A Survey of Machine Learning for Big Data Processing

Vivek Upraity and Prabhsharan Kaur

December 3, 2021

A survey of machine learning for big data processing

Vivek Upraity(20MCA1465)

University Institute of Computing

Chandigarh University, Punjab, India

Vivekupraity94@gmail.com

ER. Prabhsharan Kaur

University Institute of Computing

Chandigarh University, Punjab, India

Prabhsharankaur.uic@cumail.in

Abstract:Machine Learning(ML) approaches have created huge societal consequences in wide range of applications such as computer vision ,speech processing, natural language comprehension, neuroscience,health, and the Internet of Things.Big data has never promised or challenged machine learning algorithms to obtain fresh insights into a variety of corporate applications and human behaviours.On the one hand, big data supplies ML algorithms with unprecedented amounts of data from which to uncover underlying patterns and develop predictive models.Traditional machine learning methods,on the other hand, confront fundamental problems such as scalability in order to effectively unlock the value of big data.With the ever- expanding universe of big data, machine learning must evolve and adapt in order to turn large data into useful intelligence.

Keywords :-**Machine learning ,Big data ,Deep learning, Data , Algorithms**

INTRODUCTION

ML thrives on strong computer environments, efficient learning approaches (algorithms), and rich and/or huge data.

As a result, machine learning has a lot of potential and is an important aspect of big data analytics. In the context of large data and modern computing settings, machine learning techniques are used. We want to look into the benefits and drawbacks of machine learning on huge data. Big data opens up new possibilities for machine learning. The framework is based on machine learning, which is divided into three phases: preprocessing, learning, and evaluation. In addition, the framework includes four other components: big data, user, domain, and system, all of which influence and are influenced by ML. The components of MLBiD and the phases of ML point the way to identifying opportunities and problems, as well as future work in a variety of unknown or underexplored research fields**.**

A Framework of Machine Learning on Big Data

Machine learning is at the heart of MLBiD, and it interacts with four other components: big data, user, domain, and system.Figure 1 depicts the framework for machine learning on large data (MLBiD).Interactions take happen in both directions. Big data, for example, provides inputs to the learning component, which creates outputs that form part of the big data.The learning component may be interacted with by the user by giving subject knowledge, personal preferences, and usability feedback, as well as by utilising learning results to enhance decision making. The way learning algorithms should run and how efficient they should be has an influence on system architecture, and concurrently satisfying the learning demand may lead to a co-design of system architecture.Following that, we'll go through each of the components individually.

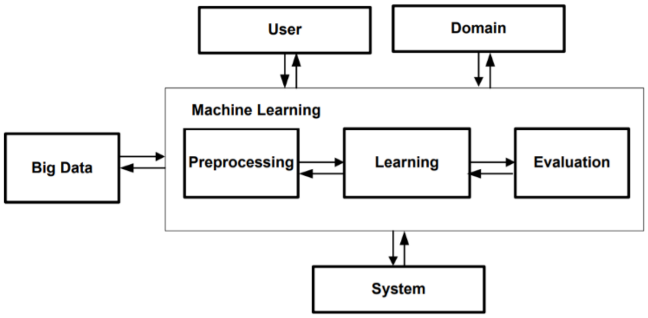


Fig 1 A framework of machine learning on big data (MLBiD)

1.Machine Learning

Data preprocessing, learning, and assessment are common steps in machine learning (see Fig. 1). Data preprocessing aids in the transformation of raw data into the "correct shape" for further learning stages.It's likely that the raw data is unstructured, noisy, incomplete, and inconsistent. Through data cleaning, extraction, transformation,and fusion,the preprocessing phase converts such data into a form that may be utilised as inputs to learning.Using the preprocessed input data, the learning phase selects learning methods and sets model parameters to create desired outputs. Data preparation can be done using several learning methods, notably representational learning.

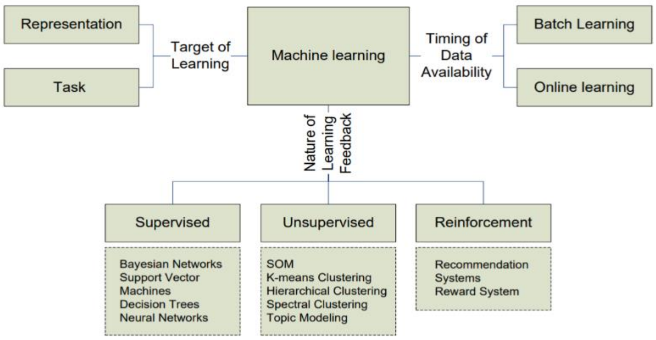


Fig. 2. A multi-dimensional taxonomy of machine learning

The form of learning feedback, the aim of learning activities, and the timeliness of data availability are all characteristics of machine learning. As a result, as illustrated in Fig. 2, we propose a multi-dimensional taxonomy of ML.

2. Big Data

Volume(quantity/amount of data),velocity(speed of data creation),diversity (kind, nature, and format of data),veracity (trustworthiness/quality of collected data), and value are the five dimensions of big data (insights and impact) .Starting at the bottom, we structured the five dimensions into a stack of large, data, and value levels (see Figure 3)

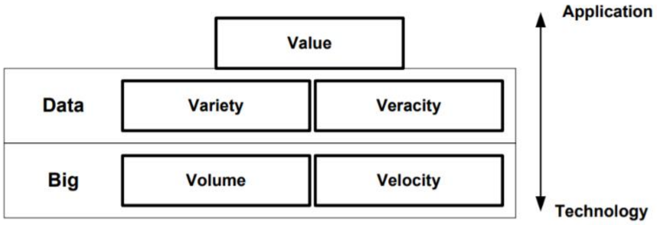


Figure 3. Big Data Stack

3.Other Components

1. Users

Domain experts, end-users, and ML researchers and practitioners are all stakeholders in machine learning systems.Traditionally, ML practitioners have made the majority of decisions in using ML, from data collection to performance evaluation. End-user engagement has been restricted to supplying data labels, answering domain-related questions, and providing feedback on the learnt outcomes, which is generally mediated by practitioners, resulting in lengthy and asynchronous iterations.

2.Domain

Domain knowledge aids machine learning in detecting intriguing patterns that might otherwise be missed by looking at datasets alone There's a chance that the training datasets aren't big enough or representative enough to find all the patterns. Obtaining sufficient and representative data is also costly,if not impossible, due to wide domain variance and application-specific needs.

3.System

The system architecture, often known as the platform, is a combination of software and hardware that creates an environment in which machine learning algorithms may execute. A multi-core computer with distributed architecture, for example, is predicted to increase ML efficiency when compared to simpler alternatives. To solve the problems of large data, new frameworks and system architectures such as Hadoop/Spark have been developed.

4.Data Preprocessing Opportunities and Challenges

The design of preprocessing pipelines and data transformations that result in a data representation that can enable successful ML takes up a large portion of the actual effort in implementing an ML system.Data preparation tries to handle difficulties such data redundancy,inconsistency, noise, heterogeneity, transformation, labelling,data imbalance, and feature representation/selection, among others.Due to the need for human work and a vast number of alternatives to choose from, data preparation and preprocessing is generally expensive.

4.1 Data Redundancy

When two or more data samples reflect the same object,duplication occurs.Data duplication or inconsistency can have a significant influence on machine learning. Despite the development of a number of approaches for detecting duplicates over the last two decades, classic methods such as pairwise similarity comparison are no longer viable for huge data.

4.2 Data Noise

Data sparsity, missing and erroneous values, and outliers may all cause noise in machine learning. When dealing with huge data, traditional solutions to noisy data problems confront obstacles.Manual techniques, for example, are no longer viable owing to their lack of scalability; replacing them with a mean would lose the benefits of big data's richness and fine granularity

4.3 Data Heterogeneity

Big data promises to provide multi-view data from a variety of sources, in a variety of formats, and from a variety of population samples, and hence is very heterogeneous. The relevance of this multi-view heterogeneous data (e.g., unstructured text, audio, and video forms) for a learning task may vary.

4.5 Data Discretization

Decision trees and Nave Bayes are two ML techniques that can only deal with discrete characteristics. Discretization converts quantitative data into qualitative data, resulting in a non overlapping domain split.

4.6 Data Labeling

Traditional data annotation methods need a lot of time and effort. To deal with the problem of large data, several different approaches have been proposed.For example, online crowd-sourced sources can provide free annotated training data with a wide range of class numbers and intra-class variation.Furthermore, probabilistic programme induction may be used to accomplish human-level idea learning.

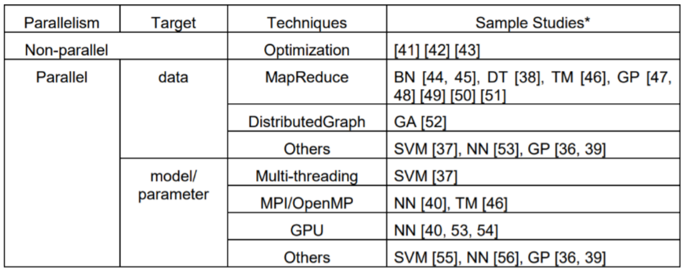
4.7 Imbalanced Data

Traditional stratified random sampling methods have addressed the problem of unbalanced data.However, if iterations of subsample creation and error metrics computation are required, the procedure might take a long time.Furthermore, standard sample algorithms are unable to enable data sampling across a user-defined subset of data, which includes value-based sampling.

5. Learning Opportunities and Challenges

Prior to the emergence of the "big data" era, developing scalable machine learning algorithms capable of handling enormous datasets was a long-standing research subject in the ML community.

We categorise research in the taxonomy based on whether or not parallelism is addressed in their algorithms/platforms.Non-parallelism approaches strive for significantly quicker optimization methods that can cope with large amounts of data without using parallelism.Traditionally,ML scalability has been primarily focused on building new algorithms that can run considerably more efficiently (e.g with greatly reduced time and/or space complexity).



Parallelism	Target	Techniques	Sample Studies*
Non-parallel		Optimization	[41] [42] [43]
Parallel	data	MapReduce	BN [44, 45], DT [38], TM [46], GP [47, 48] [49] [50] [51]
		DistributedGraph	GA [52]
		Others	SVM [37], NN [53], GP [36, 39]
	model/ parameter	Multi-threading	SVM [37]
		MPI/OpenMP	NN [40], TM [46]
		GPU	NN [40, 53, 54]
		Others	SVM [55], NN [56], GP [36, 39]

Table : A taxonomy of machine learning algorithms/platforms for big data

5.1 Non-parallelism

Most machine learning techniques are built on optimization.Combinatorial optimization (greedy search, beam search, branch-and-bound) and continuous optimization are two types of traditional optimization methods .Unconstrained optimization (e.g., gradient descent, conjugate gradient, quasi-Newton techniques) and constrained optimization (e.g., linear programming, quadratic programming) are two types of optimization.

5.2 Data Parallelism

To gain scalability, existing machine learning models might make use of big data approaches. These initiatives can be divided into two groups.One solution is to provide a generic middleware layer that reimplements current learning tasks so that they can be executed on a large data platform like Hadoop or Spark.A middleware layer like this frequently includes general primitives/operations that are beneficial for a variety of learning activities.

5.3 Models/parameter Parallelism

A lot of work has gone into figuring out how to parallelize machine learning algorithms or how to provide performance guarantees on various parallelized methods.Many machine learning algorithms are at best trivially parallel [66-68], therefore these efforts are justified. Furthermore, large data machine learning is not merely a scaled-up version of small data machine learning.To handle the accompanying technical issues, it necessitates new formulations and algorithms.The roots of parallelization of learning algorithms can be found in distributed and large-scaled machine learning.

5.3.1 Distributed machine learning

In large-scale ML, distributed ML can naturally overcome the problem of algorithm complexity and memory restriction. Distributed ML scales up learning algorithms by spreading the learning process across numerous computers or processors and addressing a distributed optimization issue to solve the inability of learning algorithms to use all of the data to learn in an acceptable amount of time. Distributed machine learning can provide not just efficiency but also fault tolerance by duplicating data across machines

5.3.2 Deep Learning

Deep neural network-based learning has recently emerged as one of the most rapidly growing and intriguing fields of big data learning.Neural networks are a class of models based on biological neural networks, which are made up of interconnected neurons with connections that may be modified and adapted to inputs.Deep neural networks are essentially neural networks with a large number of hidden layers, or deep-layered architecture, with each layer performing a nonlinear transition from its input to output.

5.4 Hybrid Approaches

Hybrid approaches integrate model and data parallelism by dividing both data and model variables at the same time, in addition to the two methods for ML on big data. This not only allows dispersed clusters to train quicker, but it also allows ML applications to run efficiently when both the data and the model are too huge to store in a single machine's memory.

Conclusion

The opportunities and problems of machine learning (ML) on massive data are discussed in this study.Big data gives new prospects for inspiring transformational and unique ML solutions to handle many associated technical difficulties and produce real-world consequences, while also posing several challenges for classical ML in terms of scalability, adaptability, and usability.These opportunities and difficulties can be used to guide future study in this field. The majority of existing work on machine learning for big data has concentrated on volume, velocity, and diversity, but nothing has been done to address the other two components of large data: truth and value.To deal with data veracity, one promising direction is to develop algorithms that can assess the trustworthiness or credibility of data or data sources, allowing untrustworthy data to be filtered out during pre-processing; another promising direction is to develop new machine learning models that can infer with unreliable or even contradictory data .

In summarise, machine learning is required to handle the problems offered by big data and to identify hidden patterns, knowledge, and insights in order to turn the latter's promise into practical value for business decision-making and scientific exploration. The combination of machine learning and big data speaks to a bright future in a new frontier.

References

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, pp. 255-260, 2015.

[2] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," Journal of Big Data, vol. 2, pp. 1-32, 2015// 2015

[3] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar,N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," Journal of Big Data, vol. 2, pp. 1-21, 2015.

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

[5] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed.: Prentice Hall, 2010.

[6] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, 2013.

[7] O. Dekel, "From Online to Batch Learning with Cutoff-Averaging," in Ofer Dekel. NIPS, page 377-384. Curran Associates, Inc., (2008), NIPS, 2008, pp. 377-384.

[8] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the People: The Role of Humans in Interactive Machine Learning," AI Magazine, vol. 35, pp. 105-120, 2014.

[9] J. E. Mason, I. Traoré, and I. Woungang, Machine Learning Techniques for Gait Biometric Recognition: Using the Ground Reaction Force. Switzerland: Springer, 2016.

[10] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," Journal of Big Data, vol. 2, pp. 1-36, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton, Imagenet classification with deep convolutional neural networks, 2012.

[12] S. Zhou, Q. Chen, and X. Wang, "Active deep learning method for semi-supervised sentiment classification," Neurocomputing, vol. 120, pp. 536-546, 11/23/ 2013

[13] L. Breiman, "Pasting Small Votes for Classification in Large Databases and On-Line," Machine Learning, vol. 36, pp. 85-103, 07/01/1999 1999

[14] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," Journal of Big Data, vol. 2, pp. 1-20, 2014// 2014

[15] T. R. Armes, M., "Using Big Data and predictive machine learning in aerospace test environments," in IEEE AUTOTESTCON, 2013.