Machine Learning with Big Data

Amit Kumar Tyagi[a], G.Rekha[b]

[a]Department of Computer Science and Engineering, Lingaya's Vidyapeeth, Faridabad - 121002, Haryana, India
[b]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522502.

ARTICLE INFO

Article History:
Received 12 January 19
Received in revised form 20 January 19
Accepted 21 February 19

Keywords:

Data Mining
Machine Learning
Unsupervised Learning
Supervised Learning
Semi-Supervised Learning
Big Data

ABSTRACT

In the past decade, machine learning techniques have been used for solving several problems with respect to big data. In current, there are several types of Machine Learning (ML) techniques available like supervised, unsupervised and semi-supervised. Similarly, several techniques like classification, Pre-processing, Association rules, Random forest, Decision tree, Support vector machines, etc. available to solve several problems like data imbalance, machine translation, enhancement in robotics, etc. Today's we need to several basic facts about machine learning techniques to solve many problems like prediction analysis in several applications, for example, in e- healthcare applications (big data: data generated from connected electronically smart devices)/ in other applications (agriculture, e-commerce, defence, etc.). For that, most of the researcher confused and hesitate to discuss/ decide which technique or metric to use in respective applications. How machine leaning techniques differs from Data mining techniques? So this article reviewed several existing work/ papers and presents a remedy for all (those) researchers, which does not solve only doubt from their first stage (selection of techniques or metrics and doubts regarding data mining and machine learning) but also mitigate several issues with respect to machine learning techniques (in compare to deep learning). Hence in summary, this work summarizes with various needful information related to Machine Learning (including Big Data). Following survey on evaluation metrics and some other related factors, this paper showed some future directions at last.

Peer review under responsibility of International Conference on Sustainable Computing in Science, Technology and Management.

1. Introduction About Data Mining, Machine Learning and Big Data

In general terms, Data Mining is a process to finding several new patterns in large collection of data sets using several techniques like classification, clustering, regression (Kumar, Tyagi, A. K., & Tyagi, 2014) etc. to predict future trends. In old days (from 1990 to 2010), data mining techniques was used quite a lot. But, when the revolution in smart technology enters, a lot of data was collected at server side (for all organisations). In other words, data mining was used in past decades to discover hidden pattern/ information/ unknown facts from a data. These hidden patterns play an essential role in increasing profit for organisation. On the other hand, a lot of data has been produced by several (integration of devices) of internet of things (via machine to machine to device to device communication). But available traditional data mining tools were not sufficient to handle this data or finding hidden patterns. So, a new term "Machine Learning" was created by Arthur Samuel (Weiss, 1992) and he defined Machine Learning as "ML is a field of study that gives computers the capability to learn without being explicitly programmed" (Weiss, 1992). The main focus of Machine Learning (ML) is classification and prediction. The ML algorithms learn based on previously known properties from the training data and perform prediction for unknown/ future properties. In general, ML algorithms need a problem formulation from a particular domain for prediction, for example, e-healthcare, agriculture, astronomy, etc. Note that "the term 'data mining' was introduced in late 1980s (the first Knowledge Discovery Database conference took place in 1989, coined by Gregory Piatetsky-Shapiro (but since it was trademarked by HNC, a San Diego-based company)), whereas the term machine learning has been in use since the 1960s", i.e., "data mining" appeared in the database community. Generally, Data mining is of two types, i.e., directed data mining, and undirected data mining (https://himadri.cmsdu.org/documents/datamining_metrics.pdf). Here directed data mining provides searching through historical (past) records to find patterns that explain a particular outcome, and uses classification, regression, prediction and profiling (for mining purpose), whereas, undirected data mining searches through the same (similar) records for finding hidden/ interesting patterns using clustering, finding association rules and description.

As discussed above, Data Mining never look for specific goal (for an area/ domain), but it always focuses on finding new or hidden and interesting knowledge. Data mining can be performed on following types of data, i.e., Relational databases, Data warehouses, Advanced Data bases and information

repositories, Object-oriented and object-relational databases, Transactional and Spatial databases, Heterogeneous and legacy databases, Multimedia and streaming database, Text databases, Text mining and Web mining. Data Mining techniques can be summarized as:

- Classification: This technique classifies the data based on different classes. It is mainly used to find the class label for new data samples. (https://data-flair.training/blogs/data-mining-techniques/).
- Clustering: Clustering analysis technique to identify data similarities by projecting the data on n-dimensional space.
- Regression: Regression analysis finds the relationship between variables by identify the likelihood of a specific variable, given the presence of other variables.
- Association Rules: This technique generates the association among different items in the form of rules. It discovers a hidden pattern in the data set.
- Outlier detection: This technique is used to find the unexpected pattern or unexpected behavior from the data. It is used in a various domain, such as intrusion detection, fraud or fault detection, etc. Another name for outlier detection is outlier Analysis or outlier mining.
- Sequential Patterns: It works on transactional data bases and helps to discover the similar patterns or trends exist in transaction data sets for certain period of time.

Hence, Data Mining is all about explaining the past and predicting the future for analysis. It helps in extraction of valuable information from huge amount of data. It is the process of mining knowledge from data. Data mining (Jackson, 2002) includes Business Understanding, Data Understanding, Data Preparation, Modelling, Evolution, Deployment. Some of important Data mining techniques are: Categorization, Classification, Clustering, Regression, Association rules, Outlier analysis, and Sequential patterns. The prominent data mining tools are Weka, R-language and Oracle Data mining. Data mining technique helps organisations/ companies to get knowledge-based information to improve their profit (with respect to their products). The application of data mining techniques has been used in several industries such as insurance, education, communication, manufacturing, banking, retail, ecommerce, service providers and many more.

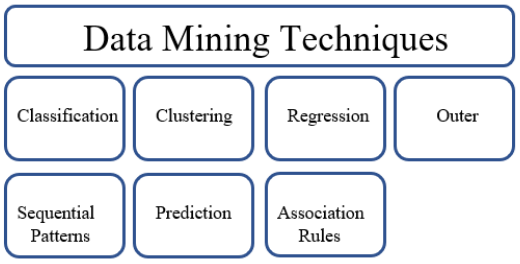

Figure 1 Collection of Data Mining Techniques

In Business understanding, business and data-mining goals are established. Several processes exist in data mining which are included as: Data understanding, Data preparation, data transformation, modelling, evaluation and deployment. Here in Data understanding, sanity check on data is performed to check whether it's appropriate for the performing data mining task. In Data preparation phase, data is made ready for generating models. Note that about 90% of the time of a project is consumed by the data preparation process. Data transformation operations would contribute toward the success of the mining process. In Modelling phase, mathematical models are used to determine data patterns. The generated pattern is evaluated against the business goals in evaluation phase. Finally, in deployment phase, the discoveries of patterns are adapted to everyday business operations.

On the other hand, machine learning uses algorithms like Supervised (Regression, Decision Tree, Random Forest, Classification) and Unsupervised (Clustering, Association Analysis, Hidden Markov Model, etc.). Here, supervised, unsupervised and re-enforcement learning are three types of machine learning. Supervised learning is used to predict value based on labelled data, whereas unsupervised learning predicts value based on unlabelled data, while re-enforcement learning learns from its own or work on feedback process. Re-enforcement learning is also called reward-based learning. ML algorithm typically consists of two phases: training and testing. Training data is used to develop a model whereas testing phase is used to validate model. In machine learning, the following steps are performed:

- Identify class attributes/ features and classes from training data.
- Identify a subset of the attributes necessary for classification (i.e., feature/dimensionality reduction).
- Learn the model using training data.
- Use the trained model to classify the unknown data.
- In the training phase, each feature with its associated class is learned by using appropriate algorithm from the training set. In the testing phase, new data are run through the model and the algorithm classifies as to whether it belongs to one of the classes specified in training data set.

Some of the examples of machine learning are: automatic reading of handwriting, assisted medical diagnosis, automatic text classification (spam filtering, classification of web pages) and financial predictions.

As discussed above, due to emerging of new technologies and all integrated devices, it is predicted that there will be a lot of data available/ generated in the next few years. The technology industry faces a great challenge in handling (produced, collected and stored) the unprecedented (large) amount of data. To handle such huge data the term "big data" has become a buzzword. It has been designed to process the Data of varying size and complexities. Laney (Laney, 2001) in 2001, described three main dimensions (challenges) faced by data management systems. These three dimensions (3V's) were introduced, i.e., volume, velocity and variety in 2001. These three dimensions are defined as follows:

- Volume (size of the data). The massive amount of data that is been generated very second and required scientists to rethink about the storage and processing models in order to develop appropriate tools to analyse it.
- Velocity (speed at which data is generated) The movement from batch processing to stream processing is needed, to handle such real-time data.
- Variety refers (data is presented in different data formats). As data can come in from many different sources and take on many different forms, the main issue is incompatibility of data format. A significant amount of time and effort is needed for designing efficient techniques to preparing the data for analysis.

Further, there two more V's were added (in past), i.e., Here a fourth V is now also sometimes added:

- Veracity: This refers to the data quality, which can vary greatly. It also consists uncertainty and inconsistencies in the data.

There are many more V's that gets added in the near future depending on the context. The next V, we added:

- Valence: This refers to how data (big data) can tie with each other, forming connections between datasets (disparate in nature).

But in general, there are 7 V's available in Big Data like Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value. Note that these V's are not final, it may decrease or increase in near future, for example, 3 V's, 2001 (and again), 4 V's, 2012 (and again), 4 V's, 2013, 7 V's, 2013, 6 V's, 2013, 5 V's, 2013, 10 V's, 2014, 8 V's, 2014, 5 V's, 2014 (and again), 7 V's, 2018 (and again). Some more V's can be in near future like Vagueness, Validity, Valor, Value, Vane, Vanilla, Vantage, Variability, Variety, Varifocal, Varmint, Varnish, Vastness, Vaticination, Vault, Veer, Veil, Velocity, Venue, Veracity, Verdict, Versed, Version Control, Vet, Vexed, Viability, Vibrant, Victual, Viral, Virtuosity, Viscosity, Visibility, Visualization, Vivify, Vocabulary, Vogue, Voice, Volatility, Volume, Voodoo, Voyage, Vulpine. Moreover this, in simple words, Machine Learning can be considered as the older sibling of Data Mining. Machine learning is a subset of Artificial intelligence, whereas deep learning is known as subset of machine learning. Therefore, this work concentrates on Machine Learning/ Data Mining methods. Moreover this, both data mining and machine learning used to predict (some) results from a large collection of data (generated from integration of billions of devices). Here, has used a combination of the different data mining techniques like classification, regression, trend analysis, sequential patterns generation, clustering, etc. It analyses past events or instances or activities in a right sequence for future prediction (about a product or user). Whereas, machine learning has used supervised (regression, decision tree, classification, etc.), unsupervised (clustering, association rule, etc.) and reinforcement learning techniques to similar work (make prediction) in efficient and reliable way.

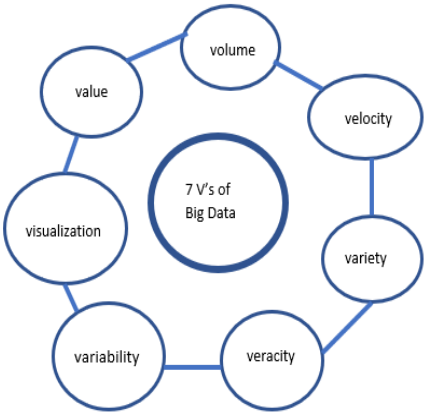

Figure 2 Big Data 7 V's

Note that data used for (by) machine learning and big data techniques can be broadly available in two types of variables: qualitative (ordinal or nominal) and quantitative (or numeric or discrete (often, integer) or continuous). Hence, finally the organization of this paper is follows as: Section 2 discusses about data mining tools, benefits of data mining with several drawbacks related to data mining. Then, essential tools, benefits of machine learning with some drawbacks are discussed in section 3. Further, section 4 discusses about big data ecosystem, its benefits and drawbacks. Further, metrics used to measure performance of analysis with respect to data mining and machine learning has been included in section 5. Then section 6

investigates several articles in past and include several challenges related to data mining, machine learning and big data. In last, Section 7 concludes this work with some future enhancements.

2. Data Mining Tools, Benefits and Drawbacks

This section discusses several tools to refine/ extract meaning data from a collection data. These mentioned tools used to find hidden pattern and information from a collection of data sets. During this process, we may face problem of class imbalance. Usually this problem occurs with machine learning techniques. In this class imbalance problem, number of a class of data (positive) is far less than other class (negative). Following tools are used to for analysis purpose (to data/ mining data) is referred as Data Mining Tools, included as:

- R-language: It is used for data mining/ analysis and visualization R language is an open source tool for statistical computing and graphics. R supports extensive variety of statistical and data mining/ machine learning techniques including classical statistical tests, time-series analysis, classification and graphical techniques. It also offers effective data management and storage facility.
- Oracle Data Mining (ODM): ODM is a component of the Oracle Advanced Analytics Database. This ODM allows data analysts to generate detailed insights/ patterns and makes predictions. It helps predict behavioural patterns of customer, develops customer profiles, and identifies cross-selling opportunities.
- Weka: It provides functions, such as data processing, feature selection, classification, regression, clustering, association rule, and visualization
- KNIME: It provides functions, such as data processing, feature selection, classification
- Excel: This tool is mostly used as data mining tools.
- R data mining tool: it is a programming language, used in analytics with respect to data mining. Note that it is a free version/ tool available on World Wide Web/ Internet.

Benefits of Data Mining:

- It helps companies to get knowledge-based information.
- It helps organizations to increase the revenue both in operation and production.
- It provides cost-effective and efficient solution compared to other statistical data applications.
- It helps in decision-making process.
- It facilitates in prediction of trends and behaviors automatically, as well as discovery of hidden patterns (in small data only).
- It can be deployed in new systems as well as existing platforms
- It analyze huge amount of data in less time.

Drawbacks of Data Mining

- In companies, there are likelihoods of selling valuable information of their clients to other companies for money, for example, American Express has sold credit card purchases of their customers to the other companies.
- The operation of many data mining analytics software requires advance training to work on with the techniques in an effective way.
- Due to different algorithms employed in data mining tools, the appropriate selection of data mining tool for an application will be a very tough task.
- If the data mining techniques are not accurately defined, it may cause serious consequences in certain conditions. The main drawback of data mining, the existing analytical tools require advance training to work (https://www.zaptox.com/tag/disadvantages-of-data-mining/https://www.zaptox.com/tag/disadvantages-of-data-mining/). Hence in the past decades, data mining is (were) used in various fields like Communications, Insurance, Education, Manufacturing, Banking, Retail, Service providers, e-commerce, Super markets, Crime investigation, and Bioinformatics, etc. Now next section will deal with definitions of machine learning, tools using in machine learning, benefits of machine learning and fields/ applications/ areas where machine learning is being used now days.

3. Machine Learning Tools, Benefits and Drawbacks

We already have discussed about machine learning in section 1. Machine learning is the process of building the products (software/hardware) that can predict the future outcome by learning from past examples. Machine Learning is training the data (collected one) using learning algorithms like Linear Regression, KNN (K-nearest neighbors), K-Means, Decision Trees, Random Forest, and SVM (Support Vector Machine) with datasets, so that the algorithms could learn to adapt to a new situation and find patterns that might be interesting and important. Again, ML is data-driven. For training Machine Learning, the dataset can be labeled. This is called a supervised learning and algorithms like Linear Regression and KNN, regression or classification used with respect to this learning. On the other hand, other datasets which are not labeled class called unsupervised learning (used Hidden Markov Model, Association Rule, etc., algorithms to train data). Each one can be discussed with an example as:

- Supervised learning: It comes with an "answer sheet", telling the computer what the right answer is, like which emails are spams and which are not. Another example is a coin is 3 gram or not.

- Unsupervised learning: K-Means to associate or cluster patterns that it finds without any answer sheet. Another example is a cricket team, which consists some bowlers and batsman.
- Semi-Supervised or re-enforcement leaning: For this, example is a given picture consist dog or cat.

Machine learning is a magic because in case of re-enforcement learning, machine learn from its feedback and try to give better result for next time. Following tools are used to for analysis purpose (mining data), machine learning Tools:

- KNIME
- RapidMiner
- Orange
- Apache Mahout
- Weka
- Apache Spark
- Others like R, C++, etc.

Some examples of machine learning tools with API's (application programming interfaces) include:

- Pylearn2 for Python
- Deeplearning4j for Java
- LIBSVM for C

Benefits of Machine Learning

- It used in Feature Learning. Using it, a system initialized (randomly) and learned on some datasets, which will learn good feature representations for a given work/ task.
- It (machine learning) also used for Parameter Optimization. It used a gradient based method of optimizing a large array of parameters. For example, a deep neural architecture (it has millions of parameters and not possible for a human to find such an optimal settings for large number of parameters by hand), so machine learning algorithms used, i.e., (stochastic gradient descent used to find an optimal setting here).
- In dynamic environment, machine learning is used to handle multi-dimensional and multi-variety data.
- It provides Fast Processing and Real-Time Predictions.
- It provides a continuous quality with large and complex process environments.
- The process of automation of tasks is easily possible.
- It is based on user's past search, Google and Facebook are using machine learning to present relevant advertisements.
- It is used to handle multi-dimensional and multi-variety data in dynamic environments.

As machine learning has many wide applications like banking, financial sector, healthcare, retail, and publishing, etc. That advertisement is based on users past search behavior. In summary, today's machine learning is playing an essential role in growth of any organisation (especially e-healthcare, e-commerce, or retail, etc.).

Drawbacks of Machine Learning

- A lot of training data is required to train the machine learning algorithms.
- ML Works with continuous loss functions, i.e., it is hard to optimize Non-differentiable discontinuous loss functions using machine learning techniques.
- For a Machine, learning is Limited, i.e., it is not a guarantee that machine learning algorithms will always work in every case imaginable. Sometimes or most of the times machine learning will fail, i.e., requires clear-cut about a problem to solve it with choosing right ML algorithm.
- Like deep learning algorithm (https://www.analyticsvidhya.com/blog/2017/04/comparison-between-deep-learning-machine-learning/), with ML it is difficult to work with it because of large data. It might be difficult or complex to work with a large amount of data.
- ML may susceptibility to errors due to lack of variability. Brynjolfsson and McAfee said that the actual problem with this inevitable fact and they also said that machine learning deals with statistical truths.
- The immediate predictions with a machine learning system may lead to fewer possibilities. As, it learns from historical data, the larger the data and the longer it needs to expose to these data, the better it will perform (more data means more accuracy).

Hence, Machine Learning works appropriately if the problem is actually solvable with the data that we possess. For example, if we want to build a model that predicts home prices based on the type of plants in the garden in each house, it's never going to work. There just is not any kind of relationship between the plants in each house and the house's sale price. So, the computer can never infer a relationship between the two. So, remember, if a human expert could not use the data to solve the problem manually, a computer probably would not be able to either. Now, Table 1 discusses several benefits and drawbacks of machine learning in brief.

Table 1: Benefits and Drawbacks of Machine Learning

Benefits	Drawbacks
Wide application	Acquisition
Advertisements	Interpretation limited
Utilization of resources	Time constraints in learning
Automation on tasks	Large data requirements
Quality, large and complex	Error diagnosis and correction
Handle multiple dimensional	Problem with verification

In summary, Machine learning is a field of research that mainly focuses on the theoretical, performance and properties of learning algorithms. It is highly interdisciplinary field combining artificial intelligence, statistics, mathematics, optimization, and many other disciplines of science and engineering. Machine learning supports wide range of applications covering almost every scientific domain. Machine learning is used in variety of problems like recommendation systems, autonomous control systems, recognition systems and many more. Hence, today's data plays a vital role in machine learning models and the new era of Big data is projecting machine learning techniques to be led in research and industry applications. The term "big data" refers to data that is too huge and complex to process on a single system/ machine. Hence this section discusses about machine learning, benefits and drawback of machine learning in brief. And now next section will deal with big data ecosystem and its benefits and drawbacks with respect to real world applications.

4. Big Data Ecosystem, its Benefits, and Drawbacks (According to Data)

Big data ecosystem consists several tools to extract/ refine big data in several scenario like in real time or task trainer simulation, manikin-based, virtual Reality or client side or to make predict based on available (collected) data. Note that Hadoop ecosystem is bigger than that, and the Big Data ecosystem is even bigger.
Figure 3 collects several tools inside in it. Now each tolls or method can be discussed as:

a. The Hadoop Distributed File System (HDFS): HDFS provides a scalable solution for storing and processing the huge amount of data in parallel and distributed manner. It divides the data into smaller chunks and stores it in different nodes.
b. MapReduce: MapReduce framework provides an interface for the distribution of sub-tasks to several mappers and finally the gathering of outputs from the reducer. When tasks are executed, MapReduce tracks the processing of each server/node.
c. PIG and PIG Latin (Pig and PigLatin): Pig programming language is configured to adapt all types of data including structured data/ unstructured data, etc. It consists of two key modules: PigLatin (the language itself), and the runtime version in which the PigLatin code is executed.
d. Hive: Hive is a runtime Hadoop support architecture with Hadoop platform powered with Structure Query Language (SQL). In other words, Apache Hive is a data warehousing tool that allows us to perform big data analytics using Hive query languages which is very similar to SQL.
e. Jaql: Jaql is a query language both with declarative and functional features. Its ability is to process a huge data. To ease the parallel processing, Jaql converts high level queries in to low level queries to support MapReduce.
f. Apache spark: It is an in-memory data processing engine. It efficiently executes streaming, machine learning or SQL workloads and needs fast iterative access to datasets.
g. Apache: Apache pig is used to analyse large data sets. It represents the huge data as data flows.
h. Graph X (Graph Computation): It is a graph computation engine. It combines data parallel and graph parallel concepts.
i. Mlib (Machine Learning): It contains machine learning libraries being built on top of spark.
j. Spark SQL (SQL): It is used for structured data. It is called spark streaming. It can hire queries on existing Hadoop deployment.
k. Spark core engine: it provides utilities and architecture for other components.
l. Zookeeper: Zookeeper provides centralized infrastructure with various services, providing synchronization across a cluster of servers. Big data analytics applications utilize these services to direct parallel processing across big clusters.
m. Apache HBase: It is a NO SQL database. It allows us to store both unstructured and semi-structured data easily and provided real time read or write access.
n. Oozie: Oozie is an open source project. It is used to streamlines the workflow and coordination among the tasks.
o. Mahout: Mahout is used to generate free applications of distributed and scalable machine learning algorithms that support big data analytics on the Hadoop platform.

Todays some of the Real-Time Big Data Analytics Tools are Storm, Cloudera, Gridgrain and SpaceCurve. There are several Real-Time Big Data Analytics tools available, but note that Real-Time Big Data Analytics is probably the ultimate usage of Big Data.

Benefits of Big Data Ecosystem (from data to decisions)

- Data creation
- Information processing
- Data acquisition
- Business process

Drawback of Big data Ecosystem

- Security
- Privacy
- Sharing of files among systems
- Complexity in data analytics
- Lack of decisions
- False predictions for critical data or applications

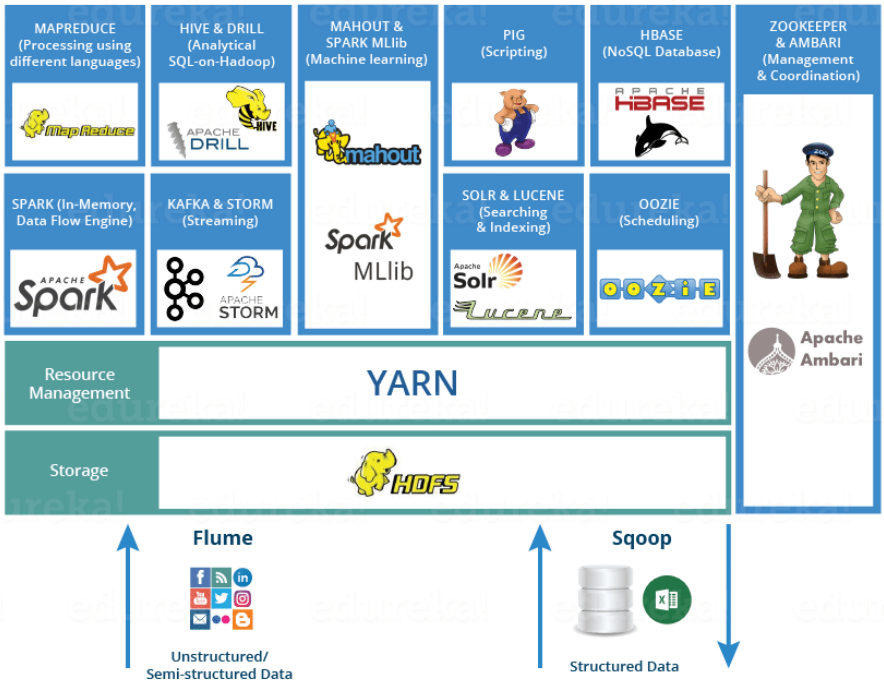

Figure 3 Big Data Ecosystem (https://www.edureka.co/blog/hadoop-ecosystem)

Apart above drawbacks, several limitations of most data analysis methods for big data are un-scalability and centralization, non-dynamic and uniform data structure. In general, Big Data is used to transform many different disciplines like improving health care, reducing carbon emissions, improving marketing efficiencies, etc. Using this big data, data mining or machine learning techniques are able to identify hidden trends and associations. In general, some of benefits of big data are providing cost reduction, faster, better decision making, new products and services, product recommendation, and fraud detection. Note that the basic use of big data is sharing the tasks across different nodes/systems to process parallelly. Hence, this section discusses about terms like data mining, machine learning, big data and its ecosystem with several benefits and drawbacks. Now next section will discuss about used metrics to measure performance in Data mining, Machine Learning and Big data.

5. Metrics used to Measure Performance in Data mining, Machine Learning and Big data

Generally, metrics for any algorithms or techniques are used to determine that "how much that particular algorithm is working perfectly in case of performance"? It is used to evaluate the results of algorithms (e.g., data mining). In general, Data mining algorithms are measures using accuracy, reliability, and usefulness. Where accuracy is a measure of how better the classifier performance relates an outcome with the attributes in the data. And reliability evaluates the way that a data mining technique performs on different data sets, whereas, usefulness includes various metrics in data mining that tell us whether the model provides useful information or not. Note that measuring the effectiveness or usefulness (with respect to data mining approach) is not always enough. Moreover this, some other evaluation metrics for data mining tasks are: cross-validation, holdout method, random sub-sampling, k-fold cross validation, leave one out method, bootstrap, confusion matrix, Receiver Operating Curves (ROC). Also, there are several metrics for association rule mining like Support, Confidence, Lift, Succinctness and Conviction.

Table 2: Performance Metrics, Formula and its Description

Metric	Formula	Description
Sensitivity/Recall	$\frac{TP}{TP + FN}$	It is the ability of the classifier in identifying the positive class correctly. Also Known as True Positive Rate (TPR). It is the measure of completeness.
Specificity	$\frac{TN}{FP + TN}$	It is the ability of the classifier in identifying the negative class correctly. Also Known as True Negative Rate (TPR).

Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$	It is the ability of the classifier to correctly predict both the classes (positive and negative). It is the proportion of true result.
False Positive Rate (FPR)	$\frac{FP}{FP + TN}$	It is the proportion between the number of negative samples wrongly categorized as positive (false positives) and the total number of actual negative samples. Also known as False Alarm Ratio.
False Negative Rate (FNR)	$\frac{FN}{FN + TP}$	It is the proportion between the number of positive samples wrongly categorized as negative (false negatives) and the total number of actual positive samples.
Precision	$\frac{TP}{TP + FP}$	It is the measure of exactness. It is the proportion of samples from positive class correctly classified as positive. It specifies the number of correct classification for positive class.
F-measure	$\frac{2 \times Precision \times Recall}{Precision + Recall}$	Also known as F-score. It measures the test's accuracy. It is the harmonic average of precision and recall. The perfect score is 1.
G-mean (Geometric Mean)	$\sqrt{Precision \times Recall}$	It is the ability of the classifier in balancing the classification accuracy between the positive and negative classes. Also known as G-Measure.

As discussed above, machine learning is used to predict fraudulent insurance, existing policy updates, admission in hospitals, uses of medicine of particular diseases, churn analysis, etc. we found that, Machine learning is proactive and specifically designed for "action and reaction" industries. There are several applications for Machine Learning (ML), the most significant of which is data mining. Machine learning algorithms are often used to establish relationships between multiple features. The main motivation of machine learning is to improve the efficiency of systems and the designs of machines. The datasets consist of some set of features. These features may be continuous, categorical or binary. If every instance in the dataset is associated with known labels (the corresponding correct outputs) then the learning is called supervised learning, in contrast to unsupervised learning, where instances are unlabelled.

Hence, this section discusses several metrics used in Machine Learning and also explains reason behind "why we use them with one small example". Basically, main objective of this section is to discuss used metrics with machine learning, for example, machine learning techniques used to solve a hot problem "class imbalance problem", and these metrics (in table 2) used to measure performance of used algorithms (Rekha, Krishna Reddy, & Tyagi, 2018).

Confusion Matrix: In the field of machine learning and specifically the problem of statistical classification, a confusion matrix (or error matrix) (Rekha, Krishna Reddy, & Tyagi, 2018) is used as metric to measure the performance of DM/ML algorithms (refer table 3).

Table 3: Confusion Matrix

	Positive Prediction	**Negative Prediction**
Positive class	TP	FN
Negative class	FP	TN

Hence to find which system is faster than another, we require metrics to measure the performances of system. For that, we will need better metrics than just counting the number of mistakes made. So, we introduce the concept of True Positive, True Negative, False Positive and False Negative. In two-class problems, the confusion matrix records the information about predicted and actual classifications done by the classifier. Table 3 shows the confusion matrix with its values.

- True positive (TP): The percentage of positive cases correctly classified as positive.
- True negative (TN): The percentage of negative cases correctly classified as negative.
- False positive (FP): The percentage of negative cases incorrectly classified as positive.
- False negative (FN): The percentage of positive cases incorrectly classified as negative.

These numerous performance metrics are used for evaluating binary classification problem based on the values obtained from the confusion matrix (refer table 3). Hence this section several metrics related to data mining and machine learning approaches/ techniques. Now next section will discuss several issues with respect to data mining, machine learning and big data (and analytics).

6. Challenges with Data Mining, Machine Learning and Big Data

As discussed above, advantages of Real-Time Big Data Analytics are any faults within the organisation are known instantly, New strategies of the competitor can be noticed immediately, in turn service improves dramatically, which could lead to higher conversion rate and extra revenue, Fraud can be detected the moment it happens and proper measures can be taken to limit the damage and cost savings (the implementation of a Real-Time Big Data Analytics tools may be expensive, it will eventually save a lot of money).

6.1 Challenges of Implementation of Data mining approaches on data, are:

- To formulate the data mining queries, a skilled expert is needed.
- Overfitting, i.e., due to small samples size in the training database.
- Large databases sometimes are difficult to manage using data mining.
- Business practices may need to be modified to determine to use the information uncovered.
- If the data set is not diverse, data mining results may not be accurate.
- Integrated information needed from heterogeneous databases and global information systems could be complex

6.2 Challenges in making prediction/ analysing using machine learning techniques, are:

- Memory Networks.
- Natural Language Processing (NLP)
- Attention.
- Understand Deep Nets Training.
- One-Shot learning.
- Deep Reinforcement Learning to Control Robots.
- Semantic Segmentation.
- Video Training Data.
- Object detection
- Democratizing Artificial Intelligence

The following are the major challenges in machine learning, like Problem Formulation (a wrong problem formulation would lead to dead ends, performing data engineering, cleaning, feature extraction/selection etc., applying the right ML methods, choosing of proper parameters to tune our algorithms, or adopting of right method/ technique/strategies, for example, cross-validation, performing fair experiments. These all the area (tasks) is where many methods fail and does not lead to any significant impact. In last making right conclusions, speed of solving problem with good accuracy is too slow.

Apart above challenges, a key challenge in machine learning is "How to represent The Input Data"? This field of study is known as Representation Learning or feature learning. Also misunderstanding the statistical or maths formulations can be critical to make widespread and scalable algorithms for a computer, which is also a one challenges.

6.3 Challenges in making analytics with Big Data, are:

- Privacy should be seen as a set of rules encompassing flows of information in ethical ways but not the ability to keep data secret.
- Shared information can still be confidential.
- Big data mining requires transparency.
- Big data can threaten privacy.

As major challenges in big data analytics are: requirement of special computer power, unstructured data and provenance of data, missing or incomplete data, quality of data, data security, and lack of experts. The standard version of Hadoop is, at the moment, not yet suitable for real-time analysis. New tools need to be bought and used. There are however quite some tools available to do the job and Hadoop will be able to process data in real-time in the future. Using real-time insights requires a different way of working within your organisation: if your organisation normally only receives insights once a week, which is very common in a lot of organisations, receiving these insights every second will require a different approach and way of working. Insights require action and instead of acting on a weekly basis this action is now in real-time required. This will have an effect on the culture. The objective should be to make your organisation an information-centric organisation.

7. Conclusion with Future Works

There are several big data analytic techniques available like data mining, social network analysis, web mining, machine learning, visualization approaches, and optimization methods. We used big analytics to reduce cost to predict or analysis large data (e.g., cost effective storage for huge data sets), to receive next generation products (e.g., automated car, healthcare, etc.), improved service and products. Big data is the only reason where we apply data mining and machine learning techniques. Hence, we can say data mining and machine learning are the techniques to extract hidden pattern from a large data, this process is called big data analytics. Machine learning is a hot topic in the past decade to solve several problems with several learning techniques. But

having confusion in selection of machine learning techniques or choosing metrics, researchers received a lot of problems. Hence, this work tries to overcome this problem and try to remove this burden and gave explanation of every terms or metrics or techniques (with application-wise). Preserving Privacy to user/ protecting leaking information during making device to device communication (during analysing data) and providing security to available information (or transfer information with enough encryption mechanism) is a critical issue. So now we invite all researchers who are willing to work in near future/ working in Machine learning with Big Data, are kindly invited to do their work according to their interests/ areas (in selection of techniques/ metrics from given above).

Acknowledgements

This research is funded by the Lingaya's Vidyapeeth and Anumit Academy's Research and Innovation Network (AARIN), India. The authors would like to thank Lingaya's Vidyapeeth and AARIN, India, an education foundation body and a research network for supporting the project through its financial assistance.

Conflict of Interest

The authors have used reference number 1 and 10 as self-citation of their (his/ her) work. No author has an objection of citing their work.

REFERENCES

Kumar, A., Tyagi, A. K., & Tyagi, S. K. (2014). data Mining: Various Issues and Challenges for Future A Short discussion on Data Mining issues for future work. *International Journal of Emerging Technology and Advanced Engineering*, *4*(1), 1-8.

Weiss, E. A. (1992). Biographies: Eloge: Arthur Lee Samuel (1901-90). *IEEE Annals of the History of Computing*, *14*(3), 55-69.

https://himadri.cmsdu.org/documents/datamining_metrics.pdf

https://data-flair.training/blogs/data-mining-techniques/

Jackson, J. (2002). Data mining; a conceptual overview. *Communications of the Association for Information Systems*, *8*(1), 19.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

https://www.zaptox.com/tag/disadvantages-of-data-mining/https://www.zaptox.com/tag/disadvantages-of-data-mining/

https://www.analyticsvidhya.com/blog/2017/04/comparison-between-deep-learning-machine-learning/

https://www.edureka.co/blog/hadoop-ecosystem

Rekha, G, Krishna Reddy, V., & Tyagi, A. K. (2018). A Novel Approach to solve class imbalance problem using Noise Filter method, in: *Proceedings of the 18th International Conference on Intelligent Systems Design and Applications (ISDA), VIT Vellore, Tamilnadu, India*.