crashed, then in that cases, also we can effortlessly retrieve and access our data from different devices. Hence it is extremely fault-tolerant. HDFS

gives faster file read and writes mechanism, as data is saved in different nodes inside a cluster.

Artificial Neural Network

It is a kind of classifier, whose model design structure and functionality is somewhat similar to human brain structure algorithmic model. For classification problem, the specific structure of neural network changes. First, the training is carried out for ANN, where the topology and number of network nodes present in the hidden layer are decided. Unlike SVM, there is no phenomenon i.e. ndimensional planes and hyperplanes. Still, training of data sets process here is time taking, produces less accurate and efficient results also.

SUPPORT VECTOR REGRESSION

As we know that the classification procedure falls into one of the category, either supervised or unsupervised classification. So, in the area of machine learning, support vector networks are supervised machine learning models. They are aimed for learning and training procedures for the data used in regression analysis and non-linear hyperplanes for separation task in classification. Some parameters like gaussian kernels, standard deviation and variance of data, kernel functions are some significant parameters which affect the performance of SVM.

Fuzzy SVM :

In FSVM, each training point belongs exactly to no more than one particular class. Some points having noise and that could not have classified by SVM, are dealt here through FSVM. Pre-knowledge information about data sets is needed, like - stochastic and probabilistic information. Here, several stochastic correlations can be identified.

Bayesian Classifiers:

In these type of classifiers, the statistical information and probabilistic knowledge is employed for metadata creation. Here, Bayes' theorem is utilized with naive independence assumptions among features. Since 1950's, it is being continuously explored. This is having applications in medical diagnosis analytics, spatial imaging data, text categorization etc. This classifier is highly scalable and it requires a number of parameters which are linear in no. of variable predictors in aspects of learning problem.

BIG DATA AND ANALYSIS

This section presents overall idea of big data sets and tools that are used for big data analysis.

BIG DATA: SCENARIO

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy.

Machine learning application to Big data

Machine learning [5] is ideal for exploiting the opportunities hidden in big data. It delivers on the promise of extracting value from big and disparate data sources with far less reliance on human direction.

It is data driven and runs at machine scale. It is well suited. And unlike traditional analysis, machine learning thrives on growing datasets. The more data fed into a machine learning system, the more it can learn and apply the results to higher quality insights.

Big data management tools

The entire data analytics industry nowadays has a buzzword, "big data," concerning how we're operating something with the enormous amount of information gathering up. "Big data" is replacing "business intelligence". To handle this massive amount of data available, we have listed out some significant tools that can be utilized to process big data.

"Pentaho Business Analytics"

It is a kind of software program that started as an engine, branching within big data by creating it simpler to absorb the information from the different sources. One can experiment with Pentaho's tool to many of the most popular NoSQL databases, they are - MongoDB Cassandrav etc. One can drag furthermore drop the columns into aspects and reports as if the information issued from the SQL databases, once the databases are connected.

"Karmasphere Studio and Analyst"

It is kind of a specialized IDE that makes it simpler to create and run Hadoop jobs. This produces something better: As we set up the workflow, the tool engine displays the status of the test data at each and every step.

"Talend Open Studio"

This tool gives an Eclipse-based Integrated Development Environment for stringing data processing operations collectively with Hadoop. Its tools are intended to help with data integration, data quality along with the data management.

"Skytree Server"

Skytree allows a bundle that delivers many extra advanced ML procedures. All it needs is typewriting the right command in command line. It is more

focused on the guts than the shiny GUI. Skytree Server is optimized to execute a no. of classical ML algorithms. It thought of as ten thousand time faster than different packages. It can explore through the data looking for clusters of similar objects, then rearrange this.

"Splunk"
It is a little distinctive from the other tools. It creates an index of the data as if the data were a part or a block of text. This approach is much alike to a text search method. Splunk will choose text strings and search around in the index. Its variant tool Shep guarantees bidirectional union of Hadoop and Splunk, enabling to interchange data within the systems and query Splunk data of Hadoop.

"Jaspersoft BI Suite"
It is one of the open source tool for mainly producing reports from database columns. The software tool is well-polished and already installed in many businesses turning SQL tables into PDFs that everyone can scrutinize at meetings. Jaspersoft is not specifically offering unique ways to look at the data, just more complicated ways to access and to locate data stored in the new locations.

Problem Identification :
With the beginning of span of Big Data, which may be considered as the next bound for modernization, competition and potency, a new boom of revolution is nearly about to onset. The volume of data today, is raging at an unusual rate as a result of advancements and developments in Web technologies, social media, and mobile devices etc. Traditional strategies are hardly suffering when faced with this massive sized data. These traditional machine learning(ML) routines and, varying types, great speed, uncertainty and incompleteness.
Based on the precious knowledge, we need to create new techniques and methods to excavate big data. Machine Learning demands to deeply discover itself for processing big data, so that the knowledge extraction and reasoning for uncertain concepts from unstructured and huge sized data can be done in a computationally efficient manner.

FUTURE RESEARCH DIRECTIONS :
The aim of our research is to develop new efficient methods for the analysis of big data sets. Our future research directions are as follows: -
We will contribute some optimal and computationally efficient big data analytics techniques to analyze different type of data sets. This may be achieved by selecting strategies of Rough set theory and Fuzzy logic evolved as an efficient machine learning methodology, which has become an important tool to perform data analytics.
As, today, processing of massive sized unstructured, inconsistent, incomplete and imprecise data by computing machines is a challenging task. In recent past years, Rough set theory and Fuzzy logic evolved as an efficient machine learning methodology, which has become an important tool to perform big data analytics. To perform operations in the data, present in higher dimensions may be more computationally. We will employ these modern machine learning techniques to process big data, which also gives the guarantee of dimensionality reduction and other parameters selection of data sets.

CONCLUSION
Big data analytics is the process of examining large and varied data sets. Learning from massively large and unstructured data brings significant opportunities for numerous sectors. Still, most of these routines are not much computationally efficient, practical or scalable enough. This paper discusses the need for the research that aims at proposing new techniques that can be used for analysis of big data. However, most of the traditional AI involved methods are not scalable to manage data with the properties of its huge volume, diverse types, inconsistency, uncertainty along with incompleteness. In response, there is a need for machine learning to revitalize itself for big data processing.
This paper started with various types of learning methods. Further it discusses about some of the significant and practical issues of machine learning for big data analytics. Later, we have listed out some tools which can be employed for big data management and analysis.

REFERENCES
[1] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu and Shuo Feng. "A survey of machine learning for big data processing", EURASIP Journal on Advances in Signal Processing (2016) 2016:67, Springer.

[2] Philip Russom. "Big data analytics", TDWI research, Fourth quarter (2011).

[3] Dunren Che, Mejdl Safran, Zhiyong Peng. "From Big Data to Big Data Mining: Challenges, Issues, and Opportunities", DASFAA Workshops, LNCS 7827, pp. 1-15, Springer-Verlag Berlin Heidelberg (2013).

[4] Joseph McKendrick. "Big Data, Big Challenges, Big Opportunities: IOUG Big Data Strategies Survey", Unisphere Research, ORACLE, September (2012).

[5] Z. Pawlak. "Information Systems Theoretical

Foundations", Information Systems, Vol. 6, No. 3, pp. 205-218, (1981).

[6] Changwon Yoo, Luis Ramirez, Juan Liuzzi. "Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine", Int. Neurology Journal (2014); 18:50-57.

[7] Alexandra L'heureux, Katarina Grolinger, Hany F. Elyamany and Miriam A. M. Capretz. "Machine Learning With Big Data: Challenges and Approaches", IEEE ACCESS, Vol. 5, June 7, (2017).