IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Machine Learning : Survey

Anamica Tejpal[1*], Kamaljit Kaur[2]

[1,2]Dept. CET, Guru Nanak Dev University, Amritsar, India

[*]*Corresponding Author: anamikatejpal123@gmail.com, Tel.: +00-12345-54321*

DOI: https://doi.org/10.26438/ijcse/v7i2.453457 | Available online at: www.ijcseonline.org

Accepted: 12/Feb/2019, Published: 28/Feb/2019

***Abstract*—**In this era, Machine Learning (ML) is persistently releasing its power in extensive variety of applications. It has been observed in previous years partly owing from advert of massive data. Huge information empowers machine learning calculations to reveal all the designs and make more precise predictions than ever before. In another way, machine learning presents challenges in field of data mining and big data. In this paper, we discussed what machine learning is and how it is related with big data. Here, we have introduced some phases of ML and the tools used to perform accurate prediction and how it is helpful in future tasks. This paper also has been discussed the opportunities and challenges associated with ML.

***Keywords*—**Machine Learning, Big Data, Data Mining, Knowledge Discovery. .

I. INTRODUCTION

These days, in various applications like data mining, health, Internet of things, face recognition, speech recognition, Machine Learning shows great impact. The approach of massive data has impelled wide interest in ML(Machine Learning). On other hand, the various learning algorithms have never been implemented by massive data in various industry applications. Big data is the best source to provide data for implementing modern machine learning algorithms. Some of the traditional machine learning techniques fails in various parameters like flexibility and scalability to handle big data. With the era of big data, Machine Learning technology had grown in advance to build predictive models[1] based on big data. ML give rise to various challenges like how to improve the sale of products and what parameters are require to control the drug usage.

This paper focuses on different machine learning algorithms that works on big data in new computing environment. Here, we focuses on opportunities and various challenges to handle massive data. In addition, massive data provides more challenges to machine learning algorithms. In this paper, Second section discussed the introduction of machine learning concept. Third section discussed its relation with big data. Fourth section discussed the traditional methods of machine learning. Fifth section discussed the modern machine learning strategies. Sixth section discussed the way to select the machine learning tools. Seventh section discussed the terminologies used for machine learning. Eighth section discussed the challenges and future scope of machine learning. Last section discussed the conclusion of the paper.

II. MACHINE LEARNING

In As we are living in data rising era, the main source of data are social networks, mobile gadgets and more. These sources generates the digital data and is growing very fast. According to[2], in 2020 the volume of data will reach to 37 trillion GB, which give rise to massive data. Most of the researchers have reviewed on big data topic and they have resulted big data in different area like innovation, productivity and competition[3]. Some of the researchers have related big data, machine learning and Internet of Things[4]. Various researchers have related machine learning with data mining, speech recognition also. In the previous years, machine learning algorithms have worked on complex data. But this increasing data is difficult to handle by traditional methods. The new advanced learning methods like deep learning, transfer learning, parallel and distributed learning methods come into existence for massive data.

Machine Learning is basically refer as the task of learning from various execution measures. ML strategies empower users to reveal hidden structures and make future predictions from extensive datasets. ML flourishes with efficient learning methods and various computing environments. Therefore, ML has extraordinary potential for handling enormous data analysis.

III. MACHINE LEARNING RELATED TO BIG DATA

Machine learning mainly follows pre-processing, learning and testing steps. Here, pre-processing step handles the unstructured data. In this step, generally the data that contains noise, incomplete and inconsistent is considered as

input. This phase convert data into semi-structured form by following various steps like remove missing values, data cleaning and various transformation. This produced data is used as input for learning step. The next phase known as learning phase consider the learning algorithm based on learning type and produce the output as model using input data. Some of the learning methods can be used in both pre-processing step and learning step. The testing step is helpful to get the desired output. This step tells the efficiency of the system. This phase results in type of various parameters like true positive, true negative, statistical tests, error estimation and many more[5]. The components of machine learning are shown in the figure 1.

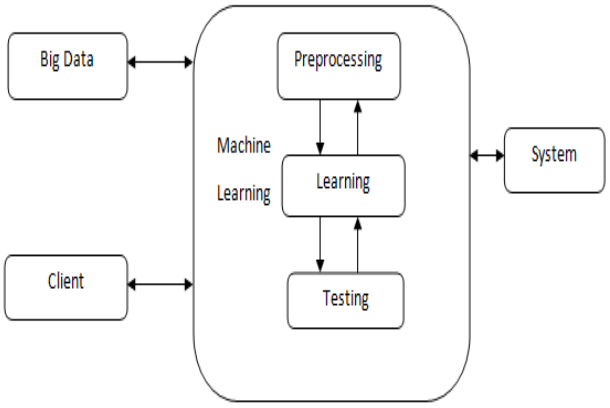

Figure 1. Relation of Machine Learning and Big data

As in the above figure relation between machine learning components and big data has been shown. The massive data is the great source as input for learning components that helps to generate a desired output, which is again a part of enormous data. Users interact with learning components by using various protocols and feedback after that, system has great effect on how learning methods should work efficiently.

IV. TAXONOMY OF MACHINE LEARNING

Machine learning growing in field of performance, prediction and theory. In this era it is using with the AI(artificial intelligence, statistics, optimal control, mathematics and many more. As a result of its usage in an extensive variety of applications, ML is used in each and every scientific field[6]. Machine Learning methods have brought major impact on society and technology. It has been used in different fields like data mining, internet of things, face and speech recognition. Generally machine learning is classified into three learning methods as shown in figure 2.

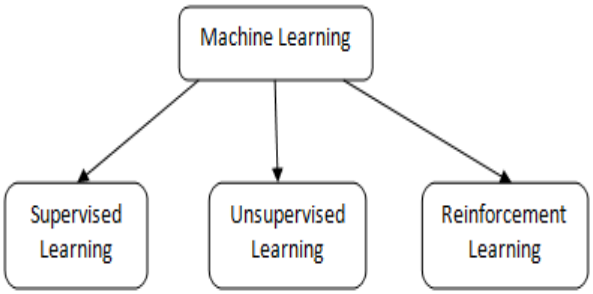

Figure 2. Classification of Machine Learning

A. Supervised Machine Learning

The learning approach, which takes labeled input to train the data and it provides the desired output. In this learning approach output is predefined. This learning method is basically used , when one wants to classify the data and arrange the data in some order. This approach performs classification, estimation and regression. Its basic principle is task driven. It follows the basic algorithm based on input and output. The main examples of supervised learning method is Naive Bayes, SVM, Random Forest, Linear Regression and many more[7].

B. Unsupervised Machine Learning

The learning approach in which there is no need of labeled input to train the data. It takes inputs from the environment. In this approach output is not predefined. The main function is performed by this learning is clustering. It can be perform on actual and synthetic data. This approach is basically based on data driven. This approach is basically used when no information about the method of classification is provided. In this approach output is not predefined. The main examples of unsupervised learning is K-means and X-means[8].

C. Reinforcement Learning

The learning approach which works on feedback taken from the environment is known reinforcement learning. It's main function is decision making. It is used when there is no idea to classify but it tries to classify is appreciable. This approach is algorithm driven. It requires real data to perform testing and prediction. The main examples of Reinforcement learning are Q-learning and TD-learning[9].

According to the above categorization, there are three main learning methods known as traditional learning methods for handling massive data. As Google[10] was also using the supervised and unsupervised learning approach for data collection and in various applications like search engines, Google maps and translators. Therefore, the supervised and unsupervised learning are used for data analysis and reinforcement learning is used for decision making. Some of the main features of machine learning are shown in table1.

Table 1. Some Features of Different Machine Learning Methods

Machine Learning Approaches	Unsupervised Machine Learning		Supervised Machine Learning			Reinforcement Machine Learning	
	K-means [16]	*X-means [17]*	*Naive Bayes [11]*	*Support Vector Machine [12][13]*	*Neural Networks[14] [15]*	*Q-Means[18]*	*TD-Learning[19]*
Classification	✔	✔	✔	✔	✔	×	×
Estimation	×	×	✔	✔	✔	×	×
Regression	×	×	✔	✔	✔	✔	✔
Clustering	✔	✔	×	×	×	×	×
Decision Making	✔	✔	✔	✔	✔	✔	✔
Data Analyis	✔	✔	✔	✔	✔	×	×
Prediction	✔	✔	✔	✔	✔	✔	✔

V. ADVANCED MACHINE LEARNING METHODS

In the above section we have discussed some traditional learning methods, sometimes these methods unable to provide better throughput. So in this section, various modern learning methods have been discussed these are somehow helpful for handling massive data. Some of the learning methods are:

A. Deep Learning

These days, Deep learning is one of the more challenging topic. In compared to traditional learning methods deep learning follows the combination of supervised and unsupervised machine learning methods. This learning strategy works in the architecture that learns the hierarchal representation. Based on deep learning some the hottest research topics are deep belief networks and convolution neural networks[20]. Deep learning strategy is considered as a best way of learning method for various technologies like face recognition. Deep learning helps to work with large datasets because of its maximum processing speed and it works with graphics processors. Some the systems works in deep learning are: IBM's Computer[21].

B. Distributed Learning

The massive data is difficult to handle as it needs more processing speed. The massive data contains some critical information that needs significant technologies in field of science and engineering. For this situation a learning method that have power to handle whole data at one time is requires[22]. As compared to traditional learning methods, distributed learning plays better role. In this strategy whole data is processed in distributed manner. Various traditional methods like decision tree, meta learning are also based on distributed strategy but with the use of distributed learning method it helps to handle massive data. The systems works on distributed learning method are: distributed computing[23].

C. Deep Learning

Active learning is very helpful for real time applications. The massive data is unlabeled data in general, that is time consuming to handle[24]. Active learning performs better to process unlabeled data and gives high accuracy. This learning is helpful in science and medical field. The systems that works based on active learning strategy are: DNA identification[25].

VI. MACHINE LEARNING TOOLS SELECTION

There are different machine learning tools like Hadoop, Map reduce, YARN and many more. But to perform machine learning methods in Hadoop, user need not to bother with any special library. For the interaction with hadoop and map reduce, a person needs to have programming skills. As discussed above, many tools come into existence for machine learning strategies but some engineers rejected them due to lack of advance features and need more expertise. Engineers have not worked more on selection of machine learning tools as compared to machine learning algorithms development. Some the steps that can be followed to select machine learning tools are as follows:

A. Criteria To Select Tool

In a machine learning environment, there are various factors to select a tool to perform machine learning technique. The following we have discussed basic main factors for selecting tools:

- **Speed:** To do any work efficiently, speed matters a lot. The platform which we are using effects the speed due to the libraries imported. But every time speed could not be considered while choosing machine learning tool. If the system don't require any dynamic update, a batch entry is preferred for its simplicity.
- **Scalability:** This factor is considered more important in field of size and complexity while selecting any tool. One ought to think about what their information looks like currently and additionally what information they may work with later on, with the end goal to decide whether a specific toolbox will be proper. Scalability should be taken in both the directions, as portion of best tools for massive data perform inadequately on small datasets. This feature is also considered in case of dimensionality also.
- **Coverage:** This refers to the scope of choices contained in the toolbox in terms of various machine learning strategies and their implementation. None of the accessible tool for enormous information give a selection as comprehensive framework like weka. As many of the tools are hard to set up and learn, it is imperative to think about future needs and also current.

VII. TERMINOLOGIES FOR MACHINE LEARNING ALGORITHM:

Classification is the process of finding a model that describes and distinguishes data class labels. A model is derived based on the analysis of a set of training data (objects for which the class labels are known). It is used to predict the class label of objects for which the class label is unknown. A well-known classifier, decision tree, is a flow chart like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represent classes or class distributions. In binary classification, there are two possible output classes. One common example of classification is spam detection in an email inbox, where an input set is an email received along with its metadata (sender's name and sent time) and the output label is either spam or not spam. A researcher use generic names for the two classes known as positive and negative or class 1 or 0. There are different ways of evaluating the performance of classification: Accuracy, confusion matrix, precision and recall are widely used evaluation metrics.

A. Accuracy

Accuracy is the number of correctly classified instances. It is characterized as the proportion of number of right predictions to the total number of predictions. In any case, if the test information isn't adjusted(when a large portion of the cases or records have a place with one of the classes), or one is more interested in the performance on either one of the classes, the accuracy metric will not be able to capture the adequacy of classifier.

B. Confusion Matrix

Confusion matrix is a table used to find the performance of the classification system. Four terms TP(true positive), TN(True negative), FP(False positive), FN(False negative) are main elements of confusion matrix. whereas TP define the cases predicted true as they are true. TN define the cases predicted false as they are actually false. FP define the cases true but in actual they are false. FN define the cases false but in actual they are true.

	Prediction False	Prediction True
Actually False	True Negative	False Positive
Actually True	False Negative	True Negative

Figure 3. Confusion Matrix

VIII. CHALLENGES AND FUTURE SCOPE

While critical advancement has been made in the most recent decades toward achieving definite objective of sensing the massive data by using machine learning methods, the consequence is that we are still not exactly there. The productive pre-processing components to make the learning framework able to manage massive information and effective learning advancements to discover the guidelines to portray the information are still of critical need. In this way the portion of open issues and challenges research areas are discussed below:

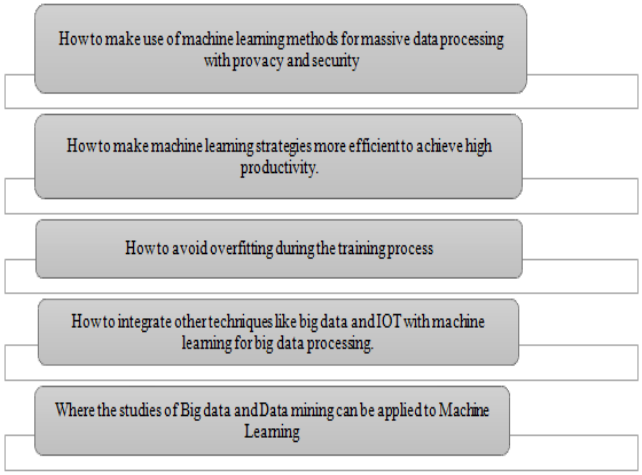

Figure 4. Challenges of Machine Learning

IX. CONCLUSION

Massive data is quickly extending in all the fields like science, medical and engineering. Prediction and testing of this big data brought many opportunities and challenges. But according the above survey, the traditional machine learning methods are not efficient for handling big data as compared to modern learning methods. The advanced machine learning methods are more scalable and efficient to handle the data with its features like velocity, volume and many more. Accordingly machine learning requires to reinvent itself for handling more data. As this paper began with some introduction of machine learning and further proceed to advance machine learning methods then ends with challenges of learning with massive data.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, "*Machine learning: Trends, perspectives, andprospects*," Science, vol. 349, pp. 255-260, 2015.
[2] J Gantz, D Reinsel, *The digital universe decade—are you ready* (EMC,Hopkinton, 2010)
[3] J Manyika, M Chui, B Brown, J Bughin, R Dobbs, C Roxburgh, AH Byers, *Bigdata: the next frontier for innovation, competition, and productivity* (McKinseyGlobal Institute, USA, 2011)

[4] Q Wu, G Ding, Y Xu, S Feng, Z Du, J Wang, K Long, *Cognitive internet ofthings: a new paradigm beyond connection.* IEEE Internet Things J1(2), 129–143 (2014)

[5] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*:Cambridge University Press, 2011.

[6] Jyoti Arora, Ambica Sood, "*A Survey on knowledge extraction Approaches from Big Data and Rectifying Misclassification strategies*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.12, pp.187-200, 2017.

[7] O. Okun, G. Valentini, (Eds.), Supervised and Unsupervised Ensemble Methods and their Applications Studies in Computational Intelligence, vol. 126, Springer, Heidelberg, 2008.

[8]] Nelles, Oliver. "Unsupervised Learning Techniques." In *Nonlinear System Identification*, pp. 137-155. Springer Berlin Heidelberg, 2001.

[9] Burch, Carl. "A survey of machine learning." *Tech. report, Pennsylvania Governor's School for the Sciences* (2001).

[10] Jones, Nicola. "The learning machines." *Nature* 505, no. 7482 (2014): 146.

[11] Mitchell, Tom M. "Machine learning. WCB." (1997).

[12] Zheng, Jun, Furao Shen, Hongjun Fan, and Jinxi Zhao. "Anonline incremental learning support vector machine for large-scale data." Neural Computing and Applications 22,no. 5 (2013): 1023-1035.

[13] Cortes, Corinna, and Vladimir Vapnik. "Support-vectornetworks." Machine learning 20, no. 3 (1995): 273-297.

[14] Burch, Carl. "A survey of machine learning." Tech. report,Pennsylvania Governor's School for the Sciences (2001).

[15] Dong, Xu, Ying Li, Chun Wu, and Yueming Cai. "Alearner based on neural network for cognitive radio."In Communication Technology (ICCT), 2010 12th IEEEInternational Conference on, pp. 893-896. IEEE, 2010.

[16] Safatly, Lise, Mario Bkassiny, Mohammed Al-Husseini,and Ali El-Hajj. "Cognitive radio transceivers: RF, spectrum sensing, and learning algorithmsreview." International Journal of Antennas andPropagation 2014 (2014).

[17] Weber, Markus, Max Welling, and Pietro Perona. "Unsupervised learning of models for recognition." In *European conference on computer vision*, pp. 18-32. Springer, Berlin, Heidelberg, 2000.

[18] Galindo-Serrano, Ana, and Lorenza Giupponi. "Distributed Q-learning for aggregated interference control in cognitive radio networks." *IEEE Transactions on Vehicular Technology* 59, no. 4 (2010): 1823-1834.

[19] Sutton, Richard S. "Learning to predict by the methods of temporal differences." *Machine learning* 3, no. 1 (1988): 9-44.

[20] D Yu, L Deng, Deep learning and its applications to signal and information processing. IEEE Signal Proc Mag 28(1), 145–154 (2011)

[21] Y Bengio, Learning deep architectures for AI. Foundations Trends Mach Learn 2(1), 1–127 (2009).

[22] D Peteiro-Barral, B Guijarro-Berdiñas, A survey of methods for distributed machine learning. Progress in Artificial Intelligence 2(1), 1–11 (2012)

[23] H Zheng, SR Kulkarni, HV Poor, Attribute-distributed learning: models, limits, and algorithms. IEEE Trans Signal Process 59(1), 386–398 (2011)

[24] Y.Fu, B Li, X Zhu, C Zhang, Active learning without knowing individual instance labels: a pairwise label homogeneity query approach. IEEE Trans Knowl Data Eng 26(4), 808–822 (2014)

[25] B Settles, Active learning literature survey (University of Wisconsin, Madison, 2010)

Authors Profile

Anamica Teipal pursed Bachelor of Technology in Computer Science from Guru Nanak Dev University , Amritsar in year 2016 .She is currently pursuing Masters of Technology in Computer Science from Guru Nanak Dev University , Amritsar and currently working as Research Scholar in Department of Computer Science . Her main research work focuses on Machine Leaning and Data Mining.

Ms. Kamaljit Kaur pursued Bachelor of Technology and Master of Technology in Computer Science from Guru Nanak Dev University. She is the Gold Medlist in Masters of Technology. She is currebntly pursuing Ph. D. and working as Assistant Professor in Computer Engineering and Technology Department. She has published more than 15 research papers in Ugc Approved Journals. Her main research area focus on Cloud Computing, Parallel Computing and Fault Tolerance.

