

REVIEW

Open Access



A survey of machine learning for big data processing

Junfei Qiu, Qihui Wu, Guoru Ding*, Yuhua Xu and Shuo Feng

Abstract

There is no doubt that big data are now rapidly expanding in all science and engineering domains. While the potential of these massive data is undoubtedly significant, fully making sense of them requires new ways of thinking and novel learning techniques to address the various challenges. In this paper, we present a literature survey of the latest advances in researches on machine learning for big data processing. First, we review the machine learning techniques and highlight some promising learning methods in recent studies, such as representation learning, deep learning, distributed and parallel learning, transfer learning, active learning, and kernel-based learning. Next, we focus on the analysis and discussions about the challenges and possible solutions of machine learning for big data. Following that, we investigate the close connections of machine learning with signal processing techniques for big data processing. Finally, we outline several open issues and research trends.

Keywords: Machine learning, Big data, Data mining, Signal processing techniques

1 Review

1.1 Introduction

It is obvious that we are living in a data deluge era, evidenced by the phenomenon that enormous amount of data have been being continually generated at unprecedented and ever increasing scales. Large-scale data sets are collected and studied in numerous domains, from engineering sciences to social networks, commerce, biomolecular research, and security [1]. Particularly, digital data, generated from a variety of digital devices, are growing at astonishing rates. According to [2], in 2011, digital information has grown nine times in volume in just 5 years and its amount in the world will reach 35 trillion gigabytes by 2020 [3]. Therefore, the term “Big Data” was coined to capture the profound meaning of this data explosion trend.

To clarify what the big data refers to, several good surveys have been presented recently and each of them views the big data from different perspectives, including challenges and opportunities [4], background and research status [5], and analytics platforms [6]. Among these surveys, a comprehensive overview of the big data from three different angles, i.e., innovation, competition, and productivity, was presented by the McKinsey Global

Institute (MGI) [7]. Besides describing the fundamental techniques and technologies of big data, a number of more recent studies have investigated big data under particular context. For example, [8, 9] gave a brief review of the features of big data from Internet of Things (IoT). Some authors also analyzed the new characteristics of big data in wireless networks, e.g., in terms of 5G [10]. In [11, 12], the authors proposed various big data processing models and algorithms from the data mining perspective.

Over the past decade, machine learning techniques have been widely adopted in a number of massive and complex data-intensive fields such as medicine, astronomy, biology, and so on, for these techniques provide possible solutions to mine the information hidden in the data. Nevertheless, 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 learning methods since the established process of learning from conventional datasets was not designed to and will not work well with high volumes of data. For instance, most traditional machine learning algorithms are designed for data that would be completely loaded into memory [13], which does not hold any more in the context of big data. Therefore, although learning from these numerous data is expected to bring significant science and engineering

* Correspondence: dingguoru@gmail.com
College of Communications Engineering, PLA University of Science and Technology, Nanjing 210007, China

advances along with improvements in quality of our life [14], it brings tremendous challenges at the same time.

The goal of this paper is twofold. One is mainly to discuss several important issues related to learning from massive amounts of data and highlight current research efforts and the challenges to big data, as well as the future trends. The other is to analyze the connections of machine learning with modern signal processing (SP) techniques for big data processing from different perspectives. The main contributions of this paper are summarized as follows:

- We first give a brief review of the traditional machine learning techniques, followed by several advanced learning methods in recent researches that are either promising or much needed for solving the big data problems.
- We then present a systematic analysis of the challenges and possible solutions for learning with big data, which are in terms of the five big data characteristics such as volume, variety, velocity, veracity, and value.
- We next discuss the great ties of machine learning with SP techniques for the big data processing.
- We finally provide several open issues and research trends.

The remainder of the paper, as the roadmap given in Fig. 1 shows, is organized as follows. In Section 1.2, we start with a review of some essential and relevant concepts about machine learning, followed by some current advanced learning techniques. Section 1.3 provides a comprehensive survey of challenges bringing by big data for machine learning, mainly from five aspects. The relationships between machine learning and signal processing techniques for big data processing are presented in Section 1.4. Section 1.5 gives some open issues and research trends. Conclusions are drawn in Section 2.

1.2 Brief review of machine learning techniques

In this section, we first present some essential concepts and classification of machine learning and then highlight a list of advanced learning techniques.

1.2.1 Definition and classification of machine learning

Machine learning is a field of research that formally focuses on the theory, performance, and properties of learning systems and algorithms. It is a highly interdisciplinary field building upon ideas from many different kinds of fields such as artificial intelligence, optimization theory, information theory, statistics, cognitive science, optimal control, and many other disciplines of science, engineering, and mathematics [15–18]. Because of its implementation in a wide range of applications, machine learning has covered almost every scientific domain, which has brought great impact on the science and society [19]. It has been used on a variety of problems, including recommendation engines, recognition systems, informatics and data mining, and autonomous control systems [20].

Generally, the field of machine learning is divided into three subdomains: supervised learning, unsupervised learning, and reinforcement learning [21]. Briefly, supervised learning requires training with labeled data which has inputs and desired outputs. In contrast with the supervised learning, unsupervised learning does not require labeled training data and the environment only provides inputs without desired targets. Reinforcement learning enables learning from feedback received through interactions with an external environment. Based on these three essential learning paradigms, a lot of theory mechanisms and application services have been proposed for dealing with data tasks [22–24]. For example, in [22], Google applies machine learning algorithms to massive chunks of messy data obtained from the Internet for Google's translator, Google's street view, Android's voice recognition, and image search engine. A simple comparison of these three machine learning technologies from different perspectives is given in Table 1

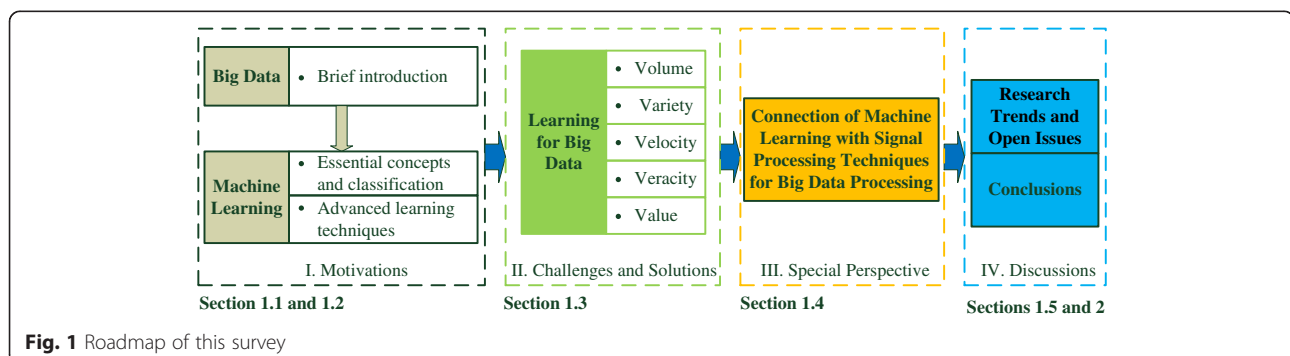


Fig. 1 Roadmap of this survey

Table 1 Comparison of machine learning technologies

Learning types	Data processing tasks	Distinction norm	Learning algorithms	Representative references
Supervised learning	Classification/Regression/Estimation	Computational classifiers	Support vector machine	[120]
			Statistical classifiers	[15]
		Connectionist classifiers	Hidden Markov model	[121]
			Bayesian networks	[122]
Unsupervised learning	Clustering/Prediction	Parametric	Neural networks	[123]
			K-means	[124]
		Nonparametric	Gaussian mixture model	[125]
			Dirichlet process mixture model	[125]
Reinforcement learning	Decision-making	Model-free	X-means	[124]
			Q-learning	[126]
		Model-based	R-learning	[127]
			TD learning	[128]
			Sarsa learning	[129]

to outline the machine learning technologies for data processing. The “Data Processing Tasks” column of the table gives the problems that need to be solved and the “Learning Algorithms” column describes the methods that may be used. In summary, from data processing perspective, supervised learning and unsupervised learning mainly focus on data analysis while reinforcement learning is preferred for decision-making problems. Another point is that most traditional machine-learning-based systems are designed with the assumption that all the collected data would be completely loaded into memory for centralized processing. However, as the data keeps getting bigger and bigger, the existing machine learning techniques encounter great difficulties when they are required to handle the unprecedented volume of data. Nowadays, there is a great need to develop efficient and intelligent learning methods to cope with future data processing demands.

1.2.2 Advanced learning methods

In this subsection, we introduce a few recent learning methods that may be either promising or much needed for solving the big data problems. The outstanding characteristic of these methods is to focus on the idea of learning, rather than just a single algorithm.

1. *Representation Learning*: Datasets with high-dimensional features have become increasingly common nowadays, which challenge the current learning algorithms to extract and organize the discriminative information from the data. Fortunately, representation learning [25, 26], a promising solution to learn the meaningful and useful representations of the data that make it easier to extract useful information when building classifiers or other

predictors, has been presented and achieved impressive performance on many dimensionality reduction tasks [27]. Representation learning aims to achieve that a reasonably sized learned representation can capture a huge number of possible input configurations, which can greatly facilitate improvements in both computational efficiency and statistical efficiency [25]. There are mainly three subtopics on representation learning: feature selection, feature extraction, and distance metric learning [27]. In order to give impetus to the multidomain learning ability of representation learning, automatic representation learning [28], biased representation learning [26], cross-domain representation learning [27], and some other related techniques [29] have been proposed in recent years. The rapid increase in the scientific activity on representation learning has been accompanied and nourished by a remarkable string of empirical successes in real-world applications, such as speech recognition, natural language processing, and intelligent vehicle systems [30–32].

2. *Deep learning*: Nowadays, there is no doubt that deep learning is one of the hottest research trends in machine learning field. In contrast to most traditional learning techniques, which are considered using shallow-structured learning architectures, deep learning mainly uses supervised and/or unsupervised strategies in deep architectures to automatically learn hierarchical representations [33]. Deep architectures can often capture more complicated, hierarchically launched statistical patterns of inputs for achieving to be adaptive to new areas than traditional learning methods and often outperform state of the art achieved by hand-made features [34]. Deep belief networks (DBNs) [33, 35] and convolutional neural

networks (CNNs) [36, 37] are two mainstream deep learning approaches and research directions proposed over the past decade, which have been well established in the deep learning field and shown great promise for future work [13].

Due to the state-of-the-art performance of deep learning, it has attracted much attention from the academic community in recent years such as speech recognition, computer vision, language processing, and information retrieval [33, 38–40]. As the data keeps getting bigger, deep learning is coming to play a pivotal role in providing predictive analytics solutions for large-scale data sets, particularly with the increased processing power and the advances in graphics processors [13]. For example, IBM's brain-like computer [22] and Microsoft's real-time language translation in Bing voice search [41] have used techniques like deep learning to leverage big data for competitive advantage.

3. *Distributed and parallel learning*: There is often exciting information hidden in the unprecedented volumes of data. Learning from these massive data is expected to bring significant science and engineering advances which can facilitate the development of more intelligent systems. However, a bottleneck preventing such a big blessing is the inability of learning algorithms to use all the data to learn within a reasonable time. In this context, distributed learning seems to be a promising research since allocating the learning process among several workstations is a natural way of scaling up learning algorithms [42]. Different from the classical learning framework, in which one requires the collection of that data in a database for central processing, in the framework of distributed learning, the learning is carried out in a distributed manner [43]. In the past years, several popular distributed machine learning algorithms have been proposed, including decision rules [44], stacked generalization [45], meta-learning [46], and distributed boosting [47]. With the advantage of distributed computing for managing big volumes of data, distributed learning avoids the necessity of gathering data into a single workstation for central processing, saving time and energy. It is expected that more widespread applications of the distributed learning are on the way [42]. Similar to distributed learning, another popular learning technique for scaling up traditional learning algorithms is parallel machine learning [48]. With the power of multicore processors and cloud computing platforms, parallel and distributed computing systems have recently become widely accessible [42]. A more detailed description about distributed and parallel learning can be found in [49].

4. *Transfer learning*: A major assumption in many traditional machine learning algorithms is that the training and test data are drawn from the same feature space and have the same distribution. However, with the data explosion from variety of sources, great heterogeneity of the collected data destroys the hypothesis. To tackle this issue, transfer learning has been proposed to allow the domains, tasks, and distributions to be different, which can extract knowledge from one or more source tasks and apply the knowledge to a target task [50, 51]. The advantage of transfer learning is that it can intelligently apply knowledge learned previously to solve new problems faster. Based on different situations between the source and target domains and tasks, transfer learning is categorized into three subsettings: inductive transfer learning, transductive transfer learning, and unsupervised transfer learning [51]. In terms of inductive transfer learning, the source and target tasks are different, no matter when the source and target domains are the same or not. Transductive transfer learning, in contrast, the target domain is different from the source domain, while the source and target tasks are the same. Finally, in the unsupervised transfer learning setting, the target task is different from but related to the source task. Furthermore, approaches to transfer learning in the above three different settings can be classified into four contexts based on "What to transfer," such as the instance transfer approach, the feature representation transfer approach, the parameter transfer approach, and the relational knowledge transfer approach [51–54]. Recently, transfer learning techniques have been applied successfully in many real-world data processing applications, such as cross-domain text classification, constructing informative priors, and large-scale document classification [55–57].
5. *Active learning*: In many real-world applications, we have to face such a situation: data may be abundant but labels are scarce or expensive to obtain. Frequently, learning from massive amounts of unlabeled data is difficult and time-consuming. Active learning attempts to address this issue by selecting a subset of most critical instances for labeling [58]. In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data [59]. It can obtain satisfactory classification performance with fewer labeled samples via query strategies than those of conventional passive learning [60]. There are three main active learning scenarios, comprising membership query synthesis, stream-based

selective sampling and pool-based sampling [59]. Popular active learning approaches can be found in [61]. They have been studied extensively in the field of machine learning and applied to many data processing problems such as image classification and biological DNA identification [61, 62].

6. *Kernel-based learning*: Over the last decade, kernel-based learning has established itself as a very powerful technique to increase the computational capability based on a breakthrough in the design of efficient nonlinear learning algorithms [63]. The outstanding advantage of kernel methods is their elegant property of implicitly mapping samples from the original space into a potentially infinite-dimensional feature space, in which inner products can be calculated directly via a kernel function [64]. For example, in kernel-based learning theory, data x in the input space \mathcal{X} is projected onto a potentially much higher dimensional feature space \mathcal{F} via a nonlinear mapping Φ as follows:

$$\Phi: \mathcal{X} \rightarrow \mathcal{F}, x \mapsto \Phi(x) \quad (1)$$

In this context, for a given learning problem, one now works with the mapped data $\Phi(x) \in \mathcal{F}$ instead of $x \in \mathcal{X}$ [63]. The data in the input space can be projected onto different feature spaces with different mappings. The diversity of feature spaces gives us more choices to gain better performance, while in practice, the choice itself of a proper mapping for any given real-world problem may generally be nontrivial. Fortunately, the kernel trick provides an elegant mathematical means to construct powerful nonlinear variants of most well-known statistical linear techniques, without knowing the mapping explicitly. Indeed, one only needs to replace the inner product operator of a linear technique with an appropriate kernel function k (i.e., a positive semi-definite symmetric function), which arises as a similarity measure that can be thought as an inner product between pairs of data in the feature space. Here, the original nonlinear problem can be transformed into a linear formulation in a higher dimensional space \mathcal{F} with an appropriate kernel k [65]:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{F}}, \forall x, x' \in \mathcal{X} \quad (2)$$

The most widely used kernel functions include Gaussian kernels and Polynomial kernels. These kernels implicitly map the data onto high-dimensional spaces, even infinite-dimensional spaces [63]. Kernel functions provide the nonlinear means to infuse correlation or side information in big data, which can obtain significant performance improvement over their linear counterparts at the price of generally higher computational complexity. Moreover, for a specific

problem, the selection of the best kernel function is still an open issue, although ample experimental evidence in the literature supports that the popular kernel functions such as Gaussian kernels and polynomial kernels perform well in most cases. At the root of the success of kernel-based learning, the combination of high expressive power with the possibility to perform the numerous analyses has been developed in many challenging applications [65], e.g., online classification [66], convexly constrained parameter/function estimation [67], beamforming problems [68], and adaptive multiregression [69]. One of the most popular surveys about introducing kernel-based learning algorithms is [70], in which an introduction of the exciting field of kernel-based learning methods and applications was given.

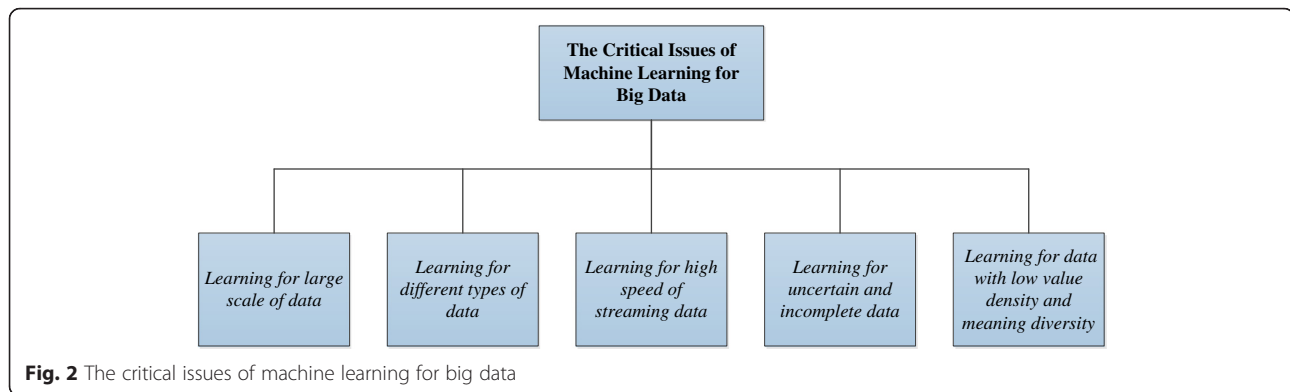
1.3 The critical issues of machine learning for big data

In spite of the recent achievement in machine learning is great as mentioned in Section 1.2, with the emergence of big data, much more needs to be done to address many significant challenges posted by big data. In this section, we present a discussion about the critical issues of machine learning techniques for big data from five different perspectives, as described in Fig. 2, including learning for large scale of data, learning for different types of data, learning for high speed of streaming data, learning for uncertain and incomplete data, and learning for extracting valuable information from massive amounts of data. Also, corresponding possible remedies to surmount the obstacles in recent researches are introduced in the discussion.

1.3.1 Critical issue one: learning for large scale of data

1.3.1.1 Critical issue It is obvious that data *volume* is the primary attribute of big data, which presents a great challenge for machine learning. Taking only the digital data as an instance, every day, Google alone needs to process about 24 petabytes (petabyte = $2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$ bytes) of data [71]. Moreover, if we further take into consideration other data sources, the data scale will become much bigger. Under current development trends, data stored and analyzed by big organizations will undoubtedly reach the petabyte to exabyte (exabyte = 2^{10} petabytes) magnitude soon [6].

1.3.1.2 Possible remedies There is no doubt that we are now swimming in an expanding sea of data that is too voluminous to train a machine learning algorithm with a central processor and storage. Instead, *distributed frameworks with parallel computing* are preferred. Alternating direction method of multipliers (ADMM) [72, 73] serving as a promising computing framework to develop distributed, scalable, online convex optimization algorithms is well suited to



accomplish parallel and distributed large-scale data processing. The key merits of ADMM is its ability to split or decouple multiple variables in optimization problems, which enables one to find a solution to a large-scale global optimization problem by coordinating solutions to smaller sub-problems. Generally, ADMM is convergent for convex optimization, but it is lack of convergence and theoretical performance guarantees for nonconvex optimization. However, vast experimental evidence in the literature supports empirical convergence and good performance of ADMM [74]. A wide variety of applications of ADMM to machine learning problems for large-scale datasets have been discussed in [74].

In addition to distributed theoretical framework for machine learning to mitigate the challenges related to high volumes, some practicable *parallel programming methods* are also proposed and applied to learning algorithms to deal with large-scale data sets. *MapReduce* [75, 76], a powerful programming framework, enables the automatic paralleling and distribution of computation applications on large clusters of commodity machines. What is more, MapReduce can also provide great fault tolerance ability, which is important for tackling the large data sets. The core idea of MapReduce is to divide massive data into small chunks firstly, then, deal with these chunks in parallel and in a distributed manner to generate intermediate results. By aggregating all the intermediate results, the final result is derived. A general means of programming machine learning algorithms on multicore with the advantage of MapReduce has been investigated in [77]. Cloud-computing-assisted learning method is another impressive progress which has been made for data systems to deal with the volume challenge of big data. *Cloud computing* [78, 79] has already demonstrated admirable elasticity that bears the hope of realizing the needed scalability for machine learning algorithms. It can enhance computing and storage capacity through cloud infrastructure. In this context, *distributed GraphLab*, a framework for machine learning in the cloud, has been proposed in [80].

1.3.2 Critical issue two: learning for different types of data

1.3.2.1 Critical issue The enormous *variety* of data is the second dimension that makes big data both interesting and challenging. This is resulted from the phenomenon that data generally come from various sources and are of different types. Structured, semi-structured, and even entirely unstructured data sources stimulate the generation of *heterogeneous, high-dimensional, and nonlinear* data with different representation forms. Learning with such a dataset, the great challenge is perceivable and the degree of complexity is not even imaginable before we deeply get there.

1.3.2.2 Possible remedies In terms of heterogeneous data, *data integration* [81, 82], which aims to combine data residing at different sources and provide the user with a unified view of these data, is a key method. An effect solution to address the data integration problem is to learn good data representations from each individual data source and then to integrate the learned features at different levels [13]. Thus, representation learning is preferred in this issue. In [83], the authors proposed a data fusion theory based on statistical learning for the two-dimensional spectrum heterogeneous data. In addition, deep learning methods have also been shown to be very effective in integrating data from different sources. For example, Srivastava and Salakhutdinov [84] developed a novel application of deep learning algorithms to learn a unified representation by integrating real-valued dense image data and text data.

Another challenge associated with high variety is that the data are often high dimensional and nonlinear, such as global climate patterns, stellar spectra, and human gene distributions. Clearly, to deal with high-dimensional data, *dimensionality reduction* is an effective solution through finding meaningful low-dimensional structures hidden in their high-dimensional observations. Common approaches are to employ feature selection or extraction to reduce the data dimensions. For example, Sun et al. [85] proposed a local-learning-based feature selection algorithm for high-dimensional data analysis. The existing typical machine learning algorithms for data dimensionality reduction

include principal component analysis (PCA), linear discriminant analysis (LDA), locally linear embedding (LLE), and laplacian Eigenmaps [86]. Most recently, *low-rank matrix* plays a more and more central role in large-scale data analysis and dimensionality reduction [8, 87]. The problem of recovering a low-rank matrix is a fundamental problem with applications in machine learning [88]. Here, we provide a simple example of using low-rank matrix recovery algorithms for high-dimensional data processing. Let us assume that we are given a large data matrix \mathbf{N} and know that it may be decomposed as $\mathbf{N} = \mathbf{M} + \mathbf{\Lambda}$, where \mathbf{M} has low rank and $\mathbf{\Lambda}$ is a noise matrix. Due to the low-dimensional column or row space of \mathbf{M} , not even their dimensions are not known, it is necessary to recover the matrix \mathbf{M} from the data matrix \mathbf{N} and the problem can be formulated as classical PCA [8, 89]:

$$\begin{aligned} \min_{\{\mathbf{M}\}} \|\mathbf{M}\|_* \\ \text{s.t. } \|\mathbf{N} - \mathbf{M}\|_F \leq \varepsilon \end{aligned} \quad (3)$$

where ε is a noise related parameter, $\|\cdot\|_*$ and $\|\cdot\|_F$ is defined by the nuclear norm and the Frobenius norm of a matrix, respectively. The problem formulated in (3) shows the fundamental task of the research on matrix recovery for high-dimensional data processing, which can be efficiently solved by some existing algorithms including augmented Lagrange multipliers (ALM) algorithm and accelerated proximal gradient (APG) algorithm [90]. As for nonlinear properties of data related to high variety, kernel-based learning methods can provide commendable solutions which have been discussed in Section 1.2.2; thus, the repetitious details will not be given here. Of course, in terms of challenges brought by different types, transfer learning is also a very good choice owing to its powerful knowledge transfer ability which enables multidomain learning to be possible.

1.3.3 Critical issue three: learning for high speed of streaming data

1.3.3.1 Critical issue For big data, speed or *velocity* really matters, which is another emerging challenge for learning. In many real-world applications, we have to finish a task within a certain period of time; otherwise, the processing results become less valuable or even worthless, such as earthquake prediction, stock market prediction and agent-based autonomous exchange (buying/selling) systems, and so on. In these time-sensitive cases, the potential value of data depends on data freshness that needs to be processed in a real-time manner.

1.3.3.2 Possible remedies One promising solution for learning from such high speed of data is online learning approaches. Online learning [91–94] is a well-established learning paradigm whose strategy is learning one instance

at a time, instead of in an offline or batch learning fashion, which needs to collect the full information of training data. This sequential learning mechanism works well for big data as current machines cannot hold the entire dataset in memory. To speed up learning, recently, a novel learning algorithm for single hidden-layer feed forward neural networks (SLFNs) named extreme learning machine (ELM) [95] was proposed. Compared with some other traditional learning algorithms, ELM provides extremely faster learning speed, better generalization performance, and with least human intervention [96]. Thus, ELM has strong advantages in dealing with high velocity of data.

Another challenging issue associated with the high velocity is that data are often nonstationary [13], i.e., data distribution is changing over time, which needs the learning algorithms to learn the data as a stream. To tackle this problem, the potential superiority of streaming processing theory and technology [97] have been found out compared with batch-processing paradigm, as they aim to analyze data as soon as possible to derive its results. Representative streaming processing systems include Borealis [98], S4 [99], Kafka [100], and many other recent architectures proposed to provide real-time analytics over big data [101, 102]. A scalable machine learning online service with the power of streaming processing for big data real-time analysis is introduced in [103]. In addition, the professor G. B. Giannakis have paid more attention to the real-time processing of streaming data by using machine learning techniques in recent studies; more details can be referred to in [87, 104].

1.3.4 Critical issue four: learning for uncertain and incomplete data

1.3.4.1 Critical issue In the past, machine learning algorithms were typically fed with relatively accurate data from well-known and quite limited sources, so the learning results tend to be unerring, too; thus, veracity has never been a serious issue for concern. However, with the sheer size of data available today, the precision and trust of the source data quickly become an issue, due to the data sources are often of many different origins and data quality is not all verifiable. Therefore, we include *veracity* as the fourth critical issue for learning with big data to emphasize the importance of addressing and managing the uncertainty and incompleteness on data quality.

1.3.4.2 Possible remedies Uncertain data are a special type of data reality where data readings and collections are no longer deterministic but are subject to some random or probability distributions. In many applications, data uncertainty is common. For example, in wireless networks, some spectrum data are inherently uncertain resulted from ubiquitous noise, fading, and shadowing and the technology barrier of the GPS sensor equipment also limits the accuracy of the data to certain levels. For

uncertain data, the major challenge is that the data feature or attribute is captured not by a single point value but represented as sample distributions [11]. A simple way to handle data uncertainty is to apply summary statistics such as means and variances to abstract sample distributions. Another approach is to utilize the complete information carried by the probability distributions to construct a decision tree, which is called distribution-based approach in [105]. In [105], the authors firstly discussed the sources of data uncertainty and gave some examples and then devised an algorithm for building decision trees from uncertain data using the distribution-based approach. At last, a theoretical foundation was established on which pruning techniques were derived which can significantly improve the computational efficiency of the distribution-based algorithms for uncertain data.

The incomplete data problem, in which certain data field values or features are missing, exists in a wide range of domains with the emerging big data, which may be caused by different realities, such as data device malfunction. Learning from these imperfect data is a challenging task, due to most existing machine learning algorithms that cannot be directly applied. Taking classifier learning as an example, dealing with incomplete data is an important issue, since data incompleteness not only impacts interpretations of the data or the models created from the data but may also affect the prediction accuracy of learned classifiers. To tackle the challenges associated with data incompleteness, Chen and Lin [13] investigated to apply the advanced deep learning methods to handle noisy data and tolerate some messiness. Furthermore, integrating the matrix completion technologies into machine learning to solve the problem of incomplete data is also a very promising direction [106]. In the following, we provide a case of using matrix completion for incomplete data processing. In this case, it is assumed that a noise matrix $\tilde{\mathbf{Y}}$ is defined by

$$\mathcal{P}_{\Omega}(\tilde{\mathbf{Y}}) = \mathcal{P}_{\Omega}(\mathbf{A} + \mathbf{Z}) \quad (4)$$

where \mathbf{A} is a sampled set of entries we would like to know as precisely as possible, \mathbf{Z} is a noise term which may be stochastic or deterministic, Ω is the set of indices of the acquired entries, and \mathcal{P}_{Ω} is the orthogonal projection onto the linear subspace of matrices supported on Ω [8]. To recover the unknown matrix, the problem can be formulated as [8]:

$$\begin{aligned} \min_{\{\mathbf{M}\}} \|\mathbf{A}\|_* \\ \text{s.t. } \|\mathcal{P}_{\Omega}(\mathbf{A} - \mathbf{Y})\|_F \leq \varepsilon \end{aligned} \quad (5)$$

To efficiently solve the problem (5), existing algorithms have been explained in [90] in detail. Furthermore, in terms of the abnormal data, the authors in [107] also investigated to use the statistical learning

theory of sparse matrix with data cleansing for the robust spectrum sensing.

1.3.5 Critical issue five: learning for data with low value density and meaning diversity

1.3.5.1 Critical issue In fact, by exploiting a variety of learning methods to analyze big datasets, the final purpose is to extract valuable information from massive amounts of data in the form of deep insight or commercial benefits. Therefore, *value* is also characterized as a salient feature of big data [2, 6]. However, to derive significant value from high volumes of data with a *low value density* is not straightforward. For example, the police often need to look through some surveillance videos to handle criminal cases. Unfortunately, a few valuable data frames are frequently hidden in a large amount of video sources.

1.3.5.2 Possible remedies To handle this challenge, knowledge discovery in databases (KDD) and data mining technologies [9, 11, 108] come into play, for these technologies provide possible solutions to find out the required information hidden in the massive data. In [9], the authors reviewed studies on applying data mining and KDD technologies to the IoT. Particularly, utilizing clustering, classification, and frequent patterns technologies to mine value from massive data in IoT, from the perspective of infrastructures and from the perspective of services were discussed in detail. In [11], Wu et al. characterized the features of the big data revolution and proposed big data processing methods with machine learning and data mining algorithms.

Another challenging problem associated with the value of big data is the *diversity of data meaning*, i.e., the economic value of different data varies significantly, even the same data have different value if considering from different perspectives or contexts. Therefore, some new cognition-assisted learning technologies should be developed to make current learning systems more flexible and intelligent. The most dramatic example of such devices is IBM's "Watson" [109], constructed with several subsystems that use different machine learning strategies with the great power of cognitive technologies to analyze the questions and arrive at the most likely answer. With the scientists' ingenuity, it is possible for this system to excel at a game which requires both encyclopedic knowledge and lightning-quick recall. Some humanlike characteristics—learning, adapting, interacting, and understanding enable Watson to be smarter and gain more computing power to deal with complexity and big data. It is expected that the era of cognitive computing will come [109].

1.3.6 Discussions

In summary, the five aspects mentioned above reflect the primary characteristics of big data, which refers to *volume*, *variety*, *velocity*, *veracity*, and *value* [2, 4–6, 13]. The five salient features bring different challenges for machine learning techniques, respectively. To surmount these obstacles, machine learning in the context of big data is significantly different from the traditional learning methods, as discussed above, some scalable, multidomain, parallel, flexible, and intelligent learning methods are preferred. What is more, several enabling technologies are needed to be integrated into the learning progress to improve the effectiveness of learning. A hierarchical framework is described in Fig. 3 to summarize the efficient machine learning for big data processing.

In fact, for big data processing, most machine learning techniques are not universal, that is to say, we often need to use specific learning methods according to different data. For example, in terms of high-dimensional datasets, representation learning seems to be a promising solution, which can learn the meaningful representations of the

data that make it easier to extract useful information for achieving impressive performance on many dimensionality reduction tasks. While for large volumes of data, distributed and parallel learning methods have stronger advantages. If the data needed to be processed are drawn from different feature spaces and have different distributions, transfer learning will be a good choice which can intelligently apply knowledge learned previously to solve new problems faster. Frequently, in the context of big data, we have to face such a situation: data may be abundant but labels are scarce or expensive to obtain. To tackle this issue, active learning can achieve high accuracy using as few labeled instances as possible. In addition, nonlinear data processing is also another thorny problem, at this moment, kernel-based learning will be here with its powerful computational capability. Of course, if we want to deal with some data in a timely or (nearly) real-time manner, online learning and extreme learning machine can give us more help.

Therefore, such a context is needed to be clear, in other words, what are the data tasks, data analysis or

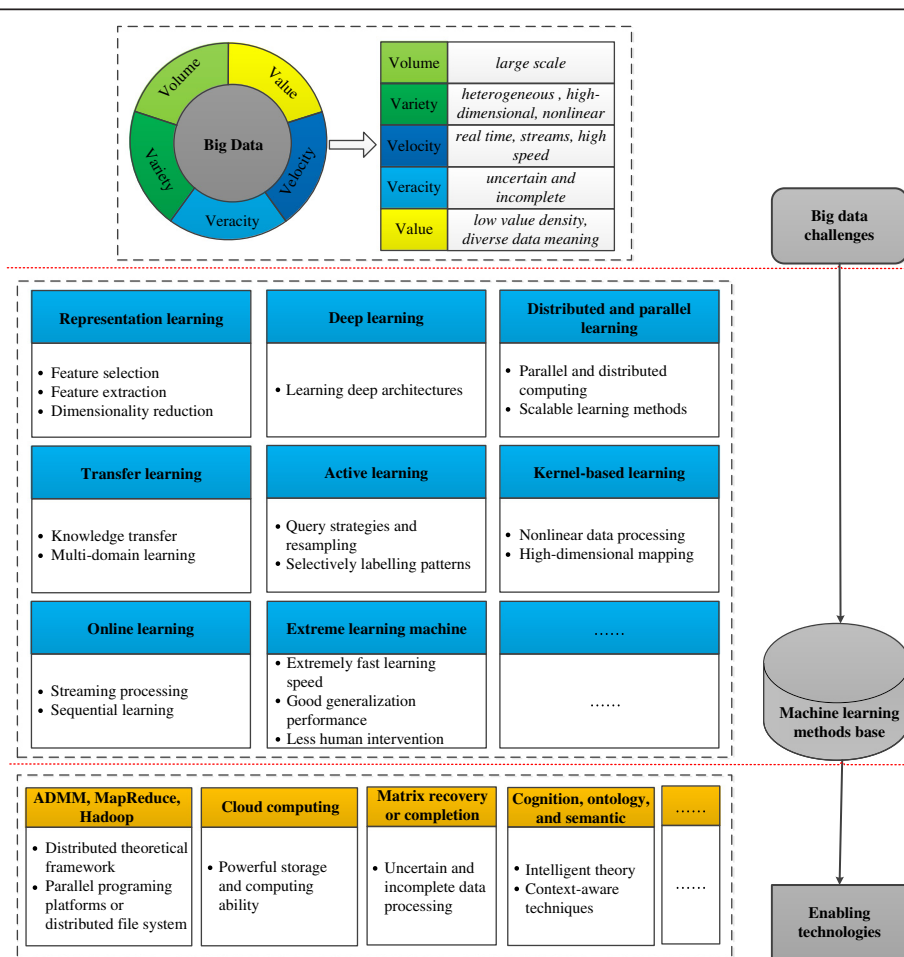


Fig. 3 Hierarchical framework of efficient machine learning for big data processing

decision making?; what are the data types, video data or text data?; what are the data characteristics, high volume or high velocity?; and so on. In terms of different data tasks, types, and characteristics, the required learning techniques are different, even a machine learning methods base is needed for big data processing. The learning systems can fast refer to the algorithm base to handle data. What is more, in order to improve the effectiveness of data processing, the combination of machine learning with some other techniques have been proposed in recent years. For example, in [80], the authors presented a cloud-assisted learning framework to enhance store and computing abilities. A general means of programming machine learning algorithms on multi-core with the advantage of MapReduce were investigated to enable the parallel and distributed processing to be possible [77]. IBM's brain-like computer, Watson, applied cognition techniques to machine learning field to make learning systems more intelligent [109]. Such enabling technologies have brought great benefits for machine learning, especially for large data processing, which are more worthy of study.

1.4 Connection of machine learning with SP techniques for big data

There is no doubt that SP is of uttermost relevance to timely big data applications such as real-time medical imaging, sentiment analysis from online social media, smart cities, and so on [110]. The interest in big-data-related research from the SP community is evident from the increasing number of papers submitted on this topic to SP-oriented journals, workshops, and conferences. In this section, we mainly discuss the close connections of machine learning with SP techniques for big data processing. Specifically, in Section 1.4.1, we analyze the existing studies on SP for big data from four different perspectives. Several representative literatures are presented. In Section 1.4.2, we provide a review of the latest research progress which is based on these typical works.

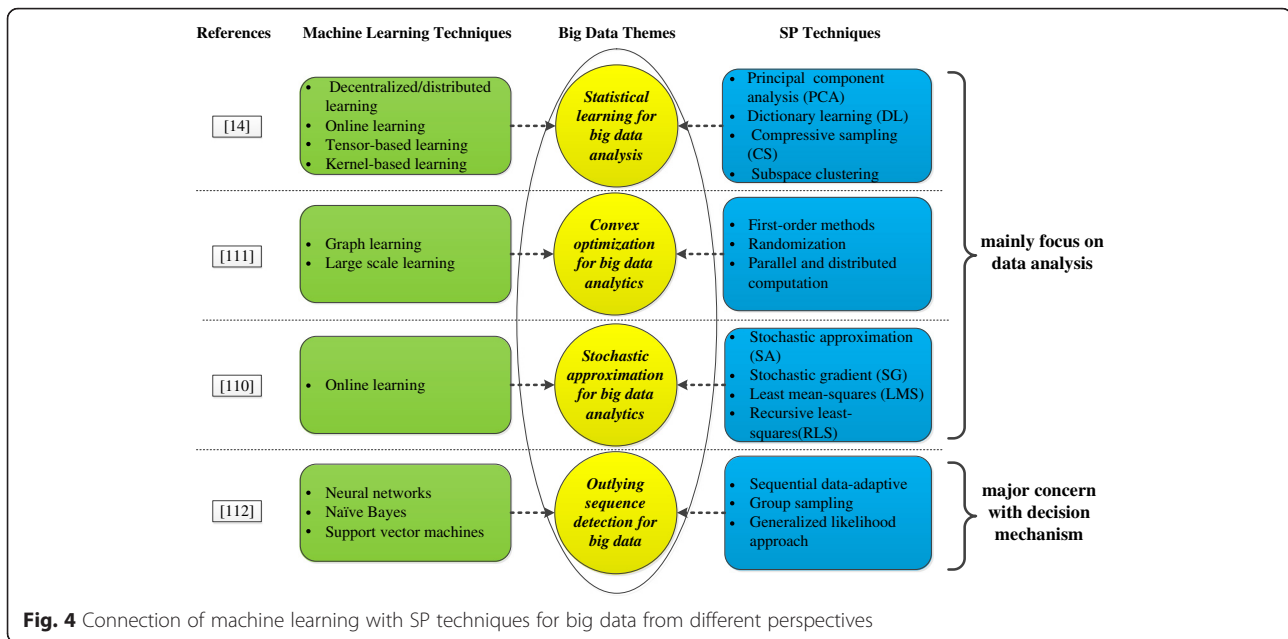
1.4.1 An overview of representative work

In this section, we analyze the relationships between machine learning and SP techniques for big data processing from four perspectives: (1) statistical learning for big data analysis, (2) convex optimization for big data analytics, (3) stochastic approximation for big data analytics, and (4) outlying sequence detection for big data. The diagram is summarized in Fig. 4. Several typical research papers are presented, which delineate the theoretical and algorithmic underpinnings together with the relevance of SP tools to the big data and also show the challenges and opportunities for SP research on large-scale data analytics.

- *Statistical learning for big data analysis*: There is no doubt this is an era of data deluge where learning from these large volumes of data by central processors and storage units seems infeasible. Therefore, the SP and statistical learning tools have to be re-examined. It is preferable to perform learning in real time for the advent of streaming data sources, typically without a chance to revisit past entries. In [14], the authors mainly focused on the modeling and optimization for big data analysis by using statistical learning tools. We can conclude from [14] that, from the SP and learning perspective, big data themes in terms of tasks, challenges, models, and optimization can be revealed as follows. SP-relevant big data tasks mainly comprise massive scale, outliers and missing values, real-time constraints, and cloud storage. There are great big data challenges we have to face, such as prediction and forecasting, cleansing and imputation, dimensionality reduction, regression, classification, and clustering. In terms of these tasks and challenges, outstanding models and optimization with the SP and learning techniques for big data include parallel and decentralized, time or data adaptive, robust, succinct, and sparse technologies.
- *Convex optimization for big data analytics*: While the importance of convex formulations and optimization has increased dramatically in the last decade and these formulations have been employed in a wide variety of signal processing applications, due to the data size of optimization problems that are too large to process locally in the context of big data, thus convex optimization needs reinvent itself. Cevher et al. [111] reviewed recent advances in convex optimization algorithms tailored for big data, having as ultimate goal to markedly reduce the computational, storage, and communication bottlenecks. For example, given a big data optimization problem formulated as

$$F^* = \min_x \{F(x) = f(x) + g(x); x \in \mathbb{R}^p\} \quad (6)$$

where f and g are convex functions. To obtain an optimal solution x^* of (6) and the required assumptions on f and g , in this article, the authors presented three efficient big data approximation techniques, including first-order methods, randomization and parallel and distributed computation. They mainly referred to the scalable, randomized, and parallel algorithms for big data analytics. In addition, for the optimization problem in (6), ADMM can provide a simple distributed algorithm to solve its composite form, by leveraging powerful augmented Lagrangian and dual decomposition techniques. Although there are two caveats for



ADMM, i.e., one is that closed-form solutions do not always exist and the other is that no convergence guarantees for more than two optimization objective terms, there are several recent solutions to address the two drawbacks, such as proximal gradient methods and parallel computing [111]. Specifically, from machine learning perspective, those bright techniques like scalable, parallel, and distributed mechanisms are also necessitated, and some applications of employing the recent convex optimization algorithms in learning methods such as support vector machines and graph learning have been appeared in recent years.

- Stochastic approximation for big data analytics:** Although many of online learning approaches were developed within the machine-learning discipline, they had strong connections with workhorse SP techniques. Reference [110] is a lecture note which presented recent advances in online learning for big data analytics, where the authors highlighted the relations and differences between online learning methods and some prominent statistical SP tools such as stochastic approximation (SA) and stochastic gradient (SG) algorithms. Through perusing [110], we can know that, on the one hand, the seminal works on SA, such as by Robbins–Monro and Widrow algorithms, and the workhorse behind several classical SP tools, such as LMS and RLS algorithms, carried rich potential in modern learning tasks for big data analytics. On the other hand, it was also demonstrated that online learning schemes together with random sampling or data sketching methods were expected to play instrumental roles in solving large-scale

optimization tasks. In summary, the recent advances in online learning methods and several SP techniques mentioned in this lecture note have the unique and complementary strengths with each other.

- Outlying sequence detection for big data:** As the data scale grows, so does the chance to involve outlying observations, which in turn motivates the demand for outlier-resilient learning algorithms scaling to large-scale application settings. In this context, data-driven outlying sequence detection algorithms have been proposed by some researchers. In [112], the authors investigated the robust sequential detection schemes for big data. In contrast to the aforementioned three articles [14, 110, 111] that mostly focus on big data analysis, this article paid more attention to the decision mechanisms. Outlier detection has immediate application in a broad range of contexts, particularly, for machine learning techniques, effective decision on the observations with categorizing them as normal or outlying are important for the improvement of learning performance. As mentioned in [112], the class of supervised outlier detection had been studied extensively under neural networks, naïve Bayes, and support vector machines.

1.4.2 The latest research progress

These representative literatures discussed in Section 1.4.1 provide us a lot of heuristic analysis on both machine learning and SP techniques for big data. Based on the ideas proposed in these works, many new studies are increasing continuously. In this section, we provide a review of the latest research progress which is based on these typical works mentioned above.

- *The latest progress based on [14]*: Based on the statistical learning tools for big data analysis proposed by Slavakis et al. in [14], a lot of new study work has emerged. For example, in [113], two distributed learning algorithms for training random vector functional-link (RVFL) networks through interconnected nodes were presented, where training data were distributed under a decentralized information structure. To tackle the huge-scale convex and nonconvex big data optimization problems, a novel parallel, hybrid random/deterministic decomposition scheme with the power of dictionary learning was investigated in [114]. In [87], the authors developed a low-complexity, real-time online algorithm for decomposing low-rank tensors with missing entries to deal with the incomplete streaming data, and the performance of the proposed subspace learning was also validated. All these new work presents the application of machine learning and SP technologies in processing big data well.
- *The latest progress based on [111]*: A broad class of machine learning and SP problems can be formally stated as optimization problem. Based on the idea of convex optimization for big data analytics in [111], a randomized primal-dual algorithm was proposed in [115] for composite optimization, which could be used in the framework of large-scale machine learning applications. In addition, a consensus-based decentralized algorithm for a class of nonconvex optimization problems was investigated in [116], with the application to dictionary learning.
- *The latest progress based on [110]*: Several classical SP tools such as the stochastic approximation methods, have carried rich potential for solving large-scale learning tasks under low computational expense. The SP and online learning techniques for big data analytics described in [110] provides a good research direction for future work. Based on this, in [117], the authors developed online algorithms for large-scale regressions with application to streaming big data. In addition, Slavakis and Giannakis further used accelerated stochastic approximation method with online and modular learning algorithms to deal with a large class of nonconvex data models [118].
- *The latest progress based on [112]*: The outlying sequence detection approach proposed in [112] provides a desirable solution to some big data application problems. In [119], the authors mainly investigated the big data analytics over the communication system with discussions about statistical analysis and machine learning techniques. The authors pointed out that one of the critically associated challenges ahead was how to detect outliers in the context of big data. It so happened

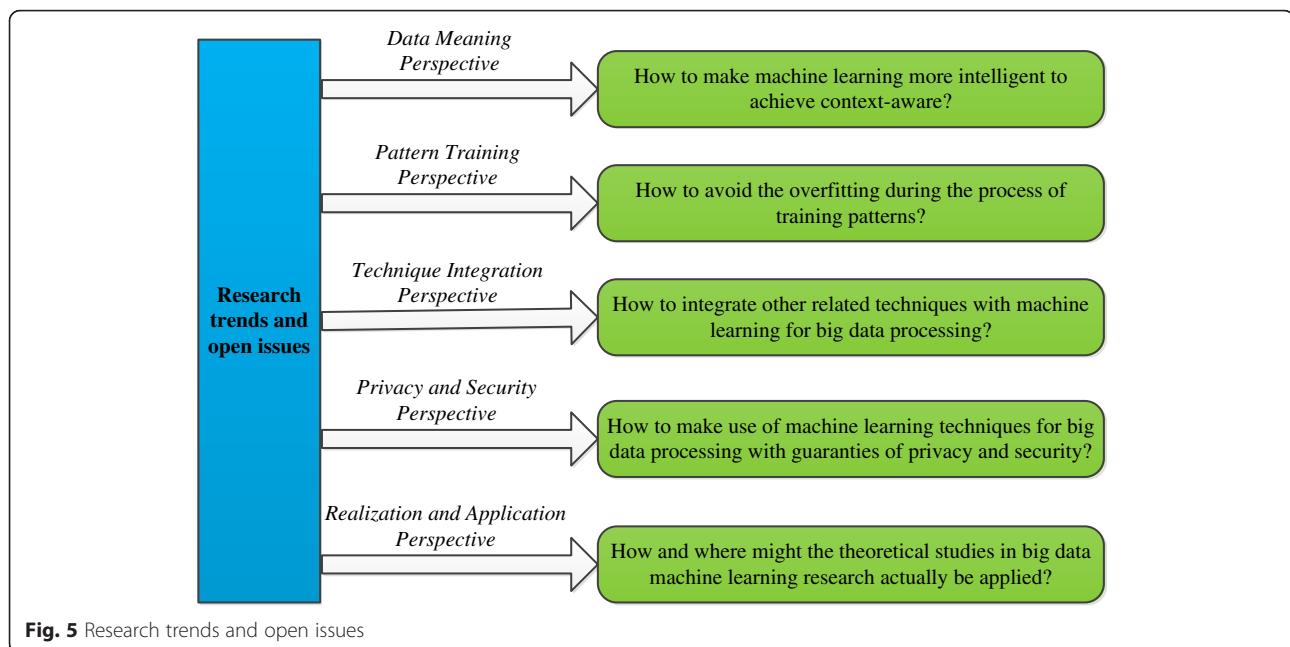
that the theoretic methodology described in [112] gave the answers.

To sum up, it can be seen from the above presented articles in Section 1.4.1 and Section 1.4.2 that the connection of machine learning with modern SP techniques is very strong. SP techniques are originally developed to analyze and handle discrete and continuous signals through using a set of methods from electrical engineering and applied mathematics. In contrast, machine learning research mainly focuses on the design and development of algorithms which allow computers to evolve behavior based on empirical data, whose major concern is to recognize complex patterns and make intelligent decisions based on data by automatically learning. Both the machine learning and SP techniques have the unique and complementary strengths for big data processing. Furthermore, combining SP and machine learning techniques to explore the emerging field of big data are expected to have a bright future. Quoting a sentence from [110], “Consequently, ample opportunities arise for the SP community to contribute in this growing and inherently cross-disciplinary field, spanning multiple areas across science and engineering”.

1.5 Research trends and open issues

While significant progress has been made in the last decade toward achieving the ultimate goal of making sense of big data by machine learning techniques, the consensus is that we are still not quite there. The efficient preprocessing mechanisms to make the learning system capable of dealing with big data and effective learning technologies to find out the rules to describe the data are still of urgent need. Therefore, some of the open issues and possible research trends are given in Fig. 5.

1. *Data meaning perspective*: Due to the fact that, nowadays, most data are dispersed to different regions, systems, or applications, the “meaning” of the collected data from various sources may not be exactly the same, which may significantly impact the quality of the machine learning results. Although the previous mentioned techniques such as transfer learning with the power of knowledge transfer and the cognition-assisted learning methods provide some possible solutions to this problem, it is obvious that they are absolutely not catholicons owing to the limitations of these techniques for achieving context-aware. Ontology, semantic web, and other related technologies seem to be preferred on this issue. Based on ontology modeling and semantic derivation, some valuable patterns or rules can be discovered as knowledge as well, which is a necessity for learning systems to be, or appear to be intelligent.



But the problem that arises now is, although the ontology and semantic web technologies can benefit the big data analysis, these two technologies are not mature enough, thus how to employ them in machine learning methods to process big data will be a meaningful research.

2. *Pattern training perspective:* In general, for most machine learning techniques, the more the training patterns are, the higher the accuracy rate of learning results is. However, a dilemma we have to face is that, on the one hand, the labeled patterns play a pivotal role for the learning algorithms; but on the other hand, labeling patterns is often expensive in terms of the computation time or cost, particularly for the large-scale streaming data, which is intractable. How many patterns are needed to train the classifier depends to a large extent on the desire to achieve a balance between cost and accuracy. Therefore, the so-called overfitting is another critical open issue.
3. *Technique integration perspective:* Once mentioning big data processing, we always like to put data mining, KDD, SP, cloud computing, and machine learning techniques together, partially because these issues and their products may play principal roles for extracting valuable information from massive data, and partially because they have strong ties with each other. It is important to note that each approach has its own merits and faults. That is to say, to get more values out of the big data, a composite model is more needed. As a result, how to integrate several related techniques with machine learning will also become a further research trend.

4. *Privacy and security perspective:* The concern of data privacy has become extremely serious with using data mining and machine learning technologies to analyze personal information in order to produce relevant or accurate results. For example, in order to increase the volume and revenue of sales, some companies today try to collect as many personal data of consumers as possible from various kinds of sources or devices and then use data mining and machine learning methods to find highly interconnected information which is conducive to make marketing tactics. However, if all pieces of the information about a person were dug out through the mining and learning technologies and put together, any privacy about that individual instantly would disappear, which will make most people uncomfortable, and even frightened. Thus, an efficient and effective method needs to preserve the performance of mining and learning while protecting the personal information. Hence, how to make use of data mining and machine learning techniques for big data processing with guaranties of privacy and security is very worthy of study.
5. *Realization and application perspective:* The ultimate goal of groping for various learning methods to handle big data is to provide better environment for people; thus, more attention should be focused on building the bridge from theory to practice. For instance, how and where might the theoretical studies in big data machine learning research actually be applied?

2 Conclusions

Big data are now rapidly expanding in all science and engineering domains. Learning from these massive data is expected to bring significant opportunities and transformative potential for various sectors. However, most traditional machine learning techniques are not inherently efficient or scalable enough to handle the data with the characteristics of large volume, different types, high speed, uncertainty and incompleteness, and low value density. In response, machine learning needs to reinvent itself for big data processing. This paper began with a brief review of conventional machine learning algorithms, followed by several current advanced learning methods. Then, a discussion about the challenges of learning with big data and the corresponding possible solutions in recent researches was given. In addition, the connection of machine learning with modern signal processing technologies was analyzed through studying several latest representative research papers. To stimulate more interests for the audience of the paper, at last, open issues and research trends were presented.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We gratefully acknowledge the financial support from the National Natural Science Foundation of China (Grant No. 61301160 and No. 61172062).

Received: 31 August 2015 Accepted: 22 April 2016

Published online: 28 May 2016

References

1. A Sandryhaila, JMF Moura, Big data analysis with signal processing on graphs: representation and processing of massive data sets with irregular structure. *IEEE Signal Proc Mag* **31**(5), 80–90 (2014)
2. J Gantz, D Reinsel, *Extracting value from chaos* (EMC, Hopkinton, 2011)
3. J Gantz, D Reinsel, *The digital universe decade—are you ready* (EMC, Hopkinton, 2010)
4. D Che, M Safran, Z Peng, From big data to big data mining: challenges, issues, and opportunities, in *Proceedings of the 18th International Conference on DASFAA* (Wuhan, 2013), pp. 1–15
5. M Chen, S Mao, Y Liu, Big data: a survey. *Mobile Netw Appl* **19**(2), 171–209 (2014)
6. H Hu, Y Wen, T Chua, X Li, Toward scalable systems for big data analytics: a technology tutorial. *IEEE Access* **2**, 652–687 (2014)
7. J Manyika, M Chui, B Brown, J Bughin, R Dobbs, C Roxburgh, AH Byers, *Big data: the next frontier for innovation, competition, and productivity* (McKinsey Global Institute, USA, 2011)
8. Q Wu, G Ding, Y Xu, S Feng, Z Du, J Wang, K Long, Cognitive internet of things: a new paradigm beyond connection. *IEEE Internet Things J* **1**(2), 129–143 (2014)
9. CW Tsai, CF Lai, MC Chiang, LT Yang, Data mining for internet of things: a survey. *IEEE Commun Surv Tut* **16**(1), 77–97 (2014)
10. A Imran, A Zoha, Challenges in 5G: how to empower SON with big data for enabling 5G. *IEEE Netw* **28**(6), 27–33 (2014)
11. X Wu, X Zhu, G Wu, W Ding, Data mining with big data. *IEEE Trans Knowl Data Eng* **26**(1), 97–107 (2014)
12. A Rajaraman, JD Ullman, *Mining of massive data sets* (Cambridge University Press, Oxford, 2011)
13. XW Chen, X Lin, Big data deep learning: challenges and perspectives. *IEEE Access* **2**, 514–525 (2014)
14. K Slavakis, GB Giannakis, G Mateos, Modeling and optimization for big data analytics: (statistical) learning tools for our era of data deluge. *IEEE Signal Proc Mag* **31**(5), 18–31 (2014)
15. TM Mitchell, *Machine learning* (McGraw-Hill, New York, 1997)
16. S Russell, P Norvig, *Artificial intelligence: a modern approach* (Prentice-Hall, Englewood Cliffs, 1995)
17. V Cherkassky, FM Mulier, *Learning from data: concepts, theory, and methods* (John Wiley & Sons, New Jersey, 2007)
18. TM Mitchell, *The discipline of machine learning* (Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006)
19. C Rudin, KL Wagstaff, Machine learning for science and society. *Mach Learn* **95**(1), 1–9 (2014)
20. CM Bishop, *Pattern recognition and machine learning* (Springer, New York, 2006)
21. B Adam, IFC Smith, F Asce, Reinforcement learning for structural control. *J Comput Civil Eng* **22**(2), 133–139 (2008)
22. N Jones, Computer science: the learning machines. *Nature* **505**(7482), 146–148 (2014)
23. J Langford, Tutorial on practical prediction theory for classification. *J Mach Learn Res* **6**(3), 273–306 (2005)
24. R Bekkerman, EY Ran, N Tishby, Y Winter, Distributional word clusters vs. words for text categorization. *J Mach Learn Res* **3**, 1183–1208 (2003)
25. Y Bengio, A Courville, P Vincent, Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal* **35**(8), 1798–1828 (2012)
26. F Huang, E Yates, Biased representation learning for domain adaptation, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island, 2012), pp. 1313–1323
27. W Tu, S Sun, Cross-domain representation-learning framework with combination of class-separate and domain-merge objectives, in *Proceedings of the 1st International Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining* (Beijing, 2012), pp. 18–25
28. S Li, C Huang, C Zong, Multi-domain sentiment classification with classifier combination. *J Comput Sci Technol* **26**(1), 25–33 (2011)
29. F Huang, E Yates, Exploring representation-learning approaches to domain adaptation, in *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing* (Uppsala, 2010), pp. 23–30
30. A Bordes, X Glorot, JWAY Bengio, Joint learning of words and meaning representations for open-text semantic parsing, in *Proceedings of 15th International Conference on Artificial Intelligence and Statistics* (La Palma, 2012), pp. 127–135
31. N. Boulanger-Lewandowski, Y. Bengio, P. Vincent, Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription. arXiv preprint (2012). arXiv:1206.6392
32. K Dwivedi, K Biswaranjan, A Sethi, Drowsy driver detection using representation learning, in *Proceedings of the IEEE International Advance Computing Conference* (Gurgaon, 2014), pp. 995–999
33. D Yu, L Deng, Deep learning and its applications to signal and information processing. *IEEE Signal Proc Mag* **28**(1), 145–154 (2011)
34. I Arel, DC Rose, TP Karnowski, Deep machine learning—a new frontier in artificial intelligence research. *IEEE Comput Intell Mag* **5**(4), 13–18 (2010)
35. Y Bengio, Learning deep architectures for AI. *Foundations Trends Mach Learn* **2**(1), 1–127 (2009)
36. R Collobert, J Weston, L Bottou, M Karlen, K Kavukcuoglu, P Kuksa, Natural language processing (almost) from scratch. *J Mach Learn Res* **12**, 2493–2537 (2011)
37. P Le Callet, C Viard-Gaudin, D Barba, A convolutional neural network approach for objective video quality assessment. *IEEE Trans Neural Netw* **17**(5), 1316–1327 (2006)
38. GE Dahl, D Yu, L Deng, A Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans Audio Speech Lang Proc* **20**(1), 30–42 (2012)
39. G Hinton, L Deng, Y Dong, GE Dahl, A Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, TN Sainath, B Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Proc Mag* **29**(6), 82–97 (2012)
40. DC Ciresan, U Meier, LM Gambardella, J Schmidhuber, Deep, big, simple neural nets for handwritten digit recognition. *Neural Comput* **22**(12), 3207–3220 (2010)
41. Y Wang, D Yu, Y Ju, A Acero, Voice search, in *Language understanding: systems for extracting semantic information from speech* (Wiley, New York, 2011)
42. D Peteiro-Barral, B Guijarro-Berdiñas, A survey of methods for distributed machine learning. *Progress in Artificial Intelligence* **2**(1), 1–11 (2012)
43. H Zheng, SR Kulkarni, HV Poor, Attribute-distributed learning: models, limits, and algorithms. *IEEE Trans Signal Process* **59**(1), 386–398 (2011)

44. H Chen, T Li, C Luo, SJ Horng, G Wang, A rough set-based method for updating decision rules on attribute values' coarsening and refining. *IEEE Trans Knowl Data Eng* **26**(12), 2886–2899 (2014)
45. J Chen, C Wang, R Wang, Using stacked generalization to combine SVMs in magnitude and shape feature spaces for classification of hyperspectral data. *IEEE Trans Geosci Remote* **47**(7), 2193–2205 (2009)
46. E Leyva, A González, R Pérez, A set of complexity measures designed for applying meta-learning to instance selection. *IEEE Trans Knowl Data Eng* **27**(2), 354–367 (2014)
47. M Sarnovsky, M Wronc, Distributed boosting algorithm for classification of text documents, in *Proceedings of the 12th IEEE International Symposium on Applied Machine Intelligence and Informatics (SAMII)* (Herlany, 2014), pp. 217–220
48. SR Upadhyaya, Parallel approaches to machine learning—a comprehensive survey. *J Parallel Distr Com* **73**(3), 284–292 (2013)
49. R Bekkerman, M Bilenko, J Langford, *Scaling up machine learning: parallel and distributed approaches* (Cambridge University Press, Oxford, 2011)
50. EW Xiang, B Cao, DH Hu, Q Yang, Bridging domains using world wide knowledge for transfer learning. *IEEE Trans Knowl Data Eng* **22**(6), 770–783 (2010)
51. SJ Pan, Q Yang, A survey on transfer learning. *IEEE Trans Knowl Data Eng* **22**(10), 1345–1359 (2010)
52. W Fan, I Davidson, B Zadrozny, PS Yu, An improved categorization of classifier's sensitivity on sample selection bias, in *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM)* (Brussels, 2012), pp. 605–608
53. J Gao, W Fan, J Jiang, J Han, Knowledge transfer via multiple model local structure mapping, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, 2008), pp. 283–291
54. C Wang, S Mahadevan, Manifold alignment using procrustes analysis, in *Proceedings of the 25th International Conference on Machine Learning (ICML)* (Helsinki, 2008), pp. 1120–1127
55. X Ling, W Dai, GR Xue, Q Yang, Y Yu, Spectral domain-transfer learning, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, 2008), pp. 488–496
56. R Raina, AY Ng, D Koller, 2006, Constructing informative priors using transfer learning, in *Proceedings of the 23rd International Conference on Machine Learning (ICML)* (Pittsburgh, 2006), pp. 713–720
57. J Zhang, Deep transfer learning via restricted Boltzmann machine for document classification, in *Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops (ICMLA)* (Honolulu, 2011), pp. 323–326
58. 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)
59. B Settles, *Active learning literature survey* (University of Wisconsin, Madison, 2010)
60. MM Crawford, D Tuia, HL Yang, Active learning: any value for classification of remotely sensed data? *P IEEE* **101**(3), 593–608 (2013)
61. MM Haque, LB Holder, MK Skinner, DJ Cook, Generalized query-based active learning to identify differentially methylated regions in DNA. *IEEE ACM Trans Comput Bi* **10**(3), 632–644 (2013)
62. D Tuia, M Volpi, L Copa, M Kanevski, J Munoz-Mari, A survey of active learning algorithms for supervised remote sensing image classification. *IEEE J Sel Top Sign Proces* **5**(3), 606–617 (2011)
63. G Ding, Q Wu, YD Yao, J Wang, Y Chen, Kernel-based learning for statistical signal processing in cognitive radio networks. *IEEE Signal Proc Mag* **30**(4), 126–136 (2013)
64. C Li, M Georgiopoulos, GC Anagnostopoulos, A unifying framework for typical multitask multiple kernel learning problems. *IEEE Trans Neur Net Lear Syst* **25**(7), 1287–1297 (2014)
65. G Montavon, M Braun, T Krueger, KR Muller, Analyzing local structure in kernel-based learning: explanation, complexity, and reliability assessment. *IEEE Signal Proc Mag* **30**(4), 62–74 (2013)
66. K Slavakis, S Theodoridis, I Yamada, Online kernel-based classification using adaptive projection algorithms. *IEEE Trans Signal Process* **56**(7), 2781–2796 (2008)
67. S Theodoridis, K Slavakis, I Yamada, Adaptive learning in a world of projections. *IEEE Signal Proc Mag* **28**(1), 97–123 (2011)
68. K Slavakis, S Theodoridis, I Yamada, Adaptive constrained learning in reproducing kernel Hilbert spaces: the robust beamforming case. *IEEE Trans Signal Process* **57**(12), 4744–4764 (2009)
69. K Slavakis, P Bouboulis, S Theodoridis, Adaptive multiregression in reproducing kernel Hilbert spaces: the multiaccess MIMO channel case. *IEEE Trans Neural Netw Learn Syst* **23**(2), 260–276 (2012)
70. KR Müller, S Mika, G Rätsch, K Tsuda, B Schölkopf, An introduction to kernel-based learning algorithms. *IEEE Trans Neural Netw* **12**(2), 181–201 (2001)
71. TH Davenport, P Barth, R Bean, How “big data” is different. *MIT Sloan Manage Rev* **54**(1), 22–24 (2012)
72. F Andersson, M Carlsson, JY Tournet, H Wendt, A new frequency estimation method for equally and unequally spaced data. *IEEE Trans Signal Process* **62**(21), 5761–5774 (2014)
73. F Lin, M Fardad, MR Jovanovic, Design of optimal sparse feedback gains via the alternating direction method of multipliers. *IEEE Trans Automat Contr* **58**(9), 2426–2431 (2013)
74. S Boyd, N Parikh, E Chu, B Peleato, J Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations Trends Mach Learn* **3**(1), 1–122 (2011)
75. J Dean, S Ghemawat, MapReduce: simplified data processing on large clusters. *Commun ACM* **51**(1), 107–113 (2008)
76. J Dean, S Ghemawat, MapReduce: a flexible data processing tool. *Commun ACM* **53**(1), 72–77 (2010)
77. C Chu, SK Kim, YA Lin, Y Yu, G Bradski, AY Ng, K Olukotun, Map-reduce for machine learning on multicore, in *Proceedings of 20th Annual Conference on Neural Information Processing Systems (NIPS)* (Vancouver, 2006), pp. 281–288
78. M Armbrust, A Fox, R Griffith, AD Joseph, R Katz, A Konwinski, G Lee, D Patterson, A Rabkin, I Stoica, M Zaharia, A view of cloud computing. *Commun ACM* **53**(4), 50–58 (2010)
79. MD Dikaiakos, D Katsaros, P Mehra, G Pallis, A Vakali, Cloud computing: distributed internet computing for IT and scientific research. *IEEE Internet Comput* **13**(5), 10–13 (2009)
80. Y Low, D Bickson, J Gonzalez, C Guestrin, A Kyrola, JM Hellerstein, Distributed GraphLab: a framework for machine learning and data mining in the cloud. *Proc VLDB Endow* **5**(8), 716–727 (2012)
81. M Lenzerini, Data integration: a theoretical perspective, in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (Madison, 2002), pp. 233–246
82. A Halevy, A Rajaraman, J Ordille, Data integration: the teenage years, in *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB)* (Seoul, 2006), pp. 9–16
83. Q Wu, G Ding, J Wang, YD Yao, Spatial-temporal opportunity detection for spectrum-heterogeneous cognitive radio networks: two-dimensional sensing. *IEEE Trans Wirel Commun* **12**(2), 516–526 (2013)
84. N Srivastava, RR Salakhutdinov, Multimodal learning with deep boltzmann machines, in *Proceedings of Neural Information Processing Systems Conference (NIPS)* (Nevada, 2012), pp. 2222–2230
85. Y Sun, S Todorovic, S Goodison, Local-learning-based feature selection for high-dimensional data analysis. *IEEE Trans Pattern Anal Mach Intell* **32**(9), 1610–1626 (2010)
86. LJP van der Maaten, EO Postma, HJ van den Herik, Dimensionality reduction: a comparative review. *J Mach Learn Res* **10**(1–41), 66–71 (2009)
87. M Mardani, G Mateos, GB Giannakis, Subspace learning and imputation for streaming big data matrices and tensors. *IEEE Trans Signal Process* **63**(10), 2663–2677 (2015)
88. K Mohan, M Fazel, New restricted isometry results for noisy low-rank recovery, in *Proceedings of IEEE International Symposium on Information Theory Proceedings (ISIT)* (Texas, 2010), pp. 1573–1577
89. EJ Candès, X Li, Y Ma, J Wright, Robust principal component analysis? *J ACM* **58**(3), 1–37 (2011)
90. Z Lin, R Liu, Z Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in *Proceedings of Neural Information Processing Systems Conference (NIPS)* (Granada, 2011), pp. 612–620
91. S Shalev-Shwartz, Online learning and online convex optimization. *Foundations Trends Mach Learn* **4**, 107–194 (2011)
92. J Wang, P Zhao, SC Hoi, R Jin, Online feature selection and its applications. *IEEE Trans Knowl Data Eng* **26**(3), 698–710 (2014)
93. J Kivinen, AJ Smola, RC Williamson, Online learning with kernels. *IEEE Trans Signal Process* **52**(8), 2165–2176 (2004)
94. M Bilenko, S Basil, M Sahami, Adaptive product normalization: using online learning for record linkage in comparison shopping, in *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM)* (Texas, 2005), p. 8

95. GB Huang, QY Zhu, CK Siew, Extreme learning machine: theory and applications. *Neurocomputing* **70**(1), 489–501 (2006)
96. S Ding, X Xu, R Nie, Extreme learning machine and its applications. *Neural Comput Appl* **25**(3–4), 549–556 (2014)
97. N Tatbul, Streaming data integration: challenges and opportunities, in *Proceedings of the 26th IEEE International Conference on Data Engineering Workshops (ICDEW)* (Long Beach, 2010), pp. 155–158
98. DJ Abadi, Y Ahmad, M Balazinska, U Cetintemel, M Cherniack, JH Hwang, W Lindner, A Maskey, A Rasin, E Ryvkina, N Tatbul, Y Xing, SB Zdonik, The design of the borealis stream processing engine, in *Proceedings of the Second Biennial Conference on Innovative Data Systems Research (CIDR)* (Asilomar, 2005), pp. 277–289
99. L Neumeyer, B Robbins, A Nair, A Kesari, S4: Distributed stream computing platform, in *Proceedings of IEEE International Conference on Data Mining Workshops (ICDMW)* (Sydney, 2010), pp. 170–177
100. K Goodhope, J Koshy, J Kreps, N Narkhede, R Park, J Rao, VY Ye, Building LinkedIn's real-time activity data pipeline. *IEEE Data Eng Bull* **35**(2), 33–45 (2012)
101. W Yang, X Liu, L Zhang, LT Yang, Big data real-time processing based on storm, in *Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (Melbourne, 2013), pp. 1784–1787
102. B SkieS, Streaming big data processing in datacenter clouds. *IEEE Cloud Comput* **1**, 78–83 (2014)
103. A Baldominos, E Albacete, Y Saez, P Isasi, A scalable machine learning online service for big data real-time analysis, in *Proceedings of IEEE Symposium on Computational Intelligence in Big Data (CIBD)* (Orlando, 2014), pp. 1–8
104. NY Soltani, SJ Kim, GB Giannakis, Real-time load elasticity tracking and pricing for electric vehicle charging. *IEEE Trans Smart Grid* **6**(3), 1303–1313 (2014)
105. S Tsang, B Kao, KY Yip, WS Ho, SD Lee, Decision trees for uncertain data. *IEEE Trans Knowl Data Eng* **23**(1), 64–78 (2011)
106. F Nie, H Wang, X Cai, H Huang, C Ding, Robust matrix completion via joint Schatten p-norm and lp-norm minimization, in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM)* (Brussels, 2012), p. 566
107. G Ding, J Wang, Q Wu, L Zhang, Y Zou, YD Yao, Y Chen, Robust spectrum sensing with crowd sensors. *IEEE Trans Commun* **62**(9), 3129–3143 (2014)
108. U Fayyad, G Piatetsky-Shapiro, P Smyth, From data mining to knowledge discovery in databases. *AI Mag* **17**(3), 37–54 (1996)
109. J Kelly III, S Hamm, *Smart machines: IBM's Watson and the era of cognitive computing* (Columbia University Press, New York, 2013)
110. K Slavakis, SJ Kim, G Mateos, GB Giannakis, Stochastic approximation vis-a-vis online learning for big data analytics. *IEEE Signal Proc Mag* **31**(6), 124–129 (2014)
111. V Cevher, S Becker, M Schmidt, Convex optimization for big data: scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Proc Mag* **31**(5), 32–43 (2014)
112. A Tajer, W Veeravalli, HV Poor, Outlying sequence detection in large data sets: a data-driven approach. *IEEE Signal Proc Mag* **31**(5), 44–56 (2014)
113. S Scardapane, D Wang, M Panella, A Uncini, Distributed learning for random vector functional-link networks. *Inf Sci* **301**, 271–284 (2015)
114. A Daneshmand, F Facchinei, V Kungurtsev, G Scutari, Hybrid random/deterministic parallel algorithms for nonconvex big data optimization. *IEEE Trans Signal Process* **63**(15), 3914–3929 (2015)
115. P. Bianchi, W. Hachem, F. Lutzeler, A stochastic coordinate descent primal-dual algorithm and applications to large-scale composite optimization. *arXiv preprint* (2014). [arXiv:1407.0898](https://arxiv.org/abs/1407.0898)
116. HT Wai, TH Chang, A Scaglione, A consensus-based decentralized algorithm for non-convex optimization with application to dictionary learning, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (South Brisbane, 2015), pp. 3546–3550
117. D. Berberidis, V. Kekatos, G.B. Giannakis, Online censoring for large-scale regressions with application to streaming big data. *arXiv preprint* (2015). [arXiv:1507.07536](https://arxiv.org/abs/1507.07536)
118. K. Slavakis, G.B. Giannakis, Per-block-convex data modeling by accelerated stochastic approximation. *arXiv preprint* (2015). [arXiv:1501.07315](https://arxiv.org/abs/1501.07315)
119. KC Chen, SL Huang, L Zheng, HV Poor, Communication theoretic data analytics. *IEEE J Sel Areas Commun* **33**(4), 663–675 (2015)
120. J Zheng, F Shen, H Fan, J Zhao, An online incremental learning support vector machine for large-scale data. *Neural Comput Appl* **22**(5), 1023–1035 (2013)
121. C Ghosh, C Cordeiro, DP Agrawal, M Bhaskara Rao, Markov chain existence and hidden Markov models in spectrum sensing, in *Proceedings of the IEEE International Conference on Pervasive Computing & Communications (PERCOM)* (Galveston, 2009), pp. 1–6
122. K Yue, Q Fang, X Wang, J Li, W Wei, A parallel and incremental approach for data-intensive learning of Bayesian networks. *IEEE Trans Cybern* **99**, 1–15 (2015)
123. X Dong, Y Li, C Wu, Y Cai, A learner based on neural network for cognitive radio, in *Proceedings of the 12th IEEE International Conference on Communication Technology (ICCT)* (Nanjing, 2010), pp. 893–896
124. A El-Hajj, L Safatly, M Bkassiny, M Hussein, Cognitive radio transceivers: RF, spectrum sensing, and learning algorithms review. *Int J Antenn Propag* **11**(5), 479–482 (2014)
125. M Bkassiny, SK Jayaweera, Y Li, Multidimensional dirichlet process-based non-parametric signal classification for autonomous self-learning cognitive radios. *IEEE Trans Wirel Commun* **12**(11), 5413–5423 (2013)
126. A Galindo-Serrano, L Giupponi, Distributed Q-learning for aggregated interference control in cognitive radio networks. *IEEE Trans Veh Technol* **59**(4), 1823–1834 (2010)
127. TK Das, A Gosavi, S Mahadevan, N Marchallick, Solving semi-markov decision problems using average reward reinforcement learning. *Manage Sci* **45**(4), 560–574 (1999)
128. RS Sutton, Learning to predict by the methods of temporal differences. *Mach Learn* **3**(1), 9–44 (1988)
129. S Singh, T Jaakkola, ML Littman, C Szepesvári, Convergence results for single-step on-policy reinforcement-learning algorithms. *Mach Learn* **38**, 287–308 (2000)

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)