A Research on Machine Learning Methods for Big Data Processing

Junfei Qiu[1,a*], and Youming Sun[2,b]

[1]College of Communications Engineering, PLA University of Science and Technology, Nanjing, China, 210007

[2]National Digital Switching System Engineering and Technological Research Center, Zhengzhou, China, 450000

[a*]junfeiqiu@163.com, [b]sunyouming10@163.com

Keywords: Machine learning; Big data; Data mining; Cloud computing

Abstract. Machine learning has found widespread implementations and applications in many different domains in our life. However, as the big data era is coming, some traditional machine learning techniques cannot satisfy the requirements of real-time processing for large volumes of data. In response, machine learning needs to reinvent itself for big data. In this article, we provide a review of machine learning for big data processing in recent studies. Firstly, a discussion about big data is presented, followed by the analysis of the new characteristics of machine learning in the context of big data. Then, we propose a feasible reference framework for dealing with big data based on machine learning techniques. Finally, several research challenges and open issues are addressed.

Introduction

Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed, aiming to understand computational mechanisms by which experience can lead to improved performance [1]. It is a highly interdisciplinary field building upon ideas from many different kinds of domains. In the past decades, machine learning has covered almost every domain of our life which is so pervasive that you probably use it dozens of times a day without knowing it. It is primarily influencing the broader world through its implementation in a wide range of applications, which has brought great impact on the science and society [2]. A great number of machine learning algorithms have been proposed in the last decades, such as neural network, decision tree, support vector machine, k-nearest-neighbor, genetic algorithms, Q-learning, etc.. They have been used in diverse domains such as pattern recognition, robotics, natural language processing, and autonomous control systems [3, 4].

Machine learning is a rather efficient mathematics, based on statistical algorithms that can analyze large volume of diverse data sources. However, as the time for big data is coming, the collection of data sets is so large and complex that it is difficult to deal with using traditional data processing tools and models. As a result, some traditional machine learning techniques are unsuitable to this condition and cannot satisfy the requirements of real-time processing and storage for big data. Thus this needs us to explore some new methods with the power of distributed storage and parallel computing to analyze and deal with big data. In previous work, scholars mainly focused on two aspects of researches: i) one was to design a kind of distributed parallel computing framework or platform for fast dealing with big data, such as MapReduce [5], Dryad [6], Graphlab [7], Hadoop [8], Haloop [9], and Twister [19], etc.; ii) the other was to propose a sort of new algorithms to solve a class of determined big data problems. For example, He Q et al. applied parallel extreme learning machine for dealing with regression problems based on MapReduce [10]. In [11], the authors developed a low-complexity subspace learning to handle the incomplete streaming big data. Some researchers also applied the dictionary learning for sparse representation of big data [12, 13]. However, to date, there are relatively few discussions that systemically and deeply analyze the new characteristics of machine learning in the age of big data and provide the corresponding methods based on machine learning for dealing with big data. Therefore, in this paper, we mainly study the methods of handling big data based on machine

learning and design a reasonable framework model for big data processing. The main work of this article can be summarized as follows:

- We firstly give a brief review of big data and summarize five key words to characterize it, i.e., volume, variety, velocity, veracity and value.
- We then systemically and deeply analyze the new features of machine learning in the context of big data. Several possible solutions to tackling big data challenges are also discussed.
- We finally design a kind of reference framework, which is based on machine learning with the power of distributed storage and parallel computing, for fast processing big data.

An Overview of Big Data

We now live in an era of data deluge where large volumes of data are accumulating in all aspects of our lives. Data streams coming from diverse domains contribute to the emerging paradigm of big data. It may be a great opportunity for the big data scientist amongst the vast amount and array of data. By discovering associations, analyzing patterns and predicting trends within the data, big data has the potential to change our society and improve the quality of our life. Big data typically refers to the following three types based on data sources from physical, cyber, and social worlds:

- *Nature data:* we can imagine that data coming from the nature in our earth will be a great potential data source, such as satellite data from outer space.
- *Life data:* it is a big project on the study of biological body, especially the exploration on the human body still have a lot of challenges, such as biological data.
- *Sociality data:* with the fast development of digital mobile products and network, large volumes of sociality data are generating every day in our life, such as voice abd video data.

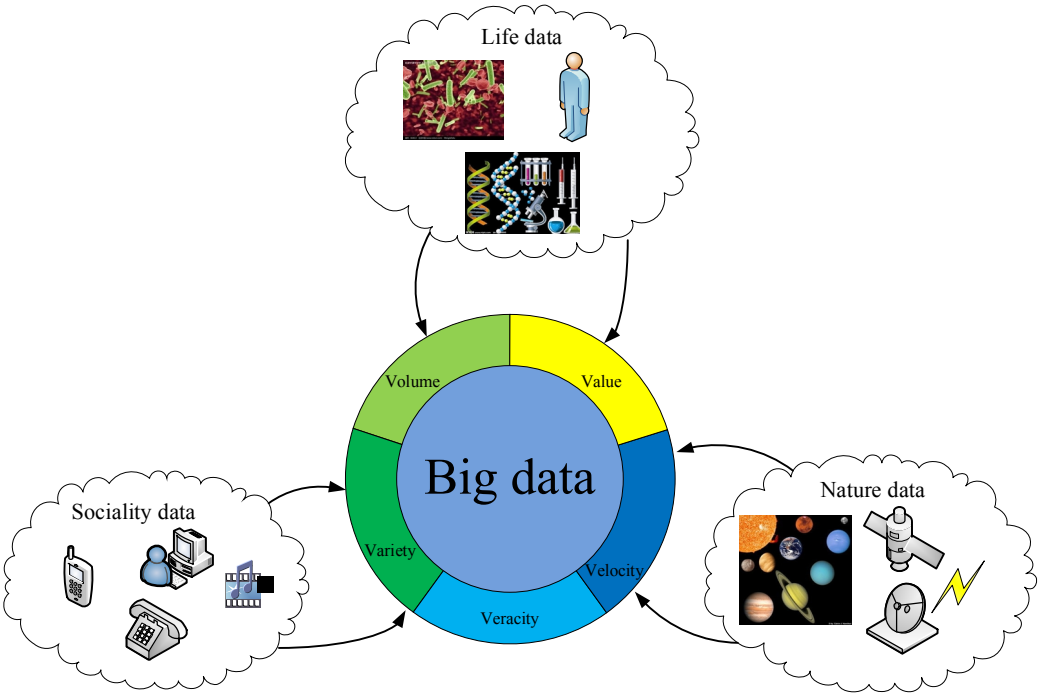

Fig. 1. Big data types and characteristics.

As shown in Fig. 1, big data can be characterized by five keywords: *volume, variety, velocity, veracity and value*. In the following, we will discuss each characteristic in detail.

- *Volume*. Volume relates to the size of data and is the primary attribute of big data [3]. It has been an indisputable fact that enormous amounts of data have been being continually generated at unprecedented scales from diverse domains in our life. The constant flow of new data accumulating at unprecedented rates brings great challenges to the traditional processing infrastructure in the side of effective capture, storage and manipulation of large volumes of data. It

requires high scalability of data management and mining tools.

- *Variety*. Variety means the different types of data [14]. Big data is generally from different sources which inherently possesses a lot of different formations including structured, unstructured and semi-structured representation forms. Mining such a heterogeneous dataset, the great challenge is perceivable, constructing a single model will not result in good-enough mining results. It is expected that specialized, more complex and multi-model systems to be constructed.
- *Velocity*. In general, the produced unprecedented data every day are often continuously generating in the form of streams that require being processed in real time or at a rapid pace [22]. In special time, we must finish some tasks within a certain period of time, otherwise, the processing results become less valuable or even worthless. To tackle this challenge, the key idea is to develop parallel processing techniques to handle data in parallelization.
- *Veracity*. It can be characterized as data accuracy [22]. In the era of big data, we may receive data from different fields with incomplete information in a great probability. These incomplete, uncertainty and dynamic data sources from many different origins greatly influence the quality of data. Therefore, the accuracy and trust of the source data quickly become a serious issue for concern. To solve this problem, data validation and provenance tracing become more and more important for data procesing systems.
- *Value*. The rise of big data is driven by the rapid development of artificial intelligence, machine learning and data mining technologies, presenting such a process: analyzing the data for information, extracting the information into knowledge and facilitating decision and action for acquiring desired values based on the knowledge. It is likely panning for gold in the sand to get valid values in terms of big data. Therefore, how to use the robust machine learning algorithms to achieve the value purification of data more quickly has become an urgent problem to be solved at present big data background [28].

While big data brings great opportunities, unpredictable challenges are on the way at the same time. It cannot be stored, analyzed and processed by traditional data management technologies and requires adaptation of some new workflows, platforms and architectures [14]. The field of machine learning which is useful to accomplish tasks of prediction, classification, and association about large amounts of data, is getting more and more attention from researchers in the current time. However, as the big data era is coming, some characteristics of big data will bring great challenges to the traditional machine learning methods. As a result, machine learning has to be provided with some new features to handle the problems that big data bringing. These new performances need to be systemically analyzed and deeply investigated.

New Features of Machine Learning with Big Data

In order to deal with the potential chanlleges posed by big data, machine learning has to possess some new properties compared with the traditional learning systems and techniques. In this section, we will highlight three aspects of abilities that are useful to deal with big data problems for machine learning techniques in detail, i.e., *sparse representation and feature selection, mining structured relations, high scalability and high speed*.

Feature Selection and Sparse Representation. Datasets with high-dimensional features have become increasingly common in big data scenarios. For the high-dimensional data, it is difficult to handle by using traditional data processing methods. Therefore, effective dimension reduction is increasingly viewed as a necessary step in dealing with these problems. In terms of high-dimensional big data, we highlight the feature selection and sparse representation methods for machine learning techniques, which are two commonly adopted approaches in dealing with high-dimensional data.

Feature selection is a key issue in building robust data processing models through the process of selecting a subset of meaningful features. Typically, many sparse based supervised binary feature selection methods can be written as the approximation of the following problem [16]:

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

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 $\{x_1, x_2, \mathbf{L}, x_n\} \in \mathbb{R}^{d\times 1}$ are training data and $\{y_1, y_2, \mathbf{L}, y_n\} \in \mathbb{R}^{m\times 1}$ 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 $x_i \in c$ denotes the input space object and $s_i \in S$ represents structured label space object. Further, $f : c \times S \to \mathbb{R}^F$ 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 $p_w(s \mid x) \propto \exp(w^T f(x, s)/e)$ with covariance e [18]. In order to find the model parameter w which best describs the possible labeling $s_i \in S$ of $x_i \in c$, we can construct a

task loss $\mathbf{l}_{(x,s)}(\hat{s})$ 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]:

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

where vector $v = \sum_{(x,s)\in D} f(x,s)$ 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 High Speed. The unprecedented volumes of big data require quite high scalability of their data mining and processing tools. In current researches, the techniques which are used to enhance the scalability issue of machine learning algorithms mainly focus on the following two aspects: i) the scalability of cloud computing makes it possible to analyze enormous datasets, which aggregates multiple workloads with varying performance goals into multi-tenanted computing clusters. Machine learning with cloud computing owns more efficient and higher performance for processing and analyzing big data; ii) distributed storage and parallel computing have helped to solve machine learning algorithms' scalability problems. Hadoop Distributed File System (HDFS) is a distributed storage system which is designed for storing very large data files, running on clusters on commodity hardware [8]. MapReduce is the programming paradigm with parallel data processing allowing massive scalability [5]. However, MapReduce suffers from an obvious weakness that it does not support iterations, leading to restricting the performances of machine learning algorithms [19]. Some extending studies for efficient iterative computations have been proposed, such as Graphlab [7], Twister [19], iMapReduce [20], and i^{2}MapReduce [21]. With the combination of cloud computing and distributed-parallel frameworks, the scalability of machine learning techniques can be greatly improved.

The capability of fast accessing and mining big data are also important abilities that the machine learning techniques have to possess. Speed is also relevant to scalability, solving anyone of them will help the other one [22]. The speed of data processing depends on two major factors: data access time and the efficiency of the learning algorithms themselves. In special time, we must finish some tasks within a certain period of time, otherwise, the processing results become less valuable or even worthless. For example, stock market prediction, earthquake prediction and agent-based autonomous exchange systems are the typical applications with real-time requests. Time-sensitive applications that require real-time response and processing need real-time computation and online implementations of machine learning algorithms. A useful approach to boost the speed of big data processing is through maximally identifying and exploiting the potential parallelism in the machine learning algorithms. High scalability and high speed can give machine learning high power to handle big data.

Discussions. In terms of these new features mentioned above, sparse representation and feature selection are mainly aimed at the characteristics of high dimension for big data by means of effective dimensionality reduction. Mining structured relations mainly focuses on the heterogeneous natures of big data with the method of inferring the structure behind the data to divid massive amounts of data into smaller chunks. And high scalability and high speed are aimed at large scale and real-time features of big data, respectively. In the age of big data, machine learning techineques have to possess these abilities to effectively and efficiently process big data problems.

A Reference Framework Based on Machine Learning for Big Data Processing

Since machine learning techniques can gradually achieve human-like learning, it is becoming tightly associated with big data. In this section, we propose a kind of reference framework model for big data processing based on machine learning with the power of distributed storage and parallel computing.

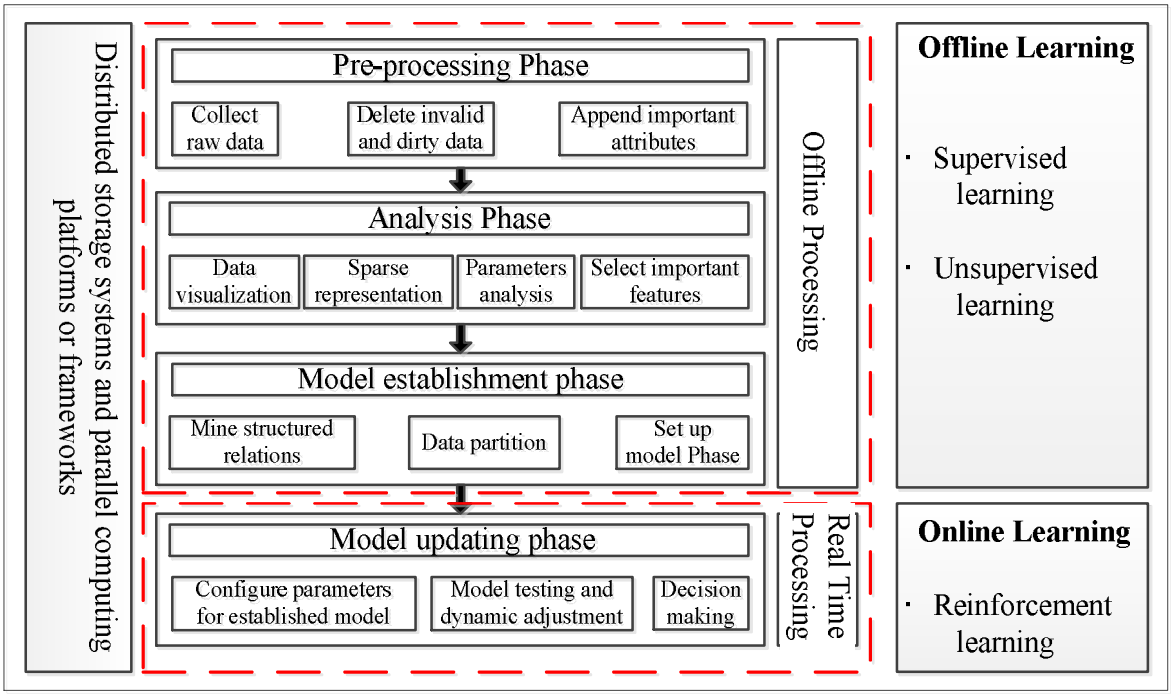

Fig. 2. The proposed reference framework based on machine learning for big data processing.

As described in Fig. 2, we suppose the big data processing procedure mainly consists of the following four phases: *pre-processing phase, analysis phase, model establishment phase and model updating phase*. Because data sources almost cover all different kinds of domains, raw big data collecting from the environment are greatly complex and has tremendous redundancies. Therefore, we need delete the invalid and dirty data at first in pre-processing phase. In addition, we frequently have to face massive uncertain and incomplete data in real life and we need append some important attributes to improve their processing practicability. After raw data pre-processing phase, we need analyze these valid and useful data to find out how to utilize the data through trial and error. Data visualization is a fundamental problem in the analysis of big data, and we can adopt sparse representation to achieve effective dimension reduction for the high-dimensional data. Through essential parameters analysis, we should be able to select some important features to establish the feasible model for dealing with real problems. In terms of model establishment phase, we try to mine the structured relations between data to obtain statistical information and trend at first, and then split data into training and testing sets. In the end, we can decide what kind of model should be generated for utilization and build up the corresponding model. While the model is established, we need configure parameters for the model and apply the generated model obtained from the model establishment phase into actual operations to test the performance of the big data processing model. In this phase, we emphasize the input data is real-time. We should make dynamic adjustments to update the model based on effects of model application.

In terms of the four phases in the procedure of big data processing, the anterior three phases are offline processing. In these phases, we are able to adopt offline learning methods which include two categories of supervised learning and unsupervised learning. In the model testing and updating phase, we mainly focus on the real-time characteristic of input data. To deal with the problem of real-time processing, online learning methods are necessary and the reinforcement learning is preferred. Although the study in the field of traditional machine learning techniques has reached a state of relative maturity, some advanced learning methods need to be developed for making up the deficiencies of

traditional learning methods in the context of big data. *Extreme learning machine* is an important algorithm which has been shown to achieve better generalization performance than other conventional learning algorithms at extremely high learning speed [10, 11, 23]. *Deep learning* is currently another extremely active research area in machine learning, impacting a wide range of data processing work [24, 25]. Due to their unique advantages for dealing with big data problems, these two novel machine learning techniques have found widespread applications in big data processing problems [10, 11, 25]. To promote the scalability of machine learning algorithms, all these processing procedures need to run on the distributed storage systems and parallel computing platforms or frameworks. Several attempts have been made on exploiting massive distributed storage systems which can manage large scale datasets, such as Google File System (GFS) [26], BigTable [27], and Hadoop Distributed File System (HDFS) [8]. Some parallel processing platforms and tools include MapReduce, Twister, Dryad, Graphlab, Hadoop, and Haloop [5-9, 19]. The framework model based on machine learning with the ability of distributed storage and parallel computing has strong performance for managing big data tasks.

Research Challenges and Open Issues

Machine learning is a powerful and essential tool for accomplishing many tasks and problems associated with big data, but current researches and developments are still faced with a lot of great research challenges for big data processing. In order to realize the full potential of big data, we need to address several major research challenges and open issues, including the following (but not limited to):

- How to explore and exploit the useful information hidden in big data by using machine learning techniques should draw further attention, as large quantities of useful data are getting lost since new data is largely untagged and unstructured data.
- In most existing machine learning applications, the researchers just apply single learning algorithm or technique to deal with practical problems, but it is important to realize that each approach has strengths and weaknesses. Thus the idea of hybrid learning should be further considered at present big data background.
- The characteristics of big data make the data visualization an enormously challenging task. The recent visualization techniques like dimension reduction can only give an abstract view of the data. Therefore, how to use machine learning techniques to give true geometric representations for big data also needs to be investigated.

Conclusions

In this paper, we firstly provided an overview about big data and summarized the types and characteristics of big data. In order to highlight the differences of machine learning techniques in the context of big data, we then analyzed the new features of machine learning with big data. We also proposed a reference framework for processing big data based on machine learning techniques with the power of distributed storage and parallel computing. Finally, we presented sevetal research challenges and open issues. We hope that this survey can stimulate more interests in research and development of techniques based on machine learning for big data processing.

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.

[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.