Journal of Engineering Sciences

Vol11,Issue2,FEB/2020
ISSN NO:0377-9254

MACHINE LEARNING ALGORITHMS PROCESSING IN BIG DATA ANALYTICS

K. VAMSHEE KRISHNA[1], DR. X. S ASHA SHINY[2], H NAGA CHANDRIKA[3]

1. **Research Scholar,** Department of Computer Science and Engineering, Dr. A.P.J Abdul Kalam University ,Indore, M.P India, vamshik825@gmail.com

2. Department of Computer Science and Engineering, Nalla Malla Reddy Engineering College, Ghatkesar, Hyderabad, India.

3. Research Scholar, Department of Computer Science and Engineering, Dr. A.P.J Abdul Kalam University ,Indore, M.P India.

Abstract:

Big data refers to size of data produced not only in terabytes but in Exabyte or even beyond this and it is a large data combination of both structured and unstructured data that it is very difficult to process usingtraditional database and mostsoftware techniques . It envelope the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered. The developments and new methodologies in researches on machine learning for processing big data. It has a learning methods and various types of data types, basic issues in big data processing and application of machine learning approaches in big data. Finally, we processing algorithms in this domain and our further research aims and directions.

The data being produced is structured, unstructured or even semi- structured. It has become difficult to find meaningful patterns hidden in these various data sets. The data is being produced at lightning speed. In this review paper, various machine learning algorithm have been reviewed for Big Bata processing.

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

INTRODUCTION

As the name says ”Big Data” means the data is uncountable and is so massive that it can go beyond the size of hundreds of terabytes. Big data accounts structured as well as unstructured data. Big data is used in fields like science, engineering and technology. The size of big data is increasing rapidly at each passing second. Traditional database system has no capacity to handle such huge data. The world is creating 2.8 quintillion of data per day from unstructured data sources like social media platform like Facebook, Instagram, gmail, hike, photos, files, etc. Big data forms the three level structures of data produced that is structured data, unstructured data and semi-structured data.

In unstructured data, the elements within the data have no defined structure. The Big data Commission at the Tech American Foundation offers the following definition: “Big data is a term that describes large volume of high-velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.” Big Data is basically collection of heterogeneous data from various pool of data resources. These data carry hidden pattern which can give fruitful results if recognized. Various Machine learning techniques have given optimized results. Pattern recognition, K-means algorithm, clustering, line regression, Bayesian algorithm, Decision tree, neural network are the machine learning algorithm.

The vital key features which define characteristics of Big data are as follows:-

DATA VOLUME: It is the amount of data that is available in terabytes or more. The data comes from social sites, media, post etc. With the ever increasing amount of data volume, it needs to be stored and analyzed and has to come up with optimal solution to solve its storage problem.

DATA VELOCITY: It is the amount of data that is produced at great pace. Data is increasing at every second with lightning speed and this makes time an important factor in several organization.

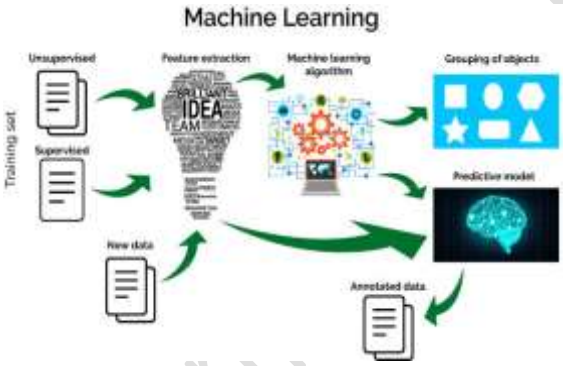
DATA VALUE: Big data comprises of heterogeneous collection of data and each data has value associated

with it which varies economically. Finding hidden patterns and meaningful data from the bag of wide-ranging data and transforming it for analysis is a head to toe task.

DATA VARIETY: It is the heterogeneous trait of data that makes it complex, variable data coming from internet, e-mails, videos etc.

1. Machine Learning

Machine Learning is a unique approach that yields answers to the problem of Big Data. Machine Learning is a field of computer science that is associated with Artificial intelligence, science, engineering and technology, bio-medical etc. Machine learning finds its root in data prediction and pattern evaluation. Various Machine Learning techniques stand in solving various problems of big data. Traditional machine learning systems were designed such that all the data would be loaded at one time and it would be processed after that. With the increasing amount of data at great speed, it has become difficult to use traditional data base systems.



1.1 SUPERVISED LEARNING: As the name suggest, it is supervised under a guider means trained data sets are readily available which consists of set of data inputs and desired outputs. Given rules are applied to the data sets to get outputs.

1.2 UNSUPERVISED LEARNING: St
ands opposite to supervised learning, no trained data sets are available here. The outputs are totally based on prediction analysis. It helps to explain hidden structure from untrained data sets.

1.3 REINFORCEMENT LEARNING: It is based on the feedback provided by the environment. It is a decision making approach. The state of a game cannot be decided it is completed and the feedback is obtained in the form of won or lost.

2. Machine Learning Techniques

2.1Supervised Machine Learning

The majority of practical machine learning uses supervised learning.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

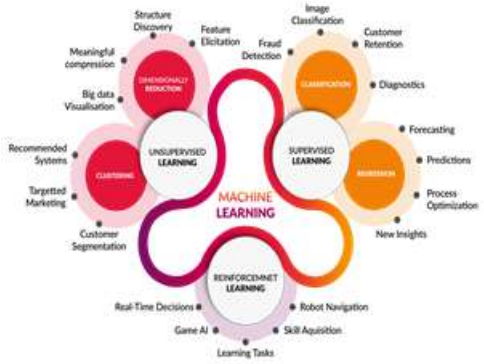


Figure: Machine Learning

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance Supervised learning problems can be further grouped into regression and classification problems.

Classification: A classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease".
Regression: A regression problem is when the output variable is a real value, such as "dollars" or "weight".
Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Some popular examples of supervised machine learning algorithms are:

- Linear regression for regression problems.
- Random forest for classification and regression problems.
- Support vector machines for classification problems.

1.2 Unsupervised Machine Learning

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.
Some popular examples of unsupervised learning algorithms are:

k-means for clustering problems.

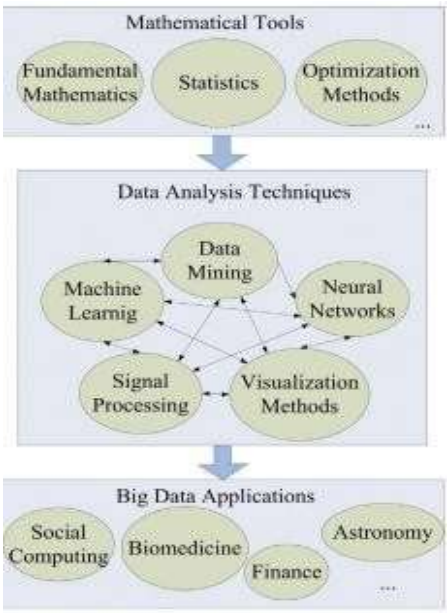
- Apriori algorithm for association rule learning problems.

1.3 Semi-Supervised Machine Learning

Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems.

These problems sit in between both supervised and unsupervised learning.

A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled.



Many real world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store.

You can use unsupervised learning techniques to discover and learn the structure in the input variables.

You can also use supervised learning techniques to make best guess predictions for the unlabeled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.

In this post you learned the difference between supervised, unsupervised and semi-supervised learning. You now know that:

Supervised: All data is labeled and the algorithms learn to predict the output from the input data.
Unsupervised: All data is unlabeled and the algorithms learn to inherent structure from the input data.
Semi-supervised: Some data is labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used.
Do you have any questions about supervised, unsupervised or semi-supervised learning? Leave a comment and ask your question and I will do my best to answer it.

1.4 What is Reinforcement Learning?

Reinforcement Learning is defined as a Machine Learning method that is concerned with how software agents should take actions in an environment. Reinforcement Learning is a part of the deep learning method that helps you to maximize some portion of the cumulative reward.

This neural network learning method helps you to learn how to attain a complex objective or maximize a specific.

Here are some important terms used in Reinforcement AI:

- **Agent:** It is an assumed entity which performs actions in an environment to gain some reward.
- **Environment (e):** A scenario that an agent has to face.
- **Reward (R):** An immediate return given to an agent when he or she performs specific action or task.
- **State (s):** State refers to the current situation returned by the environment.
- **Policy (π):** It is a strategy which applies by the agent to decide the next action based on the current state.
- **Value (V):** It is expected long-term return with discount, as compared to the short-term reward.
- **Value Function:** It specifies the value of a state that is the total amount of reward. It is an agent which should be expected beginning from that state.
- **Model of the environment:** This mimics the behavior of the environment. It helps you to make inferences to be made and also determine how the environment will behave.
- **Model based methods:** It is a method for solving reinforcement learning problems which use model-based methods.
- **Q value or action value (Q):** Q value is quite similar to value. The only difference between the two is that it takes an additional parameter as a current action.

1.4.1 Reinforcement Learning Algorithms

There are three approaches to implement a Reinforcement Learning algorithm.

a) Value-Based:

In a value-based Reinforcement Learning method, you should try to maximize a value function **V(s)**. In this method, the agent is expecting a long-term return of the current states under policy **π**.

b) Policy-based:

In a policy-based RL method, you try to come up with such a policy that the action performed in every state helps you to gain maximum reward in the future.

Two types of policy-based methods are:

- Deterministic: For any state, the same action is produced by the policy π.
- Stochastic: Every action has a certain probability, which is determined by the following equation. Stochastic Policy :

n{a\s) = P\A, = a\S, =S]

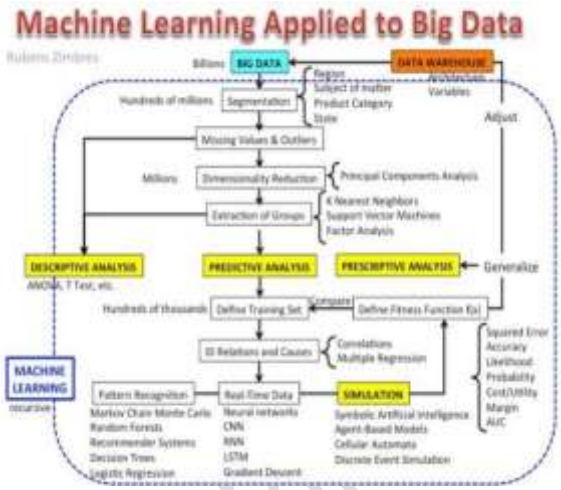
c) Model-Based:

In this Reinforcement Learning method, you need to create a virtual model for each environment. The agent learns to perform in that specific environment.

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" [2] [4]. For eg., Twitter use to process over 70 million tweets daily, through producing over 8TB in daily manner [50]. According to one research estimation, there will around 30 billion computing machines, connecting each other, by 2020 [51]. Big Data employs amazing potential for trade value in diverse fields like – "health sector, biology, medicine transportation, online advertising and financial services" [47] [52] [20]. 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 [39] [48]. 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. [11] [17] [22] 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 [6] [41] 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. The general process is represented as diagram below: -



2. 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 [39]. 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 [59], is a solution to study valuable representations of the raw data. It is comparatively simpler to get knowledge information while processing through classifiers [60]. Some variants of representational learning [61] [60] [62] are evolved in past years.

3) Active learning: This learning chooses a subset of an unstructured and critical occurrence for purpose of labeling [67]. 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)" [63] [64] 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 [66]. 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 [65]. 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.

4. Analyze Big Data

Since data is not always moved during the organization phase, the analysis may also be done in a distributed environment, where some data will stay where it was originally stored and be transparently accessed from a data warehouse. The infrastructure required for analyzing big data must be able to support deeper analytics such as statistical analysis and data mining, on a wider variety of data types stored in diverse systems; scale to extreme data volumes; deliver faster response times driven by changes in behavior; and automate decisions based on analytical models. Most importantly, the infrastructure must be able to integrate analysis on the combination of big data and traditional enterprise data. New insight comes not just from analyzing new data, but from analyzing it within the context of the old to provide new perspectives on old problems. For example, analyzing inventory data from a smart vending machine in combination with the events calendar for the venue in which the vending machine is located, will dictate the optimal product mix and replenishment schedule for the vending machine.

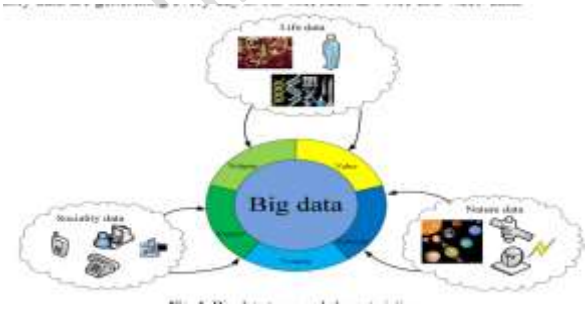


Figure: Big Data

4.1 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 [8] 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, and updating and information privacy.

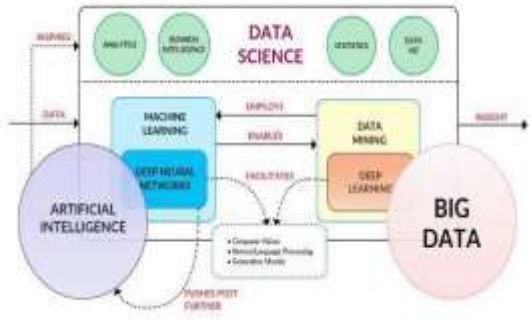


Figure : Big Data Scenario

5. Machine Learning Algorithms in Big Data Analytics:

Machine Learning is a sub-field of data science that focuses on designing algorithms that can learn from experience [5]and make predictions on the data. A computer program is said to learn from experience E with respect to some class of task T and performance measure P, if its performance as tasks T, as measured by P improves the experience E. Machine learning experience includes supervised learning, Unsupervised Learning and Reinforcement Learning methods[7]. Unsupervised methods actually start off from unlabeled data sets, so, in a way, they are directly related to finding out unknown properties in them (e.g. clusters or rules).

Machine learning focuses on prediction [8], based on known properties learned from the training data. Data mining (which is the analysis step of Knowledge Discovery in Databases) focuses on the discovery of (previously) unknown properties on the data .For instance, performance evaluation of a classifier involves dataset selection, performance measuring, error-estimation, and statistical tests [6]. The evaluation results may lead to adjusting the parameters

of chosen learning algorithms and/or selecting different algorithms.

While productive uses of machine learning can't depend entirely on packing consistently expanding measures of massive Information at calculations and seeking after the best, the capacity to use a lot of information for machine learning tasks is a categorical requirement has ability for specialist now. While much of machine learning holds true regardless of data amounts, there are aspects which are the exclusive domain of Big Data modeling, or which apply more so than they do to smaller data amounts.

Figure outlines a process for applying machine to Big Data in his original graphic. The process includes paths for descriptive, predictive, and prescriptive analysis, as well as simulation. Importantly, the machine learning process is explicitly noted as recursive, which is perhaps especially true of modeling large quantities of data, and it also breaks down the relative number of records at each successive stage of a machine learning task.

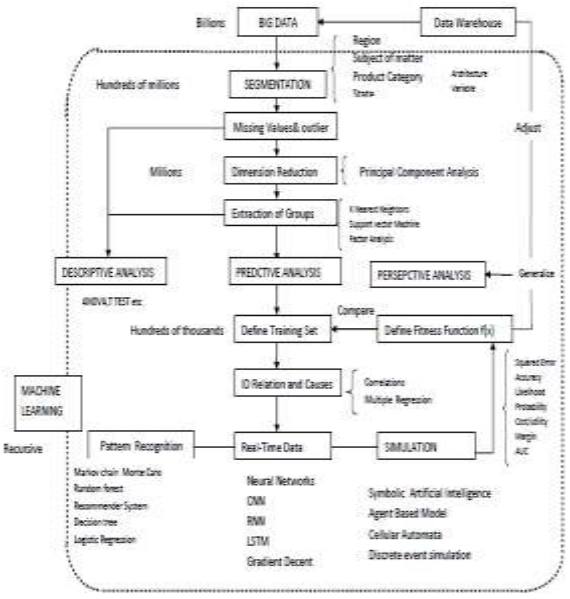


Figure : Role of Machine Learning in BigData

5.1 Tools for Machine Learning Algorithm in Big Data Analytics

In Big-Data situations, operators, managers and information researchers need to acquire data and learning from immense data sets or from wide surges of data. Keeping in mind the end goal to make this procedure simple, letting information researchers to concentrate on mechanizing this procedure and concentrate on the outcomes, a few systems have seemed to give such administration. Here a couple of them, however not by any means the only ones, are abridged.

Map-Reduce frameworks:

Apache Hadoop and Spark Most machine learning procedures can be parallelized by understanding a calculation procedure for each information and after that total the procedure yields into the arrangement. Such procedures can be explained utilizing a Map-Reduce arrangement: data is split in parts, each part is processed in parallel and the outcomes are amassed into the arrangement. Apache Hadoop [26] is a generally utilized open-source structure in Java for such purposes. Clients can send without anyone else bunches or server farms, or can get an answer from organizations offering Hadoop as Platform as a Service and concentrating just on conveying their applications. In view of Hadoop, Apache Spark [27] is an answer endeavoring to enhance execution by centering in particular functionalities like machine learning, graph analysis and data-streaming, and programmable in Scala or Python. While Hadoop has been available for long time and as of now has more capacities covering more parts of the business, Spark and Mahout develops by centering and enhancing particular issues, the vast majority of them related with machine learning and enormous information treatment.

Apache Spark

Apache Spark is a general-purpose analytics framework. It improves efficiency through in-memory computing primitives, Pipelined computation and it improves usability through APIs in Scala, but Java, Python, and R APIs also available and also works through interactive Shell. Spark provides a general middleware layer that re-implements existing learning tasks so they can run on a big data platform. Such a middleware layer often provides general primitives/operations that are useful for many learning tasks. This approach is suitable for users to try different learning tasks/algorithms within the same framework. The other category is to transform individual learning algorithms to run on a big data platform.

Apache Mahout:

Mahout is an open source project from Apache, offering Java libraries for distributed or otherwise scalable machine-learning algorithms. Popular implementation of Mahout will be done in the fashion of Latent Dirichlet Allocation. Mahout implementation of LDA are Collapsed Variation Bayes (CVB) and pipeline of Hadoop jobs.
These calculations cover exemplary machine learning undertakings, for example, arrangement, bunching, affiliation run examination, and suggestions. In spite of the fact that Mahout libraries are intended to work inside an Apache Hadoop setting, they are likewise good.

6. Big Data Mining

6.1 Local Learning and Model Fusion for Multiple Information Sources

As Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased decisions or models. Under such a circumstance, a Big Data mining system has to enable an information exchange and fusion mechanism to ensure that all distributed sites (or information sources) can work together to achieve a global optimization goal. Model mining and correlations are the key steps to ensure that models or patterns discovered from multiple information sources can be consolidated to meet the global mining objective. More specifically, the global mining can be featured with a two-step (local mining and global correlation) process, at data, model, and at knowledge levels. At the data level, each local site can calculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. At the model or pattern level, each site can carry out local mining activities, with respect to the localized data, to discover local patterns. By exchanging patterns between multiple sources, new global patterns can be synthesized by aggregating patterns across all sites. At the knowledge level, model correlation analysis investigates the relevance between models generated from different data sources to determine

how relevant the data sources are correlated with each other, and how to form accurate decisions based on models built from autonomous sources.

6.2 Mining from Sparse, Uncertain, and Incomplete Data

Sparse, uncertain, and incomplete data are defining features for Big Data applications. Being sparse, the number of data points is too few for drawing reliable conclusions. This is normally a complication of the data dimensionality issues, where data in a high dimensional space (such as more than 1,000 dimensions) do not show clear trends or distributions. For most machine learning and data mining algorithms, high-dimensional spare data significantly deteriorate the reliability of the models derived from the data. Common approaches are to employ dimension reduction or feature selection to reduce 36 *Jainendra Singh* the data dimensions or to carefully include additional samples to alleviate the data scarcity, such as generic unsupervised learning methods in data mining. Uncertain data are a special type of data reality where each data field is no longer deterministic but is subject to some random/error distributions. This is mainly linked to domain specific applications with inaccurate data readings and collections. For example, data produced from GPS equipment are inherently uncertain, mainly because the technology barrier of the device limits the precision of the data to certain levels (such as 1 meter). As a result, each recording location is represented by a mean value plus a variance to indicate expected errors. For data privacy related applications, users may intentionally inject randomness/errors into the data to remain anonymous. This is similar to the situation that an individual may not feel comfortable to let you know his/her exact income, but will be fine to provide a rough range like [120k, 160k].
For uncertain data, the major challenge is that each data item is represented as sample distributions but not as a single value, so most existing data mining algorithms cannot be directly applied. Common solutions are to take the data distributions into consideration to estimate model parameters. For

example, error aware data mining utilizes the mean and the variance values with respect to each single data item to build a Naive Bayes model for classification. Similar approaches have also been applied for decision trees or database queries. Incomplete data refer to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values (e.g., dropping some sensor node readings to save power for transmission). While most modern data mining algorithms have in-built solutions to handle missing values (such as ignoring data fields with missing values), data imputation is an established research field that seeks to impute missing values to produce improved models (compared to the ones built from the original data).

6.2.1 Mining Complex and Dynamic Data

The rise of Big Data is driven by the rapid increasing of complex data and their changes in volumes and in nature. Documents posted on WWW servers, Internet backbones, social networks, communication networks, and transportation networks, and so on are all featured with complex data. While complex dependency structures underneath the data raise the difficulty for our learning systems, they also offer exciting opportunities that simple data representations are incapable of achieving.

For example, researchers have successfully used Twitter, a well-known social networking site, to detect events such as earthquakes and major social activities, with nearly realtime speed and very high accuracy. In addition, by summarizing the queries users submitted to the search engines, which are all over the world, it is now possible to build an early warning system for detecting fast spreading flu outbreaks. Making use of complex data is a major challenge for Big Data applications, because any two parties in a complex network are potentially interested to each other with a social connection. Such a connection is quadratic with respect to the number of nodes in the network, so a million node networks may be subject to one trillion connections. For a large social network site, like Facebook, the number of active users has already reached 1 billion, and analyzing such an enormous network is a big challenge for Big Data mining. If we take daily user actions/interactions into consideration, the scale of difficulty will be even more astonishing.

7. Big Data Challenges

There are many future important challenges in Big Data management and analytics that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal during the next years:

• Analytics Architecture: It is not clear yet how an optimal architecture of analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz [4]. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real-time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, and extensible, allows ad hoc queries, minimal maintenance, and debuggable.

• Statistical Significance: It is important to achieve significant statistical results, and not be fooled by randomness.

• Distributed Mining: Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.

• Time Evolving Data: Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and in some cases to detect change first.

• Compression: Dealing with Big Data, the quantity of space needed to store it is very relevant. There are two main approaches: compression where we don't lose anything or sampling where we choose what is the data that is more representative. Using compression, we may take more time and less space, so we can consider it as a transformation from time to space. Using sampling, we are losing information, but the gains in space may be in orders of magnitude.

• Visualization: A main task of Big Data analysis is how to visualize the results.

As the data is so big, it is very difficult to find user-friendly visualizations.

• Hidden Big Data: Large quantities of useful data are getting lost since new data is largely untagged file based and unstructured data. The 2012 IDC study on Big Data [3] explains that in 2012, 23% (643 exabytes) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

8. Machine Learning Application To Big Data

Machine learning [5] is ideal for exploiting the opportunities hidden in big data. It delivers on then promise of extracting value from big and disparate data sources with far less reliance on human direction. An overview of the application to big data is given in the figure 4:

It is data driven and runs at machine scale. It is well suited to the complexity of dealing with disparate data sources and the huge variety of variables and amounts of data involved. 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

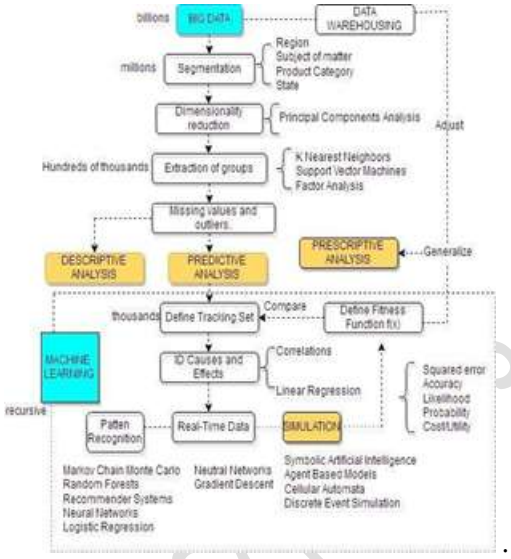


Figure.: Machine learning applications to Big data

“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.

9. Proposed Research Work

Unfortunately traditional analytics tools are not well suited to capturing the value hidden in Big Data. The volume of data is too large for comprehensive analysis. The range of potential correlations and relationships between disparate data sources, from back end customer databases through to live web based click streams, are too great for any analyst to test all hypotheses and derive all the value buried in the data. Machine learning is a rather new domain of IT and advanced mathematics, based on new statistical algorithms that could analyze big volume of diverse data sources (image, sound, video, social network, geo-localization, "traditional" structured database, etc…) in near real time. Computers, using these new types of programs, could learn from data for better future use. Whether it is health, education, trade or the environment, statistical machine learning allows to analyses and gives insight in different use cases even further. In fact, machine learning algorithms are used in very diverse contexts: to recognize handwritten text, to extract information from images, to build automatic language translation systems, to predict the behavior of customers in an online shop, to find genes that might be related to a particular disease, and so on. Generally speaking, machine learning algorithms can always be used, if we want to extract "patterns" from complex and large volume of data. In health, a lot of data are already stored on patients in various formats and it represents a huge data volume. In medical imaging, machine learning allows to see many things that we cannot see before. For example, coupling visual recognition appliance with these new ways to analyze big data helps doctors to monitor automatically elder people to see if they will fall or not (at home, in hospital, even in street). Network security is uniquely a Big Data problem. Machines on a network generate tons of data every day—within enterprises, one terabyte of data is easily generated daily. Such a large volume practically prevents commercial security tools from performing long-range analysis, such as base-lining network object behavior over a 30 day period or more for all objects on the network. Large data volume hampers researchers' ability to perform data mining experiments to gain necessary insights. Several approaches to machine learning are used to solve problems. The main focus will be on the two most commonly used ones — *supervised* and *unsupervised* learning—because they are the main ones supported by Mahout. Supervised learning is tasked with learning a function from labeled training data in order to predict the value of any valid input. Common examples of supervised learning include classifying email messages as spam, labeling Web pages according to their genre, and recognizing handwriting. Many algorithms are used to create supervised learners, the most common being neural networks, Support Vector Machines (SVMs), and Naive Bayes classifiers. Unsupervised learning, as you might guess, is tasked with making sense of data without any examples of what is correct or incorrect. It is most commonly used for clustering similar input into logical groups. It also can be used to reduce the number of dimensions in a data set in order to focus on only the most useful attributes, or to detect trends. Common approaches to unsupervised learning include k-Means, hierarchical clustering, and self-organizing maps. Apache Mahout is a new open source project by the Apache Software Foundation (ASF) with the primary goal of creating scalable machine-learning algorithms that are free to use under the Apache license. The project is entering its second year, with one public release under its belt. Mahout contains implementations for clustering, categorization, and evolutionary programming. Furthermore, where prudent, it uses the Apache Hadoop library to enable Mahout to scale effectively in the cloud.

10. 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 complex procedure as well as the computational overhead is huge in further training and testing phases of classification.
- 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.
- We will experiment our developed methods on standard datasets such as UCI ML-Repository, CORA, Reuters etc. and compare the analysis results with existing techniques

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.

Future scope of Machine learning analytics is how to make ML more declarative, so that it is easier for non-experts to specify and interact with different type of data in different streams. In the future, we will enhance and assess the performance of machine learning techniques for different types of problems. One promising direction is to extend the machine learning approaches towards big data, which are efficient and highly scalable in the way they process high-dimensional data

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. Then an extensive survey of related work and methods which have been developed in past, is presented. Later, we have listed out some tools which can be employed for big data management and analysis. To encourage more interests for the readers of the paper, in the end, some open issues in big data domain, problem identification and our future research goals were presented.
Machine learning techniques can solve such applications using a set of generic methods that differ from more traditional statistical techniques.

REFERENCES

[1] Xindong WU, Gong-Qing WU and Wei Ding, "Data Mining with Big Data," IEEE Transaction on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Dec. 2012.
[2] Wei Fan, Albert Bifet,"Mining Big Data: Current Status and Forecast to Future," SIGKDD Exploration, vol. 14,Issue 2,2013.
[3] J. Gantz and D. Reinsel. IDC: The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Dec. 2012
[4] N. Marz, J. Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications,2013.
[5] An Oracle White Paper June 2013
[6] Defending Networks with Incomplete Information: A Machine Learning Approach, BlackHat Briefings USA 2013

[7] A Trend Micro White Paper | September 2012
[8] Large-Scale Adaptive Machine Learning for Security Analytics, Ling Huang,
Joint work with ISTC and McAfee Labs ISTC Summer Retreat, 05/31/2013
[9] http://www.skytree.net/machine-learning
[10] *hadoop.apache.org/*
[11] *mahout.apache.org/*
[12] http://www.ibm.com/developerworks/library/j-mahout/
[13]http://www.networkworld.com/community/blog/defining-big-data-securityanalytics
[14] Douglas, Laney. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner. Retrieved 6 Feb 2001.
[15] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than
Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011. Retrieved 13 July 2011.
[16] The White Book of BIG DATA, FUJITSU
[17]www.mckinsey.com/mgi/publication/big_data

[18] 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.

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

[20] 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).

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

[22] Lidong Wang, Cheryl Ann Alexander. ”Machine Learning in Big Data”, International Journal of

Mathematical, Engineering and Management Sciences, Vol. 1, No. 2, 5261, (2016).

[23] Z. Pawlak. ”Information Systems Theoretical Foundations”, Information Systems, Vol. 6, No. 3, pp. 205-218, (1981).

[24] 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.

[25] 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).

[26] Christopher C Drovandi, Christopher Holmes, James M McGree, Kerrie Mengersen, Sylvia Richardson and Elizabeth G Ryan. ”A Principled Experimental Design Approach to Big Data Analysis”, QUT ePrints (2017).

[27] Yichuan Wang, LeeAnn Kung, William Yu Chung Wang, Casey G. Cegielski. ”An integrated big data analytics-enabled transformation model: Application to health care”, Information and Management, April (2017).

[28] Farzaneh Farhangmehr. ”Statistical Approaches for Big Data Analyt-ics and Machine Learning: Data-Driven Network Reconstruction and Predictive Modeling of Time Series Biological Systems”, escholarship, University of California, (2014).

[29] Yichuan Wang, Nick Hajli. ” Exploring the path to big data analytics success in healthcare”, Journal of Business Research 70 (2017) 287-299.

[30] Jacky Akoka, Isabelle Comyn-Wattiau, Nabil Laoufi. ” Research on Big Data - A systematic mapping study”, Computer Standards & Interfaces 54 (2017) 105-115.

[31] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vis-hanth Weerakkody. ” Critical analysis of Big Data challenges and analytical methods”, Journal of Business Research 70 (2017) 263-286.

[32] Shuliang Xu, Junhong Wang. ”Dynamic extreme learning machine for data stream classification”, Neurocomputing 238 (2017) 433-449.

[33] Xiaochuang Yao, Mohamed F. Mokbel, Louai Alarabi, Ahmed Eldawy, Jianyu Yang, Wenju Yun, Lin Li, Sijing Ye, Dehai Zhu. ”Spatial coding-based approach for partitioning big spatial data in Hadoop”, Computers & Geosciences 106 (2017) 60-67.

[34]. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, pp. 255-260, 2015.
[35]. Hey, T., Tansley, S., Tolle, K., editors (2009). The Fourth Paradigm: Data- Intensive Scientific Discovery. Microsoft Research. S.Hasan et a.l
[36]. Sun, Y. et al., 2014. Organizing and Querying the Big Sensing Data with Event-Linked Network in the Internet of Things. *International Journal of Distributed Sensor Networks*, 14, p.11.
[37]. Fan, J., Han, F. & Liu, H., 2014. Challenges of Big Data analysis. *National Science Review* , 1 (2), pp.293– 314.
38]. Parmar, V. & Gupta, I., 2015. Big data analytics vs Data Mining analytics. *IJITE*, 3(3), pp.258–263.
[39]. K. Sayood, Introduction to Data Compression, Morgan Kaufinarm Publishers, San Francisco, CA, 2000.
[40]. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, pp. 255-260, 2015.
[41]. N. Japkowicz and M. Shah, Evaluating Learning Algorithms: A Classification Perspective: Cambridge University Press, 2011.
[42]. S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed.: Prentice Hall, 2010.