*International Journal of Latest Trends in Engineering and Technology*
*Vol.(13)Issue(2), pp.096-101*
*DOI: http://dx.doi.org/10.21172/1.132.18*
*e-ISSN:2278-621X*

# MACHINE LEARNING ALGORITHMS TO SOLVE PROBLEMS OF HETEROGENOUS BIG DATA

Madhavi Tota[1]

**Abstract: With the revolution in Big Datait transformsthe data by enabling optimization, enhancing insight quality and improving decision making. The extraction of this heterogonous data from such massive data through data analytics; machine learning is at its important because of its ability to learn from data with different learning algorithms and provide data driven insights,decisions, and predictions. In this research wediscuss differentchallenges, the cause effects according to Big Data ordifferent dimensions ofdata. Now a days Health Care is important issue and need improvementin health science. There are multiple processes going on within health sector. As vast amount of healthcare data is increasing every day, it is believed that extracting knowledge by data analysis process is essential. Aeducation system generates massive knowledge by means of the services provided. This result in a researchers to put forward solutions for big data usage, depending on learning analytics techniques as well as the big data techniques relating to the educational field. This paper summarizes the role of big data analysis and prediction in healthcare, educational system provides a perspective on the domain, identifies research gaps and opportunities, and few machine learning techniques.**
**Keywords: Big Data, Machine Learning, Unsupervised learning, SVM, MapReduce**

## 1. INTRODUCTION

Now a day'senormous data is producing so it is absolutely important to use analytical techniques on huge, diverse big data sets to extract useful knowledge and information from it.Big data analytics is a research area that deals with the collection, storage and analysis of immense data sets toextract the unknown patterns and other important information. Big data analytics helps us to identify the data that are integral component to the future decisions. Big data analytics can be abundantly found in domains such as banking and insurance sector, healthcare, education, social media and entertainment industry, bioinformatics applications etc. Particularly, digital data generated from a variety of digital devices they are growing at high speed.

A number of techniques have been developed to work with machine learning algorithmson large

Datasets: examples are new processing such as MapReduce and distributed processing frameworks such as Hadoop [12].

This paper mainlyfocuses on machine learning (ML) as anessential component of data analytics. The McKinsey Global Institute has stated that ML will be one of the main drivers of the Big Data revolution [5]. The reason for this is its ability to learn from data and provide data driven insights, decisions, andPredictions [9]. It is based on statistics and, similarly to analysis of data, can extract patterns from data. According to the nature of the available data, the two main categories of learning tasks are: supervised learning when both inputs and their desired outputs are known and the system learns to map inputs to outputs and unsupervised learning when desired outputs are not known and the system itself discovers the structure within the data. Classification and regression are examples of supervised learning: in classification the outputs take discrete values while in regression the outputs are continuous. Some of the examples of classificationalgorithms are k-nearest neighbor, logistic regression, and Support Vector Machine (SVM) while regression examples include Support Vector Regression (SVR), linear regression, and polynomial regression. Some algorithms such as neural networks can be used for both, classification and regression.

Unsupervised learning includes clustering which group objects based on their similarity criteria; k-means is an example of such algorithm. Predictive analytics relies on machine learning to develop models built using past data in an attempt to predict the future [7]; numerous algorithms including SVR, neural networks, and Naïve Bayes can be used for this purpose.

## 2. BIG DATA

The massive data generated by online applications, education system and many moresystems are structured, semi-structured and unstructured. Given such fact, preconditionis to make an in-depth analysis focusing on all the massive data dimensions.

As educational system generates large andmultidimensional data that are[2]:

• Varied: these are structured, semi-structured and unstructured.This constraint complicates the phases of knowledge extraction and decision making.

[1] Assistant Professor, Department of Information Technology, Rajiv Gandhi College Of Engineering Research & Technology, Chandrapur

• Voluminous: these are enormous data that can reach TIRA Bit. Given this constraint,there is a large amount of data which are generated through the actor interactions.
• Distributed: these are massive data which are stored on multiple servers as well asdifferent locations. It should be noted that the problem of knowledge distributionalso constitutes a major constraint in the process of knowledge extraction.

Big data in healthcare refers to electronic health data sets as they are large and complexity are also very high so it's very difficult to manage with traditional software and/or hardware . Big data in healthcare is overwhelming not only because of its volume but also because of thediversity of data types and the speed at which it generates. The complete data related to patient healthcare makes "big data" in the healthcare sector. By discovering relations between data and understanding patterns and trends within the data, big data analytics has the potentialto improve care, save lives at lower costs. Thus, big data analytics applications in healthcare take advantage of the explosion indata to extract insights for making better informed decisions.

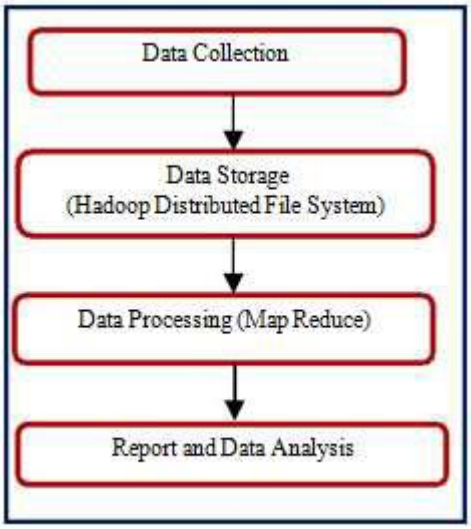


Figure (1) Big Data architecture[6]

Big data challenges in Health care[6]
- Extracting knowledge from complex or unstructured data set.
- Understanding unstructured clinical notes in the right context.
- Efficiently handling large volumes of medical imaging data and extracting potentially useful information andbiomarkers.
- Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additionallayers of complexity.
- Big data analytics platform in healthcare must support the key functions necessary for processing the data.
- Real-time big data analytics is a key requirement in healthcare.
- The lag between data collectionand processing has to be addressed.

**3. MACHINE LEARNING**

Machine leaning mainly focuseson the theory, performance, and properties oflearning systems and algorithms. It is a highly interdisciplinaryfield building upon ideas from many differentkinds of fields such as artificial intelligence, optimizationtheory, information theory, statistics, cognitive science,optimal control, and many other disciplines of science,engineering, and mathematics [14]. Because of itsimplementation in a wide range of applications, machinelearning has covered almost every scientific domain,which has brought great impact on the science and society[10]. Generally, the field of machine learning is divided intothree subdomains: supervised learning, unsupervisedlearning, and reinforcement learning [11].

Table 1: Comparison of Machine learning technologies[10]

| Learning types | Data processing tasks | Distinction norm | Learning algorithms |
|---|---|---|---|
| Supervised learning | Classification/Regression/Estimation | Computational classifiers | Support vector machine |
| | | Statistical classifiers | Naïve Bayes |
| | | | Hidden Markov model |
| | | | Bayesian networks |
| | | Connectionist classifiers | Neural networks |
| Unsupervised learning | Clustering/Prediction | Parametric | K-means |
| | | | Gaussian mixture model |
| | | Nonparametric | Dirichlet process mixture model |
| | | | X-means |
| Reinforcement learning | Decision-making | Model-free | Q-learning |
| | | | R-learning |
| | | Model-based | TD learning |
| | | | Sarsa learning |

A simple comparison of these three machine learning technologies from different perspectives is given in Table 1to outline the machine learning technologies for dataprocessing. The "Data Processing Tasks" column of thetable gives the problems that need to be solved and the"Learning Algorithms" column describes the methodsthat may be used.

**4. MACHINE LEARNING ALGORITHMS**

The field of machine learning is divided intothree subdomains: supervised learning, unsupervisedlearning, and reinforcement learning

*4.1. Supervised Learning*

This algorithm consists of a target / result variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we can generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.

*4.2. Unsupervised Learning*

In this algorithm, we do not have any target or outcome variable to predict / estimate the result. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriority algorithm, K-means.

*4.3. Reinforcement Learning:*

Using this algorithm, the machine is trained to make specific decisions. The machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions. Example of Reinforcement Learning: Markov Decision Process.

*4.4 List of Common Machine Learning Algorithms*

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. K-NN

*4.4.1. Linear Regression*

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation

Y= a *X + b.

*4.4.2. Logistic Regression*

It is a classification not a regression algorithm. It is used to estimate discrete values ( Binary values like 0/1, yes/no, true/false ) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logic function. Hence, it is also known as logic regression. Since, it predicts the probability, its output values lies between 0 and 1

*4.4.3. Decision Tree*

It is a type of supervised learning algorithm that is mostly used for classification problems. It works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

*4.4.4. SVM (Support Vector Machine)*

It is a classification method. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. The distance between the hyper plane and the closest data points is referred to as the margin. The best or optimal hyper plane that can separate the two classes is the line that has the largest margin. Only these points are relevant in defining the hyper plane and in the construction of the classifier. These points are called the support vectors.[10]

*4.4.5. Naive Bayes*

It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naive Bayesian model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

*4.4.6. k-NN (k- Nearest Neighbors)*

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K-Nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k -neighbors. The case being assigned to the class is most common amongst its K- nearest neighbors measured by a distance function.

These distance functions can be Euclidean, Manhattan, Murkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing k-NN modeling.

## 5. MACHINE LEARNING ALGORITHMS FOR BIG DATA ANATYTICS

Machine Learning is a sub-field of data science that focuses on designing algorithms that can learn from past experience [5]and make predictions on the data.[10] In Big-Data situations, operators, managers and information researchers need to extract data and learning from immense data sets or from wide varieties of data sets.

Challenges of Big data Analytics:

- learning for large scale of data
- learning for different types of data
- learning for uncertain and incomplete data
- learning for high speed of streaming data
- learning for data with low value density and meaning diversity

In terms of different data tasks, types, and characteristics, the required learning techniques are different, even a machine learning methods base is needed for big data processing. The learning systems can fast refer to the algorithm base to handle data [8]. A general means of machine learning algorithms on multicore with the advantage of MapReduce were investigated to enable the parallel and distributed processing to be possible

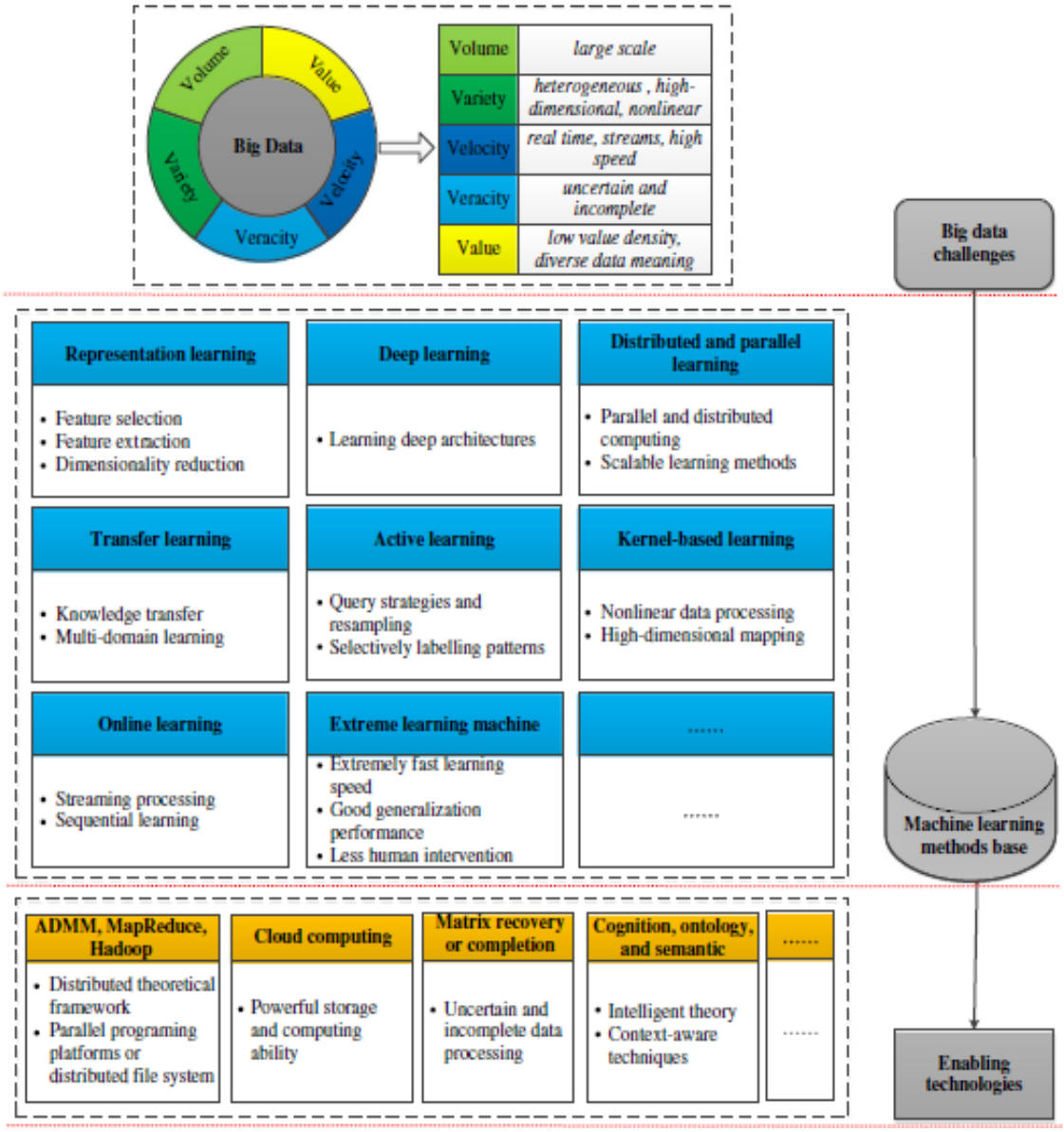


Fig. 2 Hierarchical framework of efficient machine learning for big data processing[8]

## 6. CONCLUSION

Big data are now fastly expanding in health care, education system andengineering domains. Learning from these massive datais critical and bring significant opportunities for various sectors. However, mosttraditional machine learning techniques are not abundantlyefficient or scalable enough to handle the data withthe characteristics of four V's i.e. volume, different types, highspeed, uncertainty and incompleteness, and low valuedensity. Machine learning needs to reintroduce itself for big data processing and its analytics. Therefore several advanced, efficient andintelligent learning algorithms are required to handle the hugeand heterogeneous datasets. The results obtainedthrough different analytical techniques provide more effectivesolutions to many real world problems in various domainssuch as healthcare, agriculture, social media, banking etc.This paper is abrief review of conventional machine learning algorithms for big data analytics. ML is fundamental to represent the difficulties focused by big data and extract patters, information, and bits of knowledge from enormous information.

Now also many different sectors need more attention to evaluate accurate results which are important for real world applications. 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. By using different analyzing and prediction technique with the help of machine learning algorithm,better healthcare and education system prediction can be obtained.

## 7. REFERENCE


[1] Machine Learning With Big Data: Challenges and Approaches by ALEXANDRA L'HEUREUX,KATARINA GROLINGER,AND MIRIAM A. M. CAPRETZ date of publication April 24, 2017. Digital Object Identifier 10.1109/ACCESS.2017.2696365

[2] Machine Learning Based On Big Data Extraction of Massive Educational Knowledge at https://doi.org/10.3991/ijet.v12.i11.7460 by AbdelladimHadioui!!", Nour-eddine El Faddouli, YassineBenjellounTouimi, and Samir BennaniiJET – Vol. 12, No. 11, 2017 page no.151 to 167

[3] X. W. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives", in IEEE Access, vol. 2, pp. 514-525, 2014. DOI: 10.1109/ACCESS.2014.2325029.

[4] Li Deng, "A tutorial survey of architectures, algorithms and applications for deep learning", APSIPA transactions on Signal and Information Processing, vol. 3,pp.1-29,2014.DOI: https://doi.org/10.1017/atsip.2013.9.

[5] "A Survey on Machine learning assisted Big Data Analysis for Health Care Domain" by RaolPriyankaAjaysinh, HinalSomani 2016 IJEDR | Volume 4, Issue 4 | ISSN: 2321-9939 page no. 550-554

[6] TUMK-ELM:" A Fast Unsupervised Heterogeneous Data Learning Approach" by LINGYUN XIANG, GUOHAN ZHAO, QIAN LI, WEI HAO , AND FENG LI VOLUME 6, 2018 2169-3536 2018 IEEE. Translations and content mining page no. 35305 – 35315

[7] "A survey of machine learning for big data processing" , Article in Journal on Advances in Signal Processing · December 2016 DOI: 10.1186/s13634-016-0355-x

[8] W. Tu and S. Sun, "Cross-domain representation-learningframework with combination of class-separate and domainmergeobjectives", Proceedings of the 1st InternationalWorkshop on Cross Domain Knowledge Discovery in Web andSocial Network Mining , ACM, pp. 18–25, 2012. DOI:10.1145/2351333.2351336.

[9] "A SURVEY OF MACHINE LEARNINGALGORITHMS FOR BIG DATA ANALYTICS" by Athmaja S. Hanumanthappa M. VasanthaKavitha2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)

[10] "Machine Learning Algorithms in Big data Analytics Article"in INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING · January 2018 DOI: 10.26438/ijcse/v6i1.6370

[11] White, T. (2009). Hadoop: The Definitive Guide (1st edition). O'Reilly Media, Inc. Software available from https://hadoop.apache.org

[12] WullianallurRaghupathi, VijuRaghupathi, "Big data analytics in healthcare:promise and potential", NCBI, 2014,DOI: 10.1186/2047-2501-2-3

[13] Manikandan, S.G., Ravi, S.,(2014). "Big data analysis using Apache Hadoop", in:ITconvergenceand Security (ICITCS), 2014 International Conference on. IEEE, pp. 1–4.https://doi.org/10.1109/ICITCS.2014.7021746