International Journal of Advances in Applied Sciences (IJAAS)
Vol. 4, No. 4, December 2015, pp. 117~123
ISSN: 2252-8814

Machine Learning in Big Data

Lidong Wang*, Guanghui Wang, Cheryl Ann Alexander*****
* Department of Engineering Technology, Mississippi Valley State University, USA
** State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, China
*** Technology and Healthcare Solutions, USA

Article Info

Article history:

Received Sep 28, 2015
Revised Nov 7, 2015
Accepted Nov 18, 2015

Keyword:

Big data
Big data analytics
Information technology
Machine learning
Networks
Stream processing

ABSTRACT

Machine learning is an artificial intelligence method of discovering knowledge for making intelligent decisions. Big Data has great impacts on scientific discoveries and value creation. This paper introduces methods in machine learning, main technologies in Big Data, and some applications of machine learning in Big Data. Challenges of machine learning applications in Big Data are discussed. Some new methods and technology progress of machine learning in Big Data are also presented.

Corresponding Author:

Lidong Wang,
Department of Engineering Technology,
Mississippi Valley State University,
USA.
Email: lwang22@students.tntech.edu

1. INTRODUCTION

Machine learning is an important area of artificial intelligence. The objective of machine learning is to discover knowledge and make intelligent decisions. Machine learning algorithms can be categorized into supervised, unsupervised, and semi-supervised. When big data is concerned, it is necessary to scale up machine learning algorithms [1], [2]. Another categorization of machine learning according to the output of a machine learning system includes classification, regression, clustering, and density estimation, etc. Machine learning approaches include decision tree learning, association rule learning, artificial neural networks, support vector machines (SVM), clustering, Bayesian networks, and genetic algorithms, etc., [3].

Examples of supervised learning algorithms include Naïve Bayes, boosting algorithm, support vector machines (SVM), and maximum entropy method (MaxENT), etc. Unsupervised learning takes unlabelled data and classifies by comparing the features of data. Examples of unsupervised algorithms are clustering (*k*-means, density-based, and hierarchical, etc.), self-organizing maps (SOM), and adaptive resonance theory (ART) [4].

Machine learning has been used in big data. Big data is a massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques. Big data technologies have great impacts on scientific discoveries and value creation [5]-[7]. Massive parallel-processing (MPP), distributed file systems, and cloud computing, etc. are supporting technologies of Big Data [8]. Besides general cloud infrastructure services, technologies such as Hadoop, Databases/Servers SQL, NoSQL, and MPP databases, etc. are also used to support Big Data [9].

This paper introduces machine learning, its applications in Big Data, and the challenges and technology progress of machine learning in Big Data. The organization of this paper is as follows: the next section introduces methods of machine learning and big data; Section 3 introduces machine learning

applications in big data; Section 4 discusses challenges of machine learning applications in big data; Section 5 presents technology progress of machine learning applications in big data; and the final section is conclusions.

2. METHODS OF MACHINE LEARNING AND BIG DATA

Supervised learning can be divided into classification and regression. When the class attribute is discrete, it is called classification; when the class attribute is continuous, it is regression. Decision tree learning, naive Bayes classifier, k-nearest neighbor (kNN) classifier, and classification with network information are classification methods. Linear regression and logistic regression are regression methods. Unsupervised learning is the unsupervised division of instances into groups of similar objects [10].

Clustering can be grouped into three categories. They are supervised, unsupervised, and semi-supervised [11]:

1. Supervised clustering: It identifies clusters that have high probability densities with respect to individual classes (class-uniform clusters). It is used when there is a target variable and a training set that includes the variables to cluster.
2. Unsupervised clustering: It maximizes the intracluster similarity and minimizes the intercluster similarity when a similarity/dissimilarity measure is given. It uses a specific objective function (e.g., a function that minimizes the intraclass distances to find tight clusters). *K*-means and hierarchical clustering are the most widely used unsupervised clustering techniques in segmentation.
3. Semi-supervised clustering: In addition to the similarity measure, semi-supervised clustering utilizes other guiding/adjusting domain information to improve the clustering results. This domain information can be pairwise constraints between the observations or target variables for some of the observations.

Decision trees classify examples based on their feature values. Decision trees are constructed recursively from training data using a top-down greedy approach in which features are sequentially selected [10]. Decision tree classifiers organize the training data into a tree-structure plan. Decision trees are constructed by stating with the root node having the whole data set, iteratively choosing splitting criteria and expanding leaf nodes with partitioned data subsets according to the splitting criteria. Splitting criteria are chosen based on some quality measures such as information gain, which requires handling the entire data set of each expanding nodes. This makes it difficult for decision trees to be applied to big data applications [12].

Support vector machine (SVM) is a binary classifier which finds linear classifier in higher dimensional feature space to which original data space is mapped. SVM shows very good performance for data sets in a moderate size. It has inherent limitations to big data applications [12].

Deep machine learning has become a research frontier in artificial intelligence. It is a machine learning technique, where many layers of information processing stages are exploited in hierarchical architectures. It computes hierarchical features or representations of the observational data, where the higher-level features or factors are defined from lower-level ones. Deep learning algorithms extract high-level, complex abstractions as data representations through a hierarchical learning process. While deep learning can be applied to learn from labeled data, it is primarily attractive for learning from large amounts of unlabeled/unsupervised data, making it attractive for extracting meaningful representations and patterns from big data. Deep learning algorithms and architectures are more aptly suited to address issues related to Volume and Variety of Big data analytics. Deep machine learning can be applied to big data. However, it has some restrictions in big data applications because it requires significant amount of training time [12], [13].

Parallel learner for assembling numerous ensemble trees (PLANET) is a regression tree algorithm implemented with a sequence of MapReduce jobs that run on the big data framework, Hadoop. It can deal with big volume of data, but is not applicable to data with categorical attributes [12].

One trend in machine learning is to combine results of multiple learners to obtain better accuracy. This trend is commonly known as Ensemble Learning. There are four methods of combining multiple models: bagging, boosting, stacking, and error-correcting output [14].

A comparison of several machine learning algorithms was made in Table 1 [15] according to algorithms type, algorithms trait, learning policy, learning algorithms, and classification strategy. Some features of machine learning algorithms were compared in Table 2 [16].

Table 1. Summary of several machine learning algorithms

Algorithms	Algorithms type	Algorithms characteristic	Learning policy	Learning algorithms	Classification strategy
Decision tree	Discriminant	Classification tree	Regularized maximum likelihood estimation	Feature selection, generation, prune	IF-THEN policy based on tree spitting
Non-linear support vector machine (based on libsvm)	Discriminant	Super-plane separation, kernel trick	Minimizing the loss of regular hinge, soft margin maximization	Sequential minimal optimization algorithm (SMO)	Maximum class of test samples
Linear SVM (based on liblinear)	Discriminant	Super-plane separation	Minimizing the loss of regular hinge, soft margin maximization	Sequential dual method	Maximum weighted test sample
Stochastic gradient boosting	Discriminant	Linear combination of weak classifier (based on decision tree)	Addition minimization loss	Stochastic gradient descent algorithm	Linear combination of weighted maximum weak classifiers
Naive Bayesian classifier	Generative	Joint distribution of feature and class, conditional independent assumption	Maximum likelihood estimation, Maximum posterior probability	Probabilistic computation	Maximum posterior probability

Table 2. Comparing machine learning algorithms

	Decision Trees	**Neural Networks**	**Naïve Bayes**	**kNN**	**SVM**	**Rule-learners**
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to number of attributes and the number of instances	***	*	****	****	*	**
Speed of classification	****	****	****	*	****	****
Tolerance to missing values	***	*	****	*	**	**
Tolerance to irrelevant attributes	***	*	**	**	****	**
Tolerance to redundant attributes	**	**	*	**	****	**
Tolerance to highly interdependent attributes (e.g. parity problems)	**	***	*	*	***	**
Dealing with discrete/binary/continuous attributes	****	*** (not discrete)	*** (not continuous)	*** (not directly discrete)	** (not discrete)	*** (not directly continuous)
Tolerance to noise	**	**	***	*	**	*
Dealing with danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*	****
Model parameter handling	***	*	****	***	*	***

(**** stars represent the best and * star the worst performance)

There are several frameworks, like Map/Reduce, DryadLINQ, and IBM parallel machine learning toolbox that have capabilities to scale up machine learning [1].

Mahout is an open source machine learning library from Apache for big data analysis. It aims to be the machine learning tool of choice when the collection of data is very large [4]. The Apache Mahout project aims at building a scalable machine learning library on top of Hadoop. The Mahout machine learning library was integrated, adapted, and extended by developing advanced machine learning algorithms for large scale data. Mahout may significantly help towards grouping similar items, identifying main or "hot" topics, assigning items to predefined categories, recommending important data to diverse stakeholders, and discovering frequent and meaningful patterns in a specific decision-making setting [17].

PivotalR is a package for machine learning on big data. PivotalR utilizes the full power of parallel computation and distributive storage, and thus gives the normal R user access to big data stored in distributive databases or Hadoop distributive file system (HDFS). It provides data-parallel implementations

of mathematical, statistical and machine-learning algorithms for structured and unstructured data. Thus PivotalR also enables the user to apply machine learning algorithms on big data [18].

There are a lot of technologies supporting Big Data analytics and applications. Table 3 [11] compares a number of big data technologies. The table highlights the different types of systems and their comparative strengths and weaknesses.

Table 3. Comparison of Big Data Technologies

	In-memory database	MPP database	Big Data appliance	Hadoop	NoSQL database
Consistent	W	W	W	P	P
Available	W	W	W	P	P
Fault tolerant	W	W	P	W	W
Suitable for real-time transactions	W	W	W	F	F
Suitable for analytics	P	P	W	W	F
Suitable for extremely big data	F	P	P	W	W
Suitable for unstructured data	F	F	P	W	W

W: Meets widely held expectations.
P: Potentially meets widely held expectations.
F: Fails to meet widely held expectations.

3. EXAMPLES OF MACHINE LEARNING APPLICATIONS IN BIG DATA

The combination of supervised and unsupervised machine learning techniques for efficiently analyzing a big volume of crime data was proposed. The combination includes three steps: dimensionality reduction, clustering, and classification. *R* statistical software was used because it is a powerful tool to deal with big data. The specific work is outlined as follows [19]:

1. Measure correlation between crime and social attributes. This method reduces dimensionality of the crime data.
2. Use unsupervised machine learning technique to divide crime data into groups; use *k*-means clustering algorithm to cluster the crime data into dangerous, average, and safe regions.
3. Use supervised machine learning technique to predict whether a particular region is dangerous or safe; use decision tree classification algorithm to perform predictions.

Analysis and mining of social network data for society issues was conducted using Big Data. Social data mining is the process of analyzing, representing as well as extracting actionable patterns from social network data. Machine learning and stemming algorithms were used to classify the tweets. Tweets are often in the pattern of big data. The predicting features from tweets were extracted from a collection of tweets; stopping words were removed; and all keywords were selected. As tweets are very short and may contain incomplete sentences, the meaning of the tweets may be ambiguous. In machine learning, support vector machines (SVM) are supervised models with related learning algorithms that analyze all the data which are used for classification of the tweets. Stemming algorithm uses a pre-processing task in text mining and can be used as a common requirement of natural language processing functions. Stemming algorithm was used to extract the main keywords or root words from the tweets. The stemming algorithm can be applied to predict the keywords from the tweets. All the keywords are classified by the SVM algorithm [20].

4. CHALLENGES OF MACHINE LEARNING APPLICATIONS IN BIG DATA

General challenges about machine learning are: (1) designing scalable and flexible computational architectures for machine learning; (2) the ability to understand characteristics of data before applying machine learning algorithms and tools; and (3) the ability to construct, learn and infer with increasing sample size, dimensionality, and categories of labels [21].

There are many scale machine learning algorithms, but many important specific sub-fields in large-scale machine learning, such as large-scale recommender systems, natural language processing, association rule learning, ensemble learning, still face the scalability problems [1].

The basic MapReduce framework commonly provided by first-generation "Big Data analytics" platforms like Hadoop lacks an essential feature for machine learning (ML). MapReduce does not support iteration /recursion or certain key features required to efficiently iterate "around" a MapReduce program. Programmers building ML models on such systems are forced to implement looping in ad-hoc ways outside the core MapReduce framework. This makes their programming task much harder, and it often also yields inefficient programs in the end. This lack of support has motivated the recent development of various

specialized approaches or libraries to support iterative programming on large clusters. Meanwhile, recent MapReduce extensions such as HaLoop, Twister, and PrItr aim at directly addressing the iteration outage in MapReduce; they do so at the physical level [22].

Major problems that make the machine learning (ML) techniques unsuitable for solving big data classification problems are: (1) An ML technique that is trained on a particular labeled datasets or data domain may not be suitable for another dataset or data domain – that the classification may not be robust over different datasets or data domains; (2) an ML technique is in general trained using a certain number of class types and hence a large varieties of class types found in a dynamically growing dataset will lead to inaccurate classification results; and (3) an ML technique is developed based on a single learning task, and thus they are not suitable for today's multiple learning tasks and knowledge transfer requirements of Big data analytics [23].

Traditional algorithms in machine learning generally do not scale to big data. The main difficulty lies with their memory constraint. Although algorithms typically assume that training data samples exist in main memory, big data does not fit into it. A common approach to learning from a large dataset is data distribution. By replacing batch training on the original training dataset with separated computations on the distributed subsets, one can train an alternative prediction model at a sacrifice of accuracy. Another way is to use online learning, in which memory usage does not depend on the size of the dataset. Neither online learning nor distributed learning is sufficient for learning from big data streams. There are two reasons. First is that the data size is too large to be relaxed by either online or distributed learning. Sequential online learning on big data requires too much time for training on a single machine. On the other hand, distributed learning with a large number of machines reduces the gained efficiency per machine and affects the overall performance. The second reason is that combining real-time training and prediction has not been studied. Since big data is typically used after being stored in (distributed) storage, the learning process also tends to work in a batch manner [24].

Scaling up big data to proper dimensionality is a challenge that can encounter in machine learning algorithms; and there are challenges of dealing with velocity, volume and many more for all types of machine learning algorithms. Since big data processing requires decomposition, parallelism, modularity and/or recurrence, inflexible black-box type machine learning models failed in an outset [2].

Applying the distributed data-parallelism (DDP) patterns in Big Data Bayesian Network (BN) learning faces several challenges: (1) effectively pre-processing big data to evaluate its quality and reduce the size if necessary; (2) designing a workflow capable of taking Gigabytes of big data sets and learning BNs with decent accuracy; (3) providing easy scalability support to BN learning algorithms [14].

Deep learning challenges in big data analytics lie in: incremental learning for non-stationary data, high-dimensional data, and large-scale models [13].

Because high-level data parallel frameworks, like MapReduce do not naturally or efficiently support many important data mining and machine learning algorithms and can lead to inefficient
learning systems, the GraphLab abstraction was introduced. It naturally expresses asynchronous, dynamic, graph-parallel computation while ensuring data consistency and achieving a high degree of parallel performance in the shared-memory setting [25].

5. TECHNOLOGY PROGRESS OF MACHINE LEARNING APPLICATIONS IN BIG DATA

Most of the advances for scalable machine learning (e.g. Madlib, Apache Mahout, etc.) are happening in the massively parallel database processing community. Better work can be done in the Big Data era by designing and implementing machine learning algorithms with scale-friendly predictive functions. The following methods have been exploring and evaluating [21]: (1) deep learning algorithms that automate the feature engineering process by learning to create and sift through data-driven features, (2) incremental learning algorithms in associative memory architectures that can seamlessly adapt to future data samples and sources, (3) faceted learning that can learn hierarchical structure in the data, and (4) multi-task learning that can learn several predictive functions in parallel.

The Big Data classification requires multi-domain, representation-learning (MDRL) technique because of its large and growing data domain. The MDRL technique includes feature variable learning, feature extraction learning, and distance-metric learning. Several representation-learning techniques have been proposed in machine learning. The cross-domain, representation-learning (CDRL) technique is maybe suitable for the Big Data classification along with the suggested network model [23].

A key benefit of deep learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data analytics. How deep learning can be utilized in Big Data analytics was explored; this includes extracting complex patterns from massive volumes of data, semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks. Some further research of deep

learning in Big Data was also investigated; this includes streaming data, high-dimensional data, scalability of Deep Learning models, and distributed computing [13].

As an important machine learning technique, Bayesian Network (BN) has been widely used to model probabilistic relationships among variables. An intelligent Big Data pre-processing approach and a data quality score were proposed to measure and ensure the data quality and data faithfulness; a new weight based ensemble algorithm was proposed to learn a BN structure from an ensemble of local results. For easily integrating the algorithm with distributed data-parallelism (DDP) engines, such as Hadoop, Kepler scientific workflow was employed to build the whole learning process. How Kepler can facilitate building and running the Big Data BN learning application was also demonstrated. A Scalable Bayesian Network Learning (SBNL) workflow was designed through combining machine learning, distributed computing, and workflow techniques. The workflow includes intelligent Big Data pre-processing and effective BN learning from Big Data by leveraging ensemble learning and distributed computing model [14].

For stream processing, one must process new data in real-time and in many times, considers the historical data as well to generate a value. Most often, stream processing involves the use of previously trained models to avoid too much processing and ultimately reduce response times. A novel architecture for performing machine learning over big data streams was proposed. The architecture provides reliable persistent storage of data over the Hadoop Distributed File System (HDFS) and HBase. The core of the architecture is comprised of the batch- and stream-processing modules. It provides machine learning tools and algorithms so that developers can easily take advantage of them to carry out tasks such as prediction, clustering, recommendation, and classification, etc., [26].

A distributed streaming algorithm was proposed to learn decision rules for regression tasks. The algorithm is available in Scalable Advanced Massive Online Analysis (SAMOA), an open-source platform for mining big data streams. It uses a hybrid of vertical and horizontal parallelism to distribute Adaptive Model Rules (AMRules) on a cluster. The decision rules built by AMRules are comprehensible models. SAMOA is a framework that eases the development of new distributed machine learning algorithms and the deployment of these implementations on top of state-of the-art distributed stream processing engines (DSPEs). It is also a library of distributed data mining and machine learning algorithms that allows users to use or customize existing ones [27].

Feature selection (FS) is an important topic in machine learning and data mining. The objective of feature selection is to select a subset of relevant features for building effective prediction models. Various FS methods have been proposed. Based on the selection criterion choice, these methods can be roughly divided into three categories: filter methods, wrapper methods, and embedded methods approaches. Filter methods relies on the characteristics of the data such as correlation, distance and information, without involving any learning algorithm. Wrapper methods require one predetermined learning algorithm for evaluating the performance of selected features set. Embedded methods aim to integrate the feature selection process into the model training process; they are faster than the wrapper methods; and still provide suitable feature subset for the learning algorithm. Online feature selection (OFS) for mining big data was studied to solve the feature selection problem by an online learning approach. The goal of online feature selection was to develop online classifiers that involve only a small and fixed number of features. Results show the proposed algorithms are fairly effective for feature selection tasks of online applications, and significantly more efficient and scalable than some state-of-the-art batch feature selection technique [28].

6. CONCLUSION

Splitting criteria of decision trees are chosen based on some quality measures, which requires handling the entire data set of each expanding nodes. This makes it difficult for decision trees to be used in big data applications. SVM shows very good performance to data sets in a moderate size. It has inherent limitations to big data applications. Deep learning is suited to address issues related to volume and variety of big data. However, it has some restrictions in big data because it requires much training time. PLANET can deal with big volume of data, but is not applicable to data with categorical attributes.

Machine learning applications in big data has met challenges such as memory constraint, no support (in iterations) from MapReduce, difficulty in dealing with big data due to Vs (such as high velocity, volume, and variety, etc.), and learning training limited to a certain number of class types or a particular labeled datasets, etc.

Some technology progress has been made such as faceted learning for hierarchical data structure, multi-task learning in in parallel, multi-domain/ cross-domain representation-learning, streaming data processing, high-dimensional data processing, and online feature selection, etc. These areas and the above challenges about machine learning in big data also can be further research topics.

REFERENCES

[1] C. L. P. Chen, C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", *Information Sciences*, Vol. 275, No. 10, pp. 314-347, August 2014.

[2] K. M. Tarwani, S. S. Saudagar, H. D. Misalkar, "Machine Learning in Big Data Analytics: An Overview", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 5, No. 4, pp. 270-274, April 2015.

[3] https://en.wikipedia.org/wiki/Machine_learning

[4] U. Jaswant and P.N. Kumar, "Big Data Analytics: A Supervised Approach for Sentiment Classification Using Mahout: An Illustration", *International Journal of Applied Engineering Research*, Vol. 10, No. 5, pp. 13447-13457, 2015.

[5] Y. Demchenko, P. Grosso, C. De Laat, P. Membrey, "*Addressing Big Data Issues in Scientific Data Infrastructure*", 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, pp. 48-55, May 2013.

[6] D. E. O'Leary, "'Big Data', the 'Internet of Things' and the 'Internet of Signs'", *Intelligent Systems in Accounting, Finance and Management*, Vol. 20, pp. 53-65, 2013.

[7] H. V. Jagadish, A. Labrinidis, Y. Papakonstantinou, *et al*., "Big Data and Its Technical Challenges", *Communications of the ACM*, Vol. 57, No. 7, pp. 86-94, 2014.

[8] A. Zaslavsky, C. Perera and D. Georgakopoulos, "*Sensing as a Service and Big Data*", International Conference on Advances in Cloud Computing (ACC), Bangalore, India, pp. 1-8, July 2012.

[9] M. Turk, "A chart of the big data ecosystem", take 2, 2012.

[10] R. Zafarani, M. A. Abbasi, H. Liu. "Social Media Mining: An Introduction", Cambridge University Press, UK, 2014.

[11] J. Dean, "Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners", John Wiley & Sons, Inc., 2014.

[12] K. M. Lee, "Grid-based Single Pass Classification for Mixed Big Data", *International Journal of Applied Engineering Research*, Vol. 9, No. 21, pp. 8737-8746, 2014.

[13] M. M. Najafabadi, F. Villanustre, T. M Khoshgoftaar, N. Seliya, R. Wald and E. Muharemagic, "Deep learning applications and challenges in big data analytics", *Journal of Big Data*, Vol. 2, No. 1, 2015.

[14] J. W. Wang, Y. Tang, M. Nguyen, I. Altintas, "*A Scalable Data Science Workflow Approach for Big Data Bayesian Network Learning*", BDC '14 Proceedings of the 2014 IEEE/ACM International Symposium on Big Data Computing, IEEE Computer Society, Washington, DC, USA, pp. 16-25, 2014.

[15] L. Li, "Experimental Comparisons of Multi-class Classifiers", *Informatica*, Vol. 39, pp. 71-85, 2015.

[16] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", *Informatica*, Vol. 31, pp. 249-268, 2007.

[17] N. Karacapilidis, M. Tzagarakis and S. Christodoulou, "On a meaningful exploitation of machine and human reasoning to tackle data-intensive decision making", *Intelligent Decision Technologies*, Vol. 7, pp. 225–236, 2013.

[18] H. Qian, "PivotalR: A Package for Machine Learning on Big Data", *The R Journal*, Vol. 6, No. 1, pp. 57-67, June 2014.

[19] A. Nasridinov, "Combining Unsupervised and Supervised Machine Learning to Analyze Crime Data", *International Journal of Applied Engineering Research*, Vol. 9, No. 23, pp. 18663-18669, 2014.

[20] S. Kanagavalli, S. Vaishali, J. L. Jeba, "Analysis and Mining of Social Network Data For Society Issues By Using Big Data", *International Journal of Applied Engineering Research*, Vol. 10, No. 4, pp. 10497-10506, 2015.

[21] S. R. Sukumar, "*Machine Learning in the Big Data Era: Are We There Yet?*", ACM Knowledge Discovery and Data Mining: Workshop on Data Science for Social Good, Oak Ridge National Laboratory, pp. 1-5, December 2014.

[22] Y. Bu, V. Borkar, M. J. Carey, J. Rosen, N. Polyzotis, T. Condie, M. Weimer, R. Ramakrishnan, "Scaling Datalog for Machine Learning on Big Data", March 2012.

[23] S. Suthaharan, "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning", *Performance Evaluation Review*, Vol. 41, No. 4, pp. 70-73, March 2014.

[24] S. Hido, S. Tokui, S. Oda, "Jubatus: An Open Source Platform for Distributed Online Machine Learning", Technical Report of the Joint Jubatus project by Preferred Infrastructure Inc., and NTT Software Innovation Center, Tokyo, Japan, NIPS 2013 Workshop on Big Learning, Lake Tahoe, pp. 1-6, December 2013.

[25] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, J. M. Hellerstein, "*Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud*", The 38th International Conference on Very Large Data Bases, Endowment, Vol. 5, No. 8, pp. 716-727, 2012.

[26] A. Baldominos, E. Albacete, Y. Saez and P. Isasi, "*A Scalable Machine Learning Online Service for Big Data Real-Time Analysis*", 2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD): proceedings, pp. 1-8, 2014.

[27] A. T. Vu, G. De F. Morales, J. Gama, A. Bifet, "*Distributed Adaptive Model Rules for Mining Big Data Streams*", 2014 IEEE International Conference on Big Data, Washington, DC, pp. 345-353, October 2014.

[28] S. HOI, J. Wang, P. Zhao, and R. Jin, "*Online Feature Selection for Mining Big Data*", BigMine '12 Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, ACM New York, NY, USA, pp. 93-100, 2012.