International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 21 (2017) pp. 11691-11699

Machine Learning and Statistical Approaches for Big Data: Issues, Challenges and Research Directions

D. Saidulu
Associate Professor, Department of Computer Science and Engineering,
Guru Nanak Institutions Technical Campus, Hyderabad, India.

Orcid: 0000-0001-5184-5634

Dr. R. Sasikala
Associate Professor, School of Computer Science and Engineering,
Vellore Institute of Technology,Vellore, India

Abstract
Today, as we are observing, massive sized and complex structured data is becoming available from variety of diverse sources, organizations are making attempt to utilize these plentiful resources for the purpose of enhance innovation, increase decisional and operational efficiency. Machine learning is a kind of artificial intelligence method to discover knowledge for making intelligent decisions. Big Data has vast impacts on scientific discoveries and value creation.

This paper presents an extensive literature study and review of latest advances, developments and new methodologies in researches on machine learning for processing big data. We have discussed various types of data types, learning methods, vital issues in big data processing and application of machine learning approaches in big data. Finally, we have outlined some open problems in this domain and our further research aims and directions.

Keywords: Machine learning, Data mining, Big data, Data analysis, Distributed computing, Knowledge discovery.

INTRODUCTION

Big data are expanding in a rapid manner in all engineering disciplines and science domains. Volume of data explodes at high rate today as a result in advancements of "Web technologies, social media, and mobile devices" [2] [4]. For eg., Twitter use to process over 70 million tweets daily, through producing over 8TB in daily manner [50]. According to one research estimation, there will around 30 billion computing machines, connecting each other, by 2020 [51]. Big Data employs amazing potential for trade value in diverse fields like – "health sector, biology, medicine transportation, online advertising and financial services" [47] [52] [20]. Though, traditional strategies struggles when deal with this large data. Learning from massively large data brings significant opportunities for numerous sectors. Still, most of these routines are not much practical or scalable enough [39] [48]. Therefore, ML demands to deeply discover itself for processing big data. According to a study by Oracle Company, around 90% of the world's knowledge data is held in unstructured form. [11] [17] [22] Big data may be explained in terms of three traits - velocity, volume and variety. Variety meant for heterogeneous nature, Velocity meant for the frequency at which data is being captured, and Volume meant for size of data (PB, EB and TB). Machine learning algorithms categorize the learning task in two types i.e. Supervised learning and Unsupervised learning. Mining of big data and knowledge discovery [6] [41] is the process of an efficient extraction of implicit, relevant, previously unknown, potentially useful (rules, regularities, patterns, constraints) from incomplete, noisy, random and unstructured data in large web databases. The general process is represented as diagram below: -

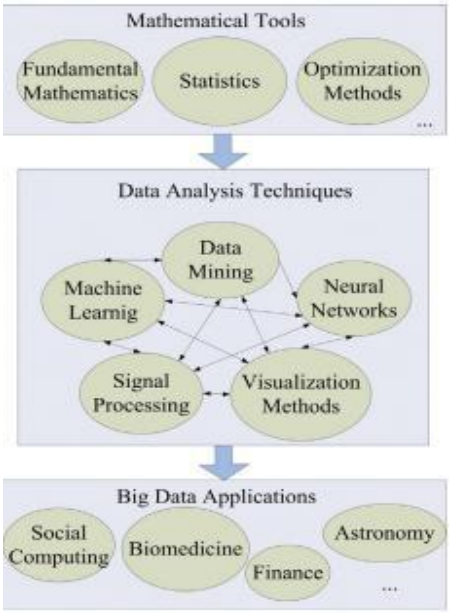

Figure.1: Tools for Big data

TYPES OF LEARNING METHODS

This subsection presents some recent learning methods that may play vital role in solving the big data problems.

1) Kernel-based learning: Kernel-based learning is proven to be very dominant methodology to efficiently enhance the computational capacity [39]. The notable advantage of this method is that both linear as well as non-linear vector kernel functional methods are present to deal with the non-linearity of data in N-dimensional feature space.

2) Depiction based learning: This kind of learning [59], is a solution to study valuable representations of the raw data. It is comparatively simpler to get knowledge information

while processing through classifiers [60]. Some variants of representational learning [61] [60] [62] are evolved in past years.

3) Active learning: This learning chooses a subset of an unstructured and critical occurrence for purpose of labeling [67]. The active learner obtains larger accuracy using reduced number of occurrences.

4) Deep learning: These designs take more complicated, compartmented statistical patterns of inputs and manages to be robust for new fields as compare to traditional learning systems. "Deep belief networks (DBNs)" [63] [64] and" convolutional neural networks (CNNs)" are two deep learning methodologies.

5) Transfer learning: The prime intention of transfer learning is to derive knowledge features from input source and later implement the knowledge to the target task [66]. The main benefit is that it can efficiently apply knowledge, which has been learned previously in order to find solution for new problems in fast manner.

6) Parallel & Distributed learning: The data which is avail-able in incomplete, inconsistent and unstructured format, is first pre-processed, then cluster forming is done [65]. Count of such distributed clusters is performed. Further one processing thread is assigned to each cluster in order to perform multi-threading in parallel and distributed manner.

Vital issues of machine learning for Big data

This section presents a review about the critical concerns of machine learning procedures for big data from diverse view-points, an overall scenario is presented in Fig-2. It includes

(i) learning for massive scaled data, (ii) learning for diverse structured data, (iii) learning for high frequency streamed data, (iv) learning for imprecise and incomplete data, (v) learning for deriving valuable knowledge from massive sized volumes of data.

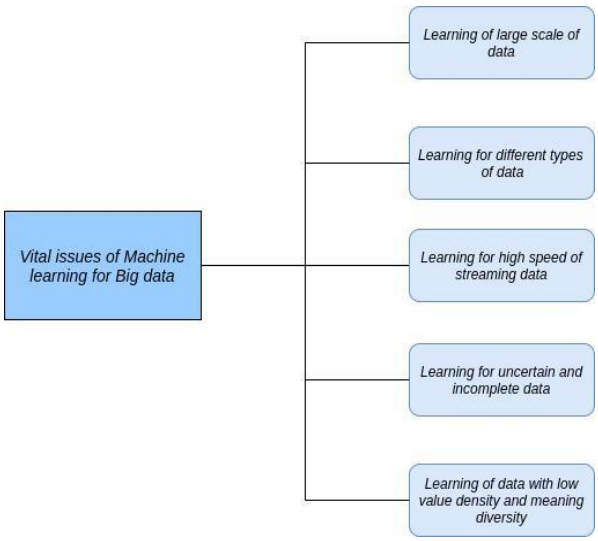

Figure.2: Learning methods for Big data

1. Learning for massive sized data: Considering only digital information, every day, Google processes approx. 24 PB data [48] [49]. Under modern development courses, data analyzed by big companies will unquestionably cross this petabyte magnitude.

   We are presently swimming in a deep and expanding ocean of data which is too bulky to train ML algorithm. Though, distributed and parallel frameworks are preferred. Cloud computing and MapReduce-assisted learning methods [68] [69] are another progress aspects which deal with core challenges of big data. It can improve computing and storage capacity through cloud infrastructure.

2. Learning for different structures of data: Immense variety of data is another aspect, that addresses big data interesting as well as challenging. It resulted of the aspect that data usually collected from diverse sources and are of varying types. Structured, semi-structured or fully unstructured data sources stimulate formation of heterogeneous, high-dimensionality, and nonlinear data [48] [23] e.g. global environment patterns, astronomical spectra, and human gene patterns with varying representation patterns.

   To deal such huge-dimensional data, reduction in dimensionality is an adequate solution through obtaining meaningful lower-dimensional constructions. Common procedures are to apply feature extraction in order to reduce dimensions.

3. Learning for high speed of streaming data: Speed or velocity really matters in big data scenario. In the time-sensitive cases like earthquake prediction, stock market prediction etc., the inherent value of data is depending

upon factor of data freshness which requires to be treated in a real-time fashion. Other challenging problem connected with high velocity is that data usually are not stationary, which requires learning procedures to determine the data as stream. The inherent superiority of streamed processing technology [2] [41] [70] been observed out compared with model of batch-processing.

4. Learning for imprecise and incomplete data: With the unmitigated sized data, the accuracy of the origin data instantly become a problem, due to data properties are not entirely verifiable. Hence, we include it as the next important problem [51] to highlight the significance of addressing as well as maintaining the incompleteness and uncertainty on data quality. Since incompleteness and inconsistency is a notable problem which surely affects the accuracy of further classification procedure. Deep learning [22] [23] is an approach to tackle with this issue.

MOTIVATION TO THE PROBLEM

As more scenarios e.g. global economy, society administration, national security involves Big Data problems, traditional strategies struggles when deal with this large data. Learning from massively large data brings significant opportunities for numerous sectors. Still, most of these routines are not much practical or scalable enough [20] [47]. From massive amount of available data, fetching (deriving) structured, useful and relevant knowledge is a significant as well as hard task in domain of big data processing.

Most of the traditional machine learning(ML) techniques are lacking computational efficiency, practicality or scalability to handle the data with traits of massive volume, varying types, great speed, uncertainty, inconsistency and incompleteness [39] [48]. So, to discover more optimal techniques which can process huge sized unstructured data efficiently are much desired.

ORGANIZATION ORDER OF THE PAPER

In rest of the paper, section 2 discusses some preliminaries requires in this domain. An extensive literature survey is presented in section 3. Section 4 summarizes some significant big data management tools. Some of the open issues in big data analytics along with our further research directions are given in section 5. Finally, section 6 concludes the paper.

Data Mining and Machine Learning Techniques

This section summarizes the mathematical, statistical techniques that are very useful while performing data mining or machine learning.

Hadoop HDFS

HDFS is Hadoop's storage layer which provides the high availability, fault tolerance and reliability. It is probable that worlds 75% of data will be stored in Hadoop HDFS by the end of 2017. Apache Hadoop HDFS is a kind of distributed file system(DFS) which affords redundant storage space for caching files which are enormous in sizes; files which are in the range of TB and PB. Files are split into blocks and diffused across junctions in a cluster. After that each block is replicated. Hence suppose a machine goes down or goes crashed, then in that cases, also we can effortlessly retrieve and access our data from different devices. Hence it is extremely fault-tolerant. HDFS gives faster file read and writes mechanism, as data is saved in different nodes inside a cluster.

Artificial Neural Network

It is a kind of classifier, whose model design structure and functionality is somewhat similar to human brain structure algorithmic model. [57] For classification problem, the specific structure of neural network changes. First, the training is carried out for ANN, where the topology and number of network nodes present in the hidden layer are decided. Unlike SVM, there is no phenomenon i.e. n-dimensional planes and hyperplanes. Still, training of data sets process here is time taking, produces less accurate and efficient results also.

SUPPORT VECTOR REGRESSION

As we know that the classification procedure falls into one of the category, either supervised or unsupervised classification. So, in the area of machine learning, support vector networks are supervised machine learning models. They are aimed for learning and training procedures for the data used in regression analysis and classification tasks. An SVM is the representation of points or attribute values in the plane, along with that the non-linear hyperplanes for separation task in classification. Some parameters like gaussian kernels [53], standard deviation and variance of data, kernel functions are some significant parameters which affect the performance of SVM.

Fuzzy SVM

In FSVM, each training point belongs exactly to no more than one particular class. Some points having noise and that could not have classified by SVM, are dealt here through FSVM. Pre-knowledge [54] [55] information about data sets is needed, like - stochastic and probabilistic information. Here, several stochastic correlations can be identified.

Bayesian Classifiers

In these type of classifiers, the statistical information and probabilistic knowledge is employed for metadata creation. Here, Bayes' theorem [58] is utilized with naive independence assumptions among features. Since 1950's, it is being continuously explored. This is having applications in medical diagnosis analytics, spatial imaging data, text categorization etc. This classifier is highly scalable and it requires a number of parameters which are linear in no. of variable predictors in aspects of learning problem.

ROUGH SET THEORY

This section contains summarization of basic details of rough set theory which was originally proposed by Z. Pawlak. [56] In some way, initial theory of rough sets is referred to as "Pawlak or classical Rough Sets".

Suppose I is an information system(IS) which is equal to set (U; A) where A: finite set of attributes and U: represents non-empty finite set of objects such that:

$$a : U \rightarrow V_a$$

for every $a \in A$,where V_a represent set of value that attribute a may take. Each attribute a and object x in U will get a value a(x) from V_a using information table. An associated equivalence relation IND(P) with any $P \subseteq A$ is following:

$$IND(P) = \{(x, y) \in U^2 \mid \forall a \in P, a(x) = a(y)\}$$

where, IND(P): P-indistinguishability relation.

LITERATURE REVIEW

This section extensively represents the research work and developments that has taken place in past years. Junfei Qiu et. al. [1] has surveyed the recent advancements in machine learning for big data analysis. Philip Russom et. al. [2] has discussed about new techniques and tools which are used for analytics with the big data in past years. Dunren Che et. al. [3] presented overview about big data mining and its challenges. Joseph McKendrick et. al. [4] described the various challenges presented by the big data. Lidong Wang et. al. [5] introduced different machine learning methods and technologies in big data. Z. Pawlak et. al. [6] has described some basic concepts about which information systems are defined in entire knowledge discovery process.

Changwon Yoo et. al. [7] introduced the concept of regression analysis named as logistic and linear regressions. Alexandra

L'heureux et. al. [8] discussed that traditional machine learning which are developed in different categeory. Christopher C Drovandi et. al. [9] has described about design approach to big data analysis, whereas, main purpose of this is to analyse bigdata by opening the discourse on use of new experimental design methods. In divide and conquer strategies, by using efficient sub modules, it has capability to add value to the other big data sampling procedures. Yichuan Wang et. al. [10] has described about an extensive BDA-enabled transformation model. Here, it mainly discusses about how big data analytics capabilities transform the organizational practices, also generating potential benefits.

Farzaneh Farhangmehr et. al. [11] has discussed on development and evolvement of algorithmic procedures and methodologies for overcoming the certain challenges in the big data analytics. Yichuan Wang et. al. [12] described about exploring the way to big data analytics success, specially in healthcare domain, authors proposed BDA-enabled business value model, explained about how does big data analytics capabilities can be developed in the health care industries. Jacky Akoka et. al. [13] have found procedures combining big data, cloud, mobility or social media. U.Sivarajah et. al. [14] has described about critical analysis of Big Data challenges and analytical methods. They presented holistic view of BD application and practices. Hence, based on existing research studies, they have presented comprehensive structured analysis on the BD and BDA. Shuliang Xu et. al. [15] given an overview about dy-namic extreme learning machine for data stream classification. They have mainly described about proposed model which is dynamic and double hidden layers learning machine for the data stream classification. Xiaochuang Yao et. al. [16] described about spatial coding-based approach for partitioning big spatial data in Hadoop. They have proposed spatial coding based approach for separating the big spatial data in the Hadoop, which is initially compressed whole data based on the spatial coding matrix for creating SIS(sensing information set). Eric P. Xing et. al. [17] given strategies and design principles of parallel and distributed machine learning on Big Data. Wan-Yu Deng et. al. [18] proposed a fast SVD-hidden-nodes based extreme learning method for large-scaled data analytics. Chen Bo-Wei et. al. [19] described about signal processing using divide and conquer, attributes extraction and machine learning for big data.

General machine learning challenges with the big data [20] [21] [22] were described by the researchers. Whereas, others discussed them in the specific methodologies point of view [21] [23]. Gandomi and Haider have categorized hurdles with big data [24] [25] [26]. Vertical and horizontal scaling platforms in big data point of view were considered by Singh and Reddy [25]. Similarly, challenges of the data mining with big data have explored in [26] [27] [28].

Traditional strategies are struggling when faced with the massive data. Learning from these huge data is assumed to bring vital opportunities and the reframing potential for numerous sectors [29] [30] [31]. In [32] have defined big data aiming at the traits of generated data, which contains both the

amount and the structure of data. In [33] [34] presents a detailed review in the big data security. In [35] presents the description of analytic methods which are focusing on the big data. In [36] listed out the challenges and their possible solutions in industries and academics etc.

In [37][38] made a discussion of categorization and conceptual view of the big data in a cloud service model. By 2030, the quantity of the sensors will surely reach to approx one trillion, therefore the internet of things data is most significant part of big data [39] [40]. [41] presented several underlying procedures and methods which are employed to handle the data deluge, like - quantum computing, cloud computing and bio-inspired computing. [42] compared different big data models and explored some alternative ways for implementing big data sets. In [43] [44] focused mainly on the security aspects and compared nine big data systems on different security criteria's. In[45] focus on the visualization tools for big data and some significant commercial systems like "Tableau, Qlik View, Spotfire" etc. are compared based on specific measures. [46] surveyed the clustering algorithmic procedures and also evaluates their actual ability for dealing with velocity, volume, and variety aspects. [47] given description on big data analytics in both healthcare and government policies. [48] reports the case study providing qualitative information on big data specificities.

BIG DATA AND ANALYSIS

This section presents overall idea of big data sets and tools that are used for big data analysis.

BIG DATA: SCENARIO

Big data [8] is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy.

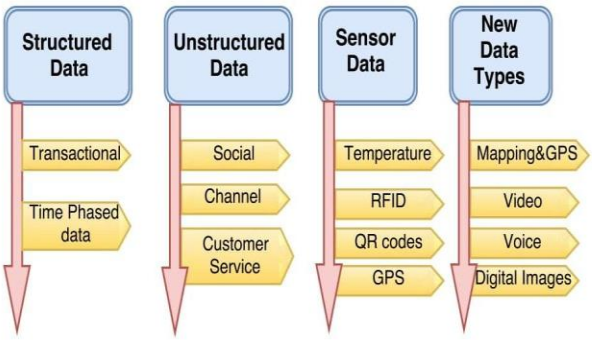

Figure.3: Big data types

Machine learning application to Big data

Machine learning [5] is ideal for exploiting the opportunities hidden in big data. It delivers on the promise of extracting value from big and disparate data sources with far less reliance on human direction. An overview of the application to big data is given in the figure 4:

It is data driven and runs at machine scale. It is well suited to the complexity of dealing with disparate data sources and the huge variety of variables and amounts of data involved. And unlike traditional analysis, machine learning thrives on growing datasets. The more data fed into a machine learning system, the more it can learn and apply the results to higher quality insights.

Big data management tools

The entire data analytics industry nowadays has a buzzword, "big data," concerning how we're operating something with the enormous amount of information gathering up. "Big data" is replacing "business intelligence". To handle this massive amount of data available, we have listed out some significant tools that can be utilized to process big data.

"Pentaho Business Analytics"

It is a kind of software program that started as an engine, branching within big data by creating it simpler to absorb the information from the different sources. One can experiment with Pentaho's tool to many of the most popular NoSQL databases, they are - MongoDB Cassandrav etc. One can drag furthermore drop the columns into aspects and reports as if the information issued from the SQL databases, once the databases are connected.

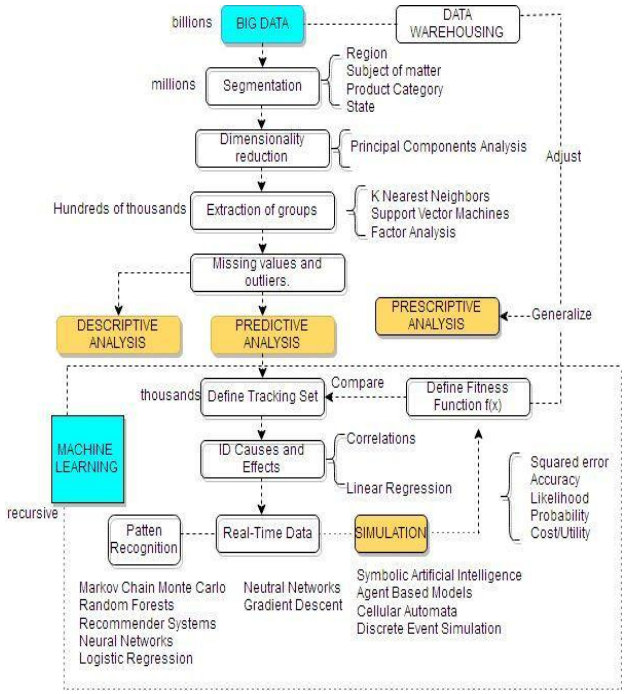

Figure.4: Machine learning applications to Big data

"Karmasphere Studio and Analyst"

It is kind of a specialized IDE that makes it simpler to create and run Hadoop jobs. This produces something better: As we set up the workflow, the tool engine displays the status of the test data at each and every step.

"Talend Open Studio"

This tool gives an Eclipse-based Integrated Development Environment for stringing data processing operations collectively with Hadoop. Its tools are intended to help with data integration, data quality along with the data management.

"Skytree Server"

Skytree allows a bundle that delivers many extra advanced ML procedures. All it needs is typewriting the right command in command line. It is more focused on the guts than the shiny GUI. Skytree Server is optimized to execute a no. of classical ML algorithms. It thought of as ten thousand time faster than different packages. It can explore through the data looking for clusters of similar objects, then rearrange this.

"Splunk"

It is a little distinctive from the other tools. It creates an index of the data as if the data were a part or a block of text. This approach is much alike to a text search method. Splunk will choose text strings and search around in the index. Its variant tool Shep guarantees bidirectional union of Hadoop and Splunk, enabling to interchange data within the systems and query Splunk data of Hadoop.

"Jaspersoft BI Suite"

It is one of the open source tool for mainly producing reports from database columns. The software tool is well-polished and already installed in many businesses turning SQL tables into PDFs that everyone can scrutinize at meetings. Jaspersoft is not specifically offering unique ways to look at the data, just more complicated ways to access and to locate data stored in the new locations.

Open Issues and Research Directions

Today, to process the huge sized unstructured, inconsistent, incomplete and vague data by computing machines is a challenging task. To perform operations in the data [2] [4], present in higher dimensions may be more computationally complex procedure as well as the computational overhead is huge in further training and testing phases of classification. In recent past years, Rough set theory and Fuzzy logic evolved as an efficient machine learning methodology, which has grown as an important tool to perform big data analytics. The problem identification and further future research directions are presented as follows:

Problem Identification

With the beginning of span of Big Data, which may be considered as the next bound for modernization, competition and potency, a new boom of revolution is nearly about to onset. The volume of data today, is raging at an unusual rate [12] [39] as a result of advancements and developments in Web technologies, social media, and mobile devices etc. Traditional strategies are hardly suffering when faced with this massive sized data. These traditional machine learning(ML) routines and procedures are not inherently practical or scalable enough to manage the data with the properties of massive volume [41], varying types, great speed, uncertainty and incompleteness.

Based on the precious knowledge, we need to create new techniques and methods to excavate big data. Machine Learning demands to deeply discover itself for processing big data, so that the knowledge extraction and reasoning for uncertain concepts from unstructured and huge sized data can be done in a computationally efficient manner.

FUTURE RESEARCH DIRECTIONS

The aim of our research is to develop new efficient methods for the analysis of big data sets. Our future research directions are as follows: -

We will contribute some optimal and computationally efficient big data analytics techniques to analyze different type of data sets. This may be achieved by selecting strategies of Rough set theory and Fuzzy logic evolved as an efficient machine learning methodology, which has become an important tool to perform data analytics.

As, today, processing of massive sized unstructured, inconsistent, incomplete and imprecise data by computing machines is a challenging task. In recent past years, Rough set theory and Fuzzy logic evolved as an efficient machine learning methodology, which has become an important tool to perform big data analytics. To perform operations in the data, present in higher dimensions may be more computationally complex procedure as well as the computational overhead is huge in further training and testing phases of classification. We will employ these modern machine learning techniques to process big data, which also gives the guarantee of

dimensionality reduction and other parameters selection of data sets.

We will experiment our developed methods on standard datasets such as UCI ML-Repository, CORA, Reuters etc. and compare the analysis results with existing techniques.

CONCLUSION

Big data analytics is the process of examining large and varied data sets. Learning from massively large and unstructured data brings significant opportunities for numerous sectors. Still, most of these routines are not much computationally efficient, practical or scalable enough. This paper discusses the need for the research that aims at proposing new techniques that can be used for analysis of big data. However, most of the traditional AI involved methods are not scalable to manage data with the properties of its huge volume, diverse types, inconsistency, uncertainty along with incompleteness. In response, there is a need for machine learning to revitalize itself for big data processing.

This paper started with various types of learning methods. Further it discusses about some of the significant and practical issues of machine learning for big data analytics. Then an extensive survey of related work and methods which have been developed in past, is presented. Later, we have listed out some tools which can be employed for big data management and analysis. To encourage more interests for the readers of the paper, in the end, some open issues in big data domain, problem identification and our future research goals were presented.

REFERENCES

[1] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu and Shuo Feng. ”A survey of machine learning for big data processing”, EURASIP Journal on Advances in Signal Processing (2016) 2016:67, Springer.

[2] Philip Russom. ”Big data analytics”, TDWI research, Fourth quarter (2011).

[3] Dunren Che, Mejdl Safran, Zhiyong Peng. ”From Big Data to Big Data Mining: Challenges, Issues, and Opportunities”, DASFAA Workshops, LNCS 7827, pp. 1-15, Springer-Verlag Berlin Heidelberg (2013).

[4] Joseph McKendrick. ”Big Data, Big Challenges, Big Opportunities: IOUG Big Data Strategies Survey”, Unisphere Research, ORACLE, September (2012).

[5] Lidong Wang, Cheryl Ann Alexander. ”Machine Learning in Big Data”, International Journal of Mathematical, Engineering and Management Sciences, Vol. 1, No. 2, 5261, (2016).

[6] Z. Pawlak. ”Information Systems Theoretical Foundations”, Information Systems, Vol. 6, No. 3, pp. 205-218, (1981).

[7] Changwon Yoo, Luis Ramirez, Juan Liuzzi. ”Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine”, Int. Neurology Journal (2014); 18:50-57.

[8] Alexandra L’heureux, Katarina Grolinger, Hany F. Elyamany and Miriam A. M. Capretz. ”Machine Learning With Big Data: Challenges and Approaches”, IEEE ACCESS, Vol. 5, June 7, (2017).

[9] Christopher C Drovandi, Christopher Holmes, James M McGree, Kerrie Mengersen, Sylvia Richardson and Elizabeth G Ryan. ”A Principled Experimental Design Approach to Big Data Analysis”, QUT ePrints (2017).

[10] Yichuan Wang, LeeAnn Kung, William Yu Chung Wang, Casey G. Cegielski. ”An integrated big data analytics-enabled transformation model: Application to health care”, Information and Management, April (2017).

[11] Farzaneh Farhangmehr. ”Statistical Approaches for Big Data Analyt-ics and Machine Learning: Data-Driven Network Reconstruction and Predictive Modeling of Time Series Biological Systems”, escholarship, University of California, (2014).

[12] Yichuan Wang, Nick Hajli. ” Exploring the path to big data analytics success in healthcare”, Journal of Business Research 70 (2017) 287-299.

[13] Jacky Akoka, Isabelle Comyn-Wattiau, Nabil Laoufi. ” Research on Big Data - A systematic mapping study”, Computer Standards & Interfaces 54 (2017) 105-115.

[14] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vis-hanth Weerakkody. ” Critical analysis of Big Data challenges and ana-lytical methods”, Journal of Business Research 70 (2017) 263-286.

[15] Shuliang Xu, Junhong Wang. ”Dynamic extreme learning machine for data stream classification”, Neurocomputing 238 (2017) 433-449.

[16] Xiaochuang Yao, Mohamed F. Mokbel, Louai Alarabi, Ahmed Eldawy, Jianyu Yang, Wenju Yun, Lin Li, Sijing Ye, Dehai Zhu. ”Spatial coding-based approach for partitioning big spatial data in Hadoop”, Computers & Geosciences 106 (2017) 60-67.

[17] Eric P. Xing, Qirong Ho, Pengtao Xie, Dai Wei. ”Strategies and Principles of Distributed Machine Learning on Big Data”, Engineering 2 (2016) 179-195.

[18] Wan-Yu Deng, Zuo Bai, Guang-Bin Huang, Qing-Hua Zheng. ”A Fast SVD-Hidden-nodes based Extreme Learning Machine for Large-Scale Data Analytics”, Neural Networks 77 (2016) 14-28.

[19] Chen Bo-Wei, Wen Ji, Seungmin Rho. ”Divide-and-conquer signal processing, feature extraction, and machine learning for big data”, Neurocomputing 174 (2016) 383-383.

[20] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha. ”Efficient machine learning for big data: A review”, Big Data Res., vol. 2, no. 3, pp. 87-93, Sep. (2015).

[21] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic. ”Deep Learning Applications and Chal-lenges in Big Data Analytics”, J. Big Data, vol. 2, no. 1, p. 1, Feb. (2015).

[22] S. R. Sukumar. ”Machine learning in the big data era: Are we there yet? in Proc. 20th ACM SIGKDD Conf. Knowl. Discovery Data Mining, Workshop Data Sci. Social Good (KDD), (2014), pp. 1-5.

[23] X.-W. Chen and X. Lin. ”Big data deep learning: Challenges and perspectives”, IEEE Access, vol. 2, pp. 514-525, (2014).

[24] A. Gandomi and M. Haider. ”Beyond the hype: Big data concepts, methods, and analytics”, Int. J. Inf. Manage., vol. 35, no. 2, pp. 137-144, Apr. (2015).

[25] D. Singh and C. K. Reddy. ”A survey on platforms for big data analytics”, J. Big Data, vol. 2, no. 1, pp. 1-20, (2015).

[26] P. D. C. de Almeida and J. Bernardino. ”Big data open source platforms”, in Proc. IEEE Int. Congr. Big Data, Jun. (2015), pp. 268-275.

[27] W. Fan and A. Bifet. ” Mining big data: Current status, and forecast to the future”, SIGKDD Explorations Newslett., vol. 14, no. 2, pp. 1-5, Dec. (2012).

[28] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding. ”Data mining with big data”, IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97-107, Jan. (2014).

[29] O.R. Team, Big Data Now: Current Perspectives from O’Reilly Radar Sebastopol, USA: O’Reilly Media, CA, (2011).

[30] M. Grobelnik, Big Data Tutorial, http://videolectures.net/eswc2012 grobelnik big data/.

[31] S. Sagiroglu, D. Sinanc. ”Big Data: a review”, in: IEEE Int. Conf. on CTS, (2013).

[32] A. Cuzzocrea, I.Y. Song, K. Davis. ”Analytics over large-scale multi-dimensional data: the big data revolution!”, in: Proceedings of the 14th international workshop on Data Warehousing and OLAP. New York, New York, USA: ACM, (2011), pp. 101-103.

[33] A. Jacobs. ”The pathologies of Big Data”, Commun. ACM 52 (8) (2009) 36.

[34] K. Kambatla, G. Kollias, V. Kumar, A. Grama. ”Trends in Big Data analytics”, J. Parallel Distrib. Comput. 74 (2014) 2561-2573.

[35] A. Gandomi, M. Haider. ”Beyond the hype: big Data concepts, methods and analytics”, Int. J. Inf. Manag. 35 (2015) 137-144.

[36] H. Fang, Z. Zhang, C.J. Wang, M. Daneshmand. ”A survey of big data research”, IEEE Netw. 29 (5) (2015) 6-9.

[37] M.D. Assunao, R.N. Calheiros, S. Bianchi, M.A.S. Netto, R. Buyya. ”Big Data computing and clouds: trends and future directions”, J. Parallel Distrib. Comput. 79-80 (2015) 3-15.

[38] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani. ”The rise of ”big data” on cloud computing: review and open research issues”, Inf. Syst. 47 (2015) 98-115 (Elsevier).

[39] M. Chen, S. Mao, Y. Zhang, V.C.M. Leung. ”Big Data: Related Tech-nologies Challenges and Future Prospects”, Springer, Cham, Heidelberg, New York, Dordrecht, London, (2014).

[40] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han et al.” Challenges and Opportunities with Big Data. A Community White Paper Developed by Researches Across the United States”.

[41] C.L. Philip Chen, C.Y. Zhang. ”Data-intensive applications, challenges, techniques and technologies: a survey on Big Data”, Inf. Sci. 275 (2014) 314-347.

[42] S. Sharma, U.S. Tim, J. Wong, S. Gadia, S. Sharma. ”A brief review on leading big data models”, Data Science Journal-jlc.jst.go.jp, (2014).

[43] D.S. Terzi, R. Terzi, S. Sagiroglu. ”A survey on security and privacy issues in big data”, in: Proceedings of the 10th International Conference for Internet Technology and Secured Transactions (ICITST), IEEE, (2015), pp. 202-207.

[44] E. Sahafizadeh, M.A. Nematbakhsh. ” A survey on security issues in Big Data and NoSQL”, ACSIJ Adv. Comput. Sci.: Int. J. 4 (4) (2015) No.16.

[45] L. Zhang, A. Stoffel, M. Behrisch, et al. ”Visual analytics for the big data era-a comparative review of state-of-the-art commercial systems”, in: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on. IEEE, (2012), pp. 173-182.

[46] A. Fahad, N. Alshatri, Z. Tari, A. Alamri. ”A survey of clustering algorithms for big data: taxonomy and empirical analysis”, Emerg. Top. Comput. IEEE Trans. 2 (3) (2014) 267-279.

[47] J. Archenaa, J. Mary Anita. ”A survey of big data analytics in healthcare and government”, Procedia Comput. Sci. 50 (2015) 408-413 (ISSN 1877-0509).

[48] S.F. Wamba, S. Akter, A. Edwards, G. Chopin. ”How ’big data’ can make big impact: findings from a systematic review and a longitudinal case study”, Int. J. Prod. Econ. 165 (2015) 234-246.

[49] Gui J, Andrew AS, Andrews P, Nelson HM, Kelsey KT, Karagas MR et al. ”A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility”, Ann Hum Genet, Wiley (2011); 75:20-8.

[50] R. Krikorian. (2010). ”Twitter by the Numbers, Twitter”. [On-line]. Available: http://www.slideshare.net/raffikrikorian/twitter-by-the-numbers?ref=http://techcrunch.com 2010/09/17/twitter-seeing-6-billion-api-calls-per-day-70k-per-second/

[51] ABI. (2013). ” Billion Devices Will Wirelessly Connect to the Internet of Everything in 2020, ABI Research.” [Online].

[52] Available: https://www.abiresearch.com/press/more-than-30 billion-devices-will-wirelessly-conn/

[53] W. Raghupathi and V. Raghupathi. ”Big data analytics in healthcare: Promise and potential”, Health Inf. Sci. Syst., vol. 2, no. 1, pp. 1-10, (2014).

[54] Cortes, C., Vapnik, V. (1995). ” Support-vector networks”. Machine Learning, 20(3), 273-297.

[55] Abonyi, J., Szeifert, F. (2003). ” Supervised fuzzy clustering for the identification of fuzzy classifiers”. Pattern Recognition Letters, 24(14), 2195-2207.

[56] C.-F. Lin and S.-D. Wang,” Fuzzy support vector machines”, IEEE Transactions on Neural Networks, vol. 13, no. 2, March 2002.

[57] Pawlak Z.(1982). ”Roughsets”. International Journal of Computer and information Sciences, 11, 341-356.

[58] S Agatonovic-Kustrin, R Beresford. ”Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical re-search”, Journal of Pharmaceutical and Biomedical Analysis, Volume 22, Issue 5, June (2000), Pages 717-727.

[59] Nir Friedman, Ron Kohavi.” Bayesian Classification”, Stanford Artificial Intelligence Laboratory, (1999) robotics.stanford.edu/ ronnyk bayesHB.pdf.

[60] Y Bengio, A Courville, P Vincent. ” Representation learning: a review and new perspectives”. IEEE Trans Pattern Anal 35(8), 1798-1828 (2012)

[61] 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.

[62] S Li, C Huang, C Zong. ”Multi-domain sentiment classification with classifier combination”. J Comput Sci Technol 26(1), 25-33 (2011).

[63] 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.

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

[65] 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).

[66] D Peteiro-Barral, B Guijarro-Berdias. ” A survey of methods for dis-tributed machine learning”. Progress in Artificial Intelligence 2(1), 1-11 (2012).

[67] EW Xiang, B Cao, DH Hu, Q Yang. ” Bridging domains using worldwide knowledge for transfer learning”. IEEE Trans Knowl Data Eng 22(6), 770783 (2010)

[68] 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).

[69] 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).

[70] 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).

[71] 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.