Journal of The Gujarat Research Society

ISSN: 0374-8588
Volume 22 Issue 1, January 2020

Machine Learning Methods for Big Data Processing Research

[*]Dr. Kulvinder Singh, [#]Ms. Nivedita

[*] *Assistant Professor (Computer Scienc), Tantia University, Sri Ganganagar(Raj.), India*

[#] *Ms.Nivedita (Scholar, Computer Scienc), University, Sri Ganganagar(Raj.), India*

Abstract: Machine learning has been widely used and applied in a variety of fields in our lives. However, as the big data era approaches, certain traditional machine learning algorithms will be unable to meet the demands of real-time data processing for massive datasets. As a result, machine learning must reinvent itself for the age of big data. In this paper, we examine recent research on machine learning for big data processing. First, a description of big data is offered, followed by an examination of novel machine learning properties in the context of large data. Then, using machine learning approaches, we present a workable reference framework for dealing with massive data. Finally, a number of research obstacles and unanswered questions are discussed.

Introduction

Machine learning is a topic of study that aims to understand the computational principles through which experience can lead to increased performance without having to be explicitly programmed. It is a very multidisciplinary field that draws on ideas from a wide range of disciplines. Machine learning has pervaded practically every aspect of our lives in recent decades, and it is now so common that you probably use it hundreds of times a day without even realizing it. It largely affects the rest of the world through its use in a wide range of applications, which has had a significant impact on science and society. In the last few decades, a slew of machine learning algorithms have been devised, including neural networks, decision trees, support vector machines, k-nearest-neighbor, genetic algorithms, and Q-learning. Pattern recognition, robotics, natural language processing, and autonomous control systems are just a few of the areas where they've been applied. Machine learning is a form of mathematics that uses statistical algorithms to analyse enormous amounts of data from a variety of sources. However, as the era of big data approaches, the gathering of data sets will become so enormous and complicated that traditional data processing methods and models will struggle to cope. As a result, several classic machine learning algorithms are ineffective in this situation and cannot meet the demands of real-time data processing and storage. As a result, we must investigate new approaches for analysing and dealing with large amounts of data using distributed storage and parallel processing. Scholars previously focused on two aspects of research: I designing a distributed parallel computing framework or platform for quickly dealing with big data, such as MapReduce, Dryad, Graphlab, Haloop, and Twister, and so on; and ii) proposing new algorithms to solve a class of determined big data problems.

He Q et al., for example, used a parallel extreme learning machine based on MapReduce to solve regression problems. To deal with incomplete streaming large data, the authors created a low-complexity subspace learning system. Dictionary learning was also used by some academics to represent huge data in a sparse way. To far, however, there have been few systematic and in-depth analyses of the new properties of machine learning in the age of big data, as well as the related machine learning-based strategies for dealing with large

data. As a result, the focus of this paper is on studying machine learning-based methods for handling huge data and developing a reasonable framework model for big data processing. This article's primary work can be summarized as follows:

- First, we'll go through a quick overview of big data and describe five essential terms that define it: volume, variety, velocity, veracity, and value.
- We then examine the new aspects of machine learning in the context of big data in a systematic and in-depth manner. Several potential strategies for dealing with large data issues are also presented.
- Finally, we create a reference framework for rapid big data processing that combines machine learning with the power of distributed storage and parallel computation.

Big Data: A Quick Overview

We now live in a data flood era, with massive amounts of data accumulating in every area of our lives. Data streams from several fields contribute to the burgeoning big data paradigm. Among the large volume and variety of data, it may be a tremendous chance for the big data scientist. Big data has the ability to revolutionize our society and improve our quality of life by uncovering associations, evaluating patterns, and anticipating trends within the data. Big data is usually defined as one of three forms of data derived from physical, cyber, and social sources:

- **Data from nature on Earth:** We may anticipate that data from nature on Earth, such as satellite data from space, will be a great potential data source.
- **Life data:** studying the biological body is a major endeavour, and exploring the human body, in particular, still has a lot of obstacles, such as biological data.
- **Sociality data:** With the rapid growth of digital mobile devices and networks, enormous volumes of sociality data, such as voice and video data, are generated every day in our lives.

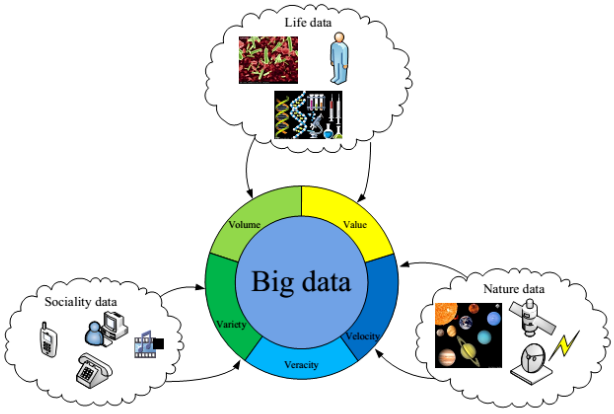

Fig. 1. Big data types and characteristics.

Big data is defined by five terms, as illustrated in Fig. 1: volume, variety, velocity, veracity, and value. We'll go through each attribute in detail in the sections that follow.

Capacity:- The key attribute of big data is volume, which refers to the magnitude of the data. It is an undeniable fact that massive amounts of data are continuously generated at unprecedented scales from a variety of disciplines in our lives. The steady flow of fresh data, which is accumulating at unprecedented rates, poses significant difficulties to traditional processing infrastructure in terms of effective data capture, storage, and manipulation. It requires high scalability of data management and mining tools.

Variety:- The term "variety" refers to the various forms of data. Big data is derived from a variety of sources and comes in a variety of formats, including structured, unstructured, and semi-structured representations. The

great issue in mining such a varied dataset is perceptible; building a single model will not result in good-enough mining outcomes. Specialized, more complicated and multi-model systems are likely to be built.

Velocity:- is a term used to describe the speed with which something moves. In general, every day, unprecedented data is continuously generated in the form of streams that must be analysed in real time or at a quick rate. We must do specific duties within a specified amount of time during special times, or the processing results will become less valuable, if not worthless. The essential idea for addressing this problem is to develop parallel processing algorithms for handling data in parallelization.

Reliability:- It can be defined as data precision. We may acquire data from several domains with incomplete information in the era of big data with a high chance. The quality of data is substantially influenced by these incomplete, imprecise, and dynamic data sources from a variety of sources. As a result, the source data's accuracy and reliability quickly become a major cause of concern. Data validation and provenance tracking are becoming increasingly critical for data processing systems to overcome this problem.

Price:- The rapid development of artificial intelligence, machine learning, and data mining technologies has fueled the rise of big data, which involves analyzing data for information, extracting that information into knowledge, and facilitating decision and action for acquiring desired values based on that knowledge. In terms of big data, valid values are likely to be found like panning for gold in the sand. As a result, how to apply strong machine learning algorithms to accomplish faster data value purification has become an urgent topic to solve in the current big data environment.

While big data has many potential, it also poses a number of unanticipated obstacles.

Traditional data management solutions cannot store, analyse, or process it, necessitating the development of new workflows, platforms, and architectures. Researchers are becoming interested in the topic of machine learning, which may be used to perform tasks such as prediction, classification, and association over enormous volumes of data. However, as the big data age approaches, several big data properties will pose significant hurdles to existing machine learning methods. As a result, machine learning will need some new features to deal with the issues that massive data will bring. These new performances must be thoroughly explored and analyzed on a systematic level.

Machine Learning with Big Data: New Features

Machine learning, in comparison to traditional learning systems and methodologies, must possess several additional qualities in order to deal with the potential issues offered by large data. In this section, we'll go over three features of machine learning approaches that are useful for dealing with huge data problems in depth, including sparse representation and feature selection, mining structured relations, high scalability, and high speed.

Sparse Representation and Feature Selection In big data scenarios, datasets with high-dimensional features are becoming more widespread. It is challenging to manage high-dimensional data using typical data processing methods. As a result, effective dimension reduction is becoming widely recognized as a vital step in resolving these issues. We cover feature selection and sparse representation methods for machine learning techniques in terms of high-dimensional big data, which are two extensively, used approaches in dealing with high-dimensional data.

Through the process of picking a subset of important features, feature selection is a key challenge in creating robust data processing models. Many sparse-based supervised binary feature selection approaches can be expressed as a simplified version of the following problem :

$$< \mathbf{w}^*, b >= \min_{\mathbf{w},b} \left\| \mathbf{y} - X^T \mathbf{w} - b\mathbf{1} \right\|_2^2, \quad s.t. \left\| \mathbf{w} \right\|_0 = k \tag{1}$$

where b is the learned biased scalar, $\mathbf{1} \in \mathbb{R}^{n\times 1}$ is a column vector with all 1 entries, $\mathbf{w} \in \mathbb{R}^{d\times 1}$ is the learned model, $X \in \mathbb{R}^{d\times n}$ is the training data, $\mathbf{y} \in \mathbb{R}^{n\times 1}$ is the binary label, and k is the number of the feature selected. While the multi-class feature selection is to learn the the bias $\mathbf{b} \in \mathbb{R}^{m\times 1}$ and projection matrix $W \in \mathbb{R}^{d\times m}$, and the function can be expressed as [16]:

$$< W^*, b >= \arg\min_{W,b} \sum_{i=1}^{n} \left\| \mathbf{y_i} - W^T \mathbf{x_i} - \mathbf{b} \right\|_2^2, \tag{2}$$

where $\{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^{d\times 1}$ are training data and $\{y_1, y_2, \cdots, y_n\} \in \mathbb{R}^{m\times 1}$ are the corresponding class labels. For some datasets with extremely large data dimension, feature selection is very necessary and useful to reduce the redundancy of features and alleviate the curse of dimensionality.

How to represent a big data set is another fundamental problem in dealing with high dimensional data. It should be able to help visualize the data, to construct better statistical models, and to improve prediction accuracy through mapping the high dimensional data into the underlying low dimensional manifold. And for high-dimensional big data, a sparse data representation is more and more important for many algorithms. Recent years have witnessed a growing interest in the study of sparse representation of data. In [15], the authors introduced the K-SVD algorithm for adapting dictionaries so as to represent data sparsely. Some optimization algorithms based on K-SVD algorithm have been also gradually proposed, such as the incremental K-SVD (IK-SVD) algorithm [12], distributed dictionary learning method [13], etc.. Through applying these methods, machine learning can achieve appropriate data representation for many big data processing tasks. With the power of feature selection and sparse representation, machine learning systems can better deal with high-dimensional big data by means of dimensionality reduction.

Mining Structured Relations. Big data is generally from different sources with obviously heterogeneous types including structured, unstructured and semi-structured representation forms. Dealing with such a heterogeneous dataset, the great challenge is perceivable, thus machine learning system needs infer the structure behind the data when it is not known beforehand. One way of structuring data is to discover the relevance based on inherent data properties through structured learning and structured prediction.

Structured machine learning refers to learning structured assumption from data with rich internal structure usually in the form of different relations [17]. In many structured learning problems, the primary inference task is to compute the variable F and F can be defined as follows [17]:

$$F = \arg\max_{Y} \Phi(X, Y; \Theta), \tag{3}$$

where X and Y are the input structure and output structure respectively, and Θ are the parameters of the scoring function Φ. In terms of structured prediction, several frameworks have been developed in the past, such as conditional random fields (CRFs), structured support vector machines (SSVMs), and their generalizations [16]. In order to design a feasible structured prediction model, we are given a data set $D = \left\{ (x_i, s_i)_{i=1}^{N} \right\}$ for training, where $x_i \in \chi$ denotes the input space object and $s_i \in S$ represents structured label space object. Further, $\phi : \chi \times S \to \mathbb{R}^F$ denotes the F-dimensional feature space. When using structured prediction methods, our interests are generally to find the parameters $w \in \mathbb{R}^F$ of a log-linear model $p_w(s \mid x) \propto \exp(w^{\mathrm{T}} \phi(x, s) / \varepsilon)$ with covariance ε [18]. In order to find the model parameter w which best describs the possible labeling $s_i \in S$ of $x_i \in \chi$, we can construct a

task loss $\ell_{(x,s)}(\hat{s})$ that measures the fitness of any labeling $\hat{s} \in S$. After training, our main purpose is to minimize the negative loss-augmented data-log-posterior [18]:

$$\Re = \min_{w} \sum_{(x,s)\in D} \varepsilon \ln \sum_{\hat{s}\in S} \exp\left(\frac{\ell_{(x,s)}(\hat{s}) + w^T \phi(x,\hat{s})}{\varepsilon}\right) - v^T w + \frac{C}{p}\|w\|_p^p, \quad (5)$$

where vector $v = \sum_{(x,s)\in D} \phi(x,s)$ denotes the empirical mean. In [18], the authors also proposed an optimization method, namely distributed structured prediction learning algorithm for large scale models, which can effectively handle the computation time and the memory demands problems for big data scenarios. The main purpose of mining structured relations from a set of data is to aggregate massive amounts of data and divide it into smaller chunks which can be easily handled by machine learning systems.

High scalability and performance. The massive amounts of big data necessitate a high level of scalability in their data mining and processing capabilities. The strategies used to improve the scalability of machine learning algorithms in current research mostly focus on the following two aspects: I Cloud computing's scalability allows for the analysis of massive datasets by combining diverse workloads with varying performance goals into multi-tenanted computing clusters.

Machine learning using cloud computing is more efficient and performs better when processing and analysing large amounts of data; ii) distributed storage and parallel computing have aided in the scalability of machine learning algorithms. HDFS (Hadoop Distributed File System) is a distributed storage system that runs on commodity hardware clusters and is meant for storing very big data files. MapReduce is a programming paradigm that allows for huge scalability by processing data in parallel. MapReduce, on the other hand, has a clear flaw in that it does not support iterations, which limits the performance of machine learning algorithms. Graphlab, Twister, iMapReduce, and i2MapReduce are some examples of extending studies for efficient iterative computing. The scalability of machine learning algorithms can be considerably increased by combining cloud computing and distributed-parallel frameworks.

Fast access to and mining of large amounts of data are also crucial capabilities for machine learning approaches. Scalability is similarly affected by speed; solving one of them will aid the other. The speed with which data is processed is determined by two key factors: the time it takes to retrieve the data and the effectiveness of the learning algorithms themselves. We must do specific duties within a specified amount of time during special times, or the processing results will become less valuable, if not worthless. Stock market forecasting, earthquake forecasting, and agent-based autonomous trading systems are just a few examples of real-time requests. Real-time computation and online implementations of machine learning algorithms are required for time-sensitive applications that demand real-time reaction and processing. Maximizing the possible parallelism in machine learning algorithms is a useful strategy for speeding up massive data processing. Machine learning can manage large amounts of data because of its scalability and speed.

Discussions. Sparse representation and feature selection, in terms of these new features outlined above, are primarily focused at the characteristics of large dimension for big data through effective dimensionality reduction. The approach of inferring the structure behind the data to divide huge volumes of data into smaller chunks is primarily focused on the heterogeneous natures of big data. High scalability and speed, on the other hand, are aimed at the large-scale and real-time aspects of big data, respectively. Machine learning techniques

in the age of big data must have these capabilities in order to process large data challenges successfully and efficiently.

A Machine Learning-Based Reference Framework for Big Data Processing

Machine learning techniques are becoming more connected with big data as they can progressively reach human-like learning. In this part, we present a reference framework model for big data processing that combines the power of distributed storage and parallel computing with machine learning.

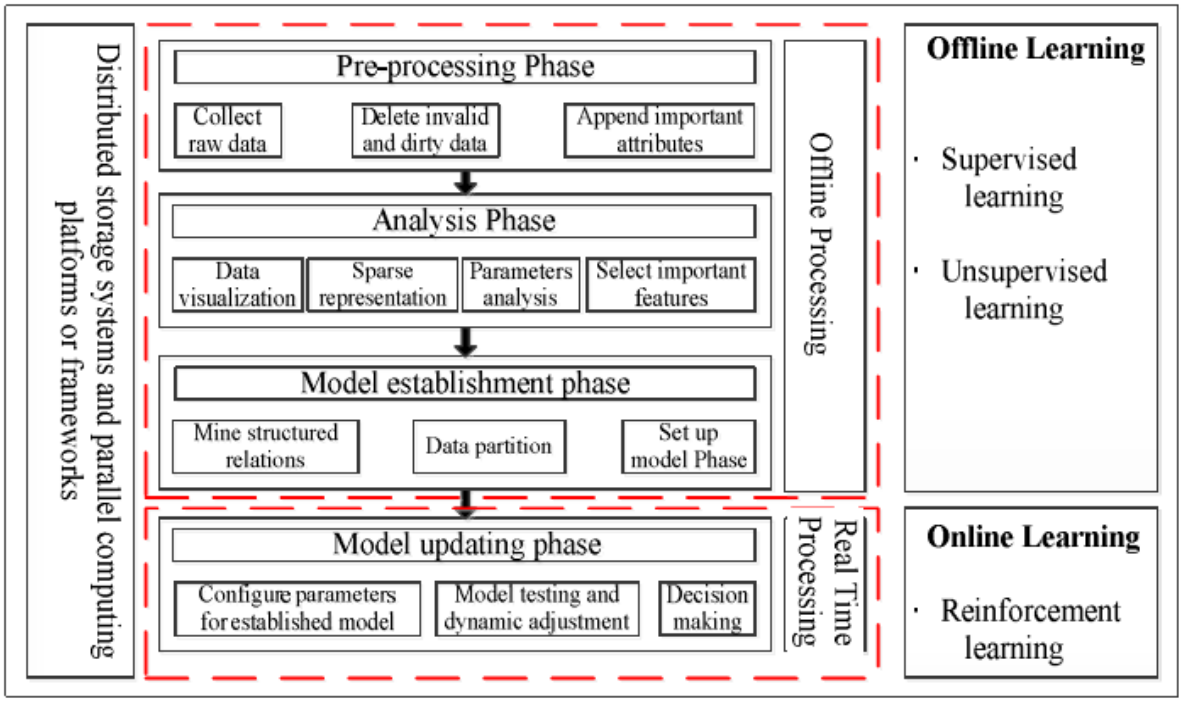

Fig. 2. The proposed reference framework based on machine learning for big data processing.

We assume that the big data processing approach consists of four steps, as shown in Fig. 2: pre-processing phase, analysis phase, model establishment phase, and model updating phase. Raw big data collection from the environment is quite complex and has a lot of redundancy because data sources practically span all kinds of fields. As a result, in the pre-processing phase, we must first eliminate the invalid and filthy data. Furthermore, in real life, we are frequently confronted with large amounts of ambiguous and partial data, and we must add certain key qualities to increase their processing feasibility. We must examine these legitimate and valuable data after the raw data pre-processing phase to learn how to use the data through trial and error. We may use sparse representation to provide effective dimension reduction for high-dimensional data. Data visualization is a basic problem in the analysis of big data, and we can use it to solve it. We should be able to select some critical features through essential parameter analysis in order to develop a feasible model for dealing with real-world challenges. In terms of the model-building phase, we first attempt to mine the structured relationships between data in order to extract statistical information and trends, and then divide the data into training and testing sets. Finally, we may determine what type of model should be developed for use and construct the appropriate model. To verify the performance of the big data processing model, we must configure parameters for the model and apply the created model received from the model establishment phase into actual activities. In this phase, we highlight the importance of real-time input data. Based on the impacts of model application, we should make dynamic adjustments to update the model.

The first three phases of the big data processing approach are offline processing. We can use offline learning approaches in these phases, which fall into two categories: supervised learning and unsupervised learning. We

concentrate on the real-time characteristic of input data during the model testing and updating phase. Online learning approaches are required to address the difficulty of real-time processing, and reinforcement learning is favoured.

Although research into classic machine learning approaches has progressed to a point where it is rather mature, some advanced learning methods are needed to compensate for the shortcomings of traditional learning methods in the context of large data. Extreme learning machine is a key algorithm that has been found to achieve greater generalisation performance than other traditional learning algorithms at extremely fast learning speeds. Deep learning is a hot topic in machine learning right now, with implications for a wide range of data processing tasks. These two revolutionary machine learning techniques have found widespread use in large data processing challenges due to their distinct benefits for dealing with big data difficulties.

All of these processing methods must execute on distributed storage systems and parallel computing platforms or frameworks to increase the scalability of machine learning algorithms. Several attempts have been made to leverage vast distributed storage systems capable of handling large datasets, such as Google File System (GFS), BigTable, and Hadoop Distributed File System (HDFS) (HDFS). MapReduce, Twister, Dryad, Graphlab, Hadoop, and Haloop are some parallel processing platforms and tools. For managing big data activities, the framework model based on machine learning with distributed storage and parallel computing provides great performance.

Open Questions and Research Challenges

Machine learning is a powerful and vital tool for completing many activities and difficulties related to big data, however current research and advancements for big data processing still confront a number of significant research hurdles. We must address several important scientific difficulties and open questions in order to fully grasp the potential of big data, including (but not limited to):

- How to employ machine learning techniques to investigate and exploit the beneficial information hidden in big data should receive more attention, as enormous amounts of relevant data are being lost as fresh data is mostly untagged and unstructured.
- In most present machine learning applications, researchers use a single learning algorithm or strategy to solve practical problems, but it's crucial to remember that each approach has its own set of advantages and disadvantages. In light of the current big data landscape, the concept of hybrid learning should be explored further.
- Because of the features of huge data, data visualisation is a difficult process. Dimension reduction and other current visualisation approaches can only provide an abstract representation of the data.
  As a result, it's also important to look into how to apply machine learning techniques to create genuine geometric representations for huge data.

Conclusions

We began by providing an overview of big data and a summary of its types and properties in this document. We next examined the new aspects of machine learning with big data in order to emphasize the distinctions between machine learning approaches in the context of large data. We also developed reference architecture for processing big data that combines the potential of distributed storage and parallel computing with machine learning techniques. Finally, we discussed a number of research challenges and unresolved topics. We hope that this survey would pique people's interest in machine learning-based big data processing research and development.

References

[1]. T.M. Mitchell, Machine Learning, McGraw Hill, New York, 1997.
[2]. C. Rudin, and K.L. Wagstaff, Machine learning for science and society, Mach Learn. 95(2014)1-9.
[3]. X.W. Chen, and X Lin, Big data deep learning: challenges and perspectives, IEEE Access 2 (2014)514-525.
[4]. N Jones, Computer science: the learning machines, Nature 505 (2014) 146-148.
[5]. J. Dean, and S. Ghemawat, MapReduce: simplified data processing on large clusters, Commun. ACM 51 (2008) 107-113.
[6]. M. Isard, M. Budiu, Y. Yu, and A. Birrell, Dryad: distributed data-parallel programs from sequential building blocks, in Proc. of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems, Lisbon, 2007, pp. 59-72.
[7]. Y.C. Low, J. Gonzalez, A. Kyrola, D. Bickson,C. Guestrin, and J.M. Hellerstein, Graphlab: a new framework for parallel machine learning, in Proc. of the 26th Conference on Uncertainty in Artificial Intelligence (UAI), Catalina Island, 2010, pp. 340-349.
[8]. T. White, Hadoop: the Definitive Guide, O'Reilly Media Inc., California, 2009.
[9]. Y. Bu, B. Howe, M. Balazinska, and M.D. Ernst, HaLoop: efficient iterative data processing on large clusters, in Proc. of the 36th International Conference on Very Large Data Bases (VLDB), Singapore, 2010, pp. 285-296.
[10]. Q. He, T.F. Shang, F.Z. Zhuang, and Z.Z. Shi, Parallel extreme learning machine for regression based on MapReduce, Neurocomputing 102 (2013) 52-58.
[11]. M. Mardani, G. Mateos, and G.B. Giannakis, Subspace learning and imputation for streaming big data matrices and tensors, IEEE Trans. Signal Process. 63 (2015), 2663-2677.
[12]. L.Z. Wang, K. Lu, P. Liu, R. Ranjan, and L. Chen, IK-SVD: dictionary learning for spatial big data via incremental atom update, Comput. Sci. Eng. 16 (2014) 41-52.
[13]. J.L. Liang, M.H. Zhang, X.Y. Zeng, and G.Y. Yu, Distributed dictionary learning for sparse representation in sensor networks, IEEE Trans. Image Process. 23 (2014) 2528-2541.
[14]. V. Mayer-Schönberger, and K. Cukier, Big data: a revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt, New York, 2013.
[15]. M. Aharon, M. Elad, and A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Proces. 54 (2006) 4311-4322.
[16]. X. Cai, Sparse and large-scale learning models and algorithms for mining heterogeneous big data, Dissertation, University of Texas, 2013.
[17]. T.G. Dietterich, P. Domingos, L. Getoor, S. Muggleton, and P. Tadepalli, Structured Machine learning: the next ten years, Mach. Learn. 73 (2008) 3-23.
[18]. A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, Distributed structured prediction for Big data, in Proc. of the NIPS, Workshop on Big Learning, 2012.
[19]. J. Ekanayake, H. Li, B.J. Zhang, T. Gunarathne, S.H. Bae, J. Qiu, and G. Fox, Twister: a runtime for iterative MapReduce, in Proc. of the 19th ACM International Symposium on High Performance Distributed Computing (HDPC), Chicago, 2010, pp. 810-818.
[20]. Y.F. Zhang, Q.X. Gao, L.X. Gao, and C.R. Wang, iMapReduce: a distributed computing framework for iterative computation, J. Grid Comput. 10 (2012) 47-68.
[21]. Y.F. Zhang, and S.M. Chen, i2MapReduce: incremental iterative MapReduce, in Proc. of the 2nd International Workshop on Cloud Intelligence, Riva del Garda, 2013.
[22]. D. Che, M. Safran, and Z.Y. Peng, From big data to big data mining: challenges, issues, and opportunities, in Proc. of the 18th International Conference on Database Systems for Advanced Applications Lecture Notes in Computer Science (LNCS), Wuhan, 2013, pp. 1-15.

ISSN: 0374-8588
Volume 22 Issue 1, January 2020

[23]. S. F. Ding, X.Z. Xu, and R. Nie, Extreme learning machine and its applications, Neural Comput. App. 25 (2013) 549-556.
[24]. G. Hinton, S. Osindero, and Y.W. Teh, A fast learning algorithm for deep belief nets, Neural comput. 18 (2006) 1527-1554.
[25]. D. Yu, and L. Deng, Deep learning and its applications to signal and information processing, IEEE Signal Proc. Mag. 28 (2011) 145-154.
[26]. S. Ghemawat, H. Gobioff, and S.T. Leung, The Google file system, in Proc.of the 19th ACM Symposium on Operating Systems Principles, Lake George, 2003, pp. 29-43.
[27]. W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R.E. Gruber, Bigtable: a distributed storage system for structured data, ACM Trans. Comput. Syst. 26 (2008) 205-218.
[28]. M. Chen, S. Mao, and Y. Liu, Big data: a survey, Mobile Netw. Appl. 19 (2014) 171-209.