ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

Survey of Machine Learning Applications

Navdeep Singh Jaggi[1]

School of Engineering and Information Technology, Federation University, Mount Helen 3350, Australia[1]

ABSTRACT: Nowadays big data machine learning algorithms are widely used in Industries, academics, and government institutions. In this era, large amount of industrial data is actively generated and collected in various fields on the world wide web. The main purpose for this research paper was to provide a systematic review of applications of machine learning in versatile field of work and its implementation in ergonomic enhancement of different systems.

KEYWORDS: Big Data, Machine Learning, Methods, and Application

I. Introduction

With increasing sources of data, the amount of data generated around the globe is increasing at a tremendous rate. This augmentation in data creation has come through various sources including such websites as Twitter, LinkedIn and Facebook or by call centres, public surveys and opinion polls. For the analysis of this huge amount of information, all companies require some sort of tool or algorithm. For instance, analysis of the datasets to predict the relationship with customers or social network users.

For working with these type of datasets, machine-learning techniques are very useful and can help to analyse big data and extract important links, rules or other important knowledge. The simplest definition of “big data” is given by [7] using different terminologies such as big data or large data. Large data is in petabytes, exabytes, and zettabytes and beyond. A zettabyte is equal to more than trillion gigabytes (1099511627776 GB exact to 10^21bytes).

Machine learning is a core issue of Artificial intelligence research [1]. According to Shank, "If a computer cannot learn, it will not be called intelligent." Since, learning is an integrative mental activity with memory, thinking, perception, feeling, and other mental activities closely related. So, researchers from different fields give a different interpretation with different disciplines respectively, and give some different points of view. [1] Machine Learning is a subject that studies how to use computers to find human learning activities, and study their behaviour for self-improvement.

II. Machine Learning

Wang Hua et al. [1] introduced the basic model and application in the field of machine learning and gained some preliminary results in machine learning’s application. Firstly, learning an algorithm is only chosen from any classic defining algorithm. Most studies, these days, are based on collecting data and solving a sample problem, but when it comes to working with massive amount of data, then it takes a lot more time to solve the same problem. Furthermore, present methods to identify the type of application are very rough, robust and limited. According to many authors, study on optimal recognition algorithms and combination methods have good prospects for future research.

[1] defines machine learning as “a subject that studies how to use computers to simulate human learning activities, and study self- improvement methods of computer that obtain new knowledge and skills, identify existing knowledge, and continuously improve the performance and achievement.”

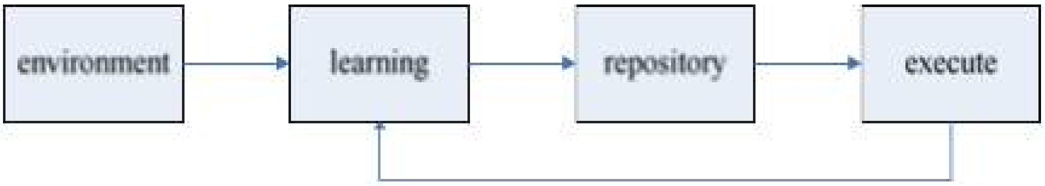


Figure 1: Basic model of Machine learning by Wang Hua and all.

Different Methods of Machine Learning

1. Route Leaning: Route Learning is a memory, that stores new knowledge and calls it, when needed. In the rote learning system, knowledge is accessed in a more stable and direct way. In this method, learning implementation is considered as a function. The function receives input of variables and then calculates the output value of functions. Same as in rote learning, it memorizes the simple storage of input and output variables. When required, the implementation part simply searches the variable from memory and re-calculates it.

The figure below explains learning model used in route learning:

$$(X_1,X_2,\ldots.X_n) \rightarrow \xrightarrow{\text{Storing}} (Y_1,Y_2,\ldots.Y_n)$$

$$((X_1,X_2,\ldots X_n) \rightarrow (Y_1,Y_2,\ldots.Y_n))$$

2. Knowledge Discovery: Knowledge discovery of repository is a senior managed process to identify effective, novel, potential, useful and understanding model from large amounts of data, the discovery process of knowledge is shown in figure 2.

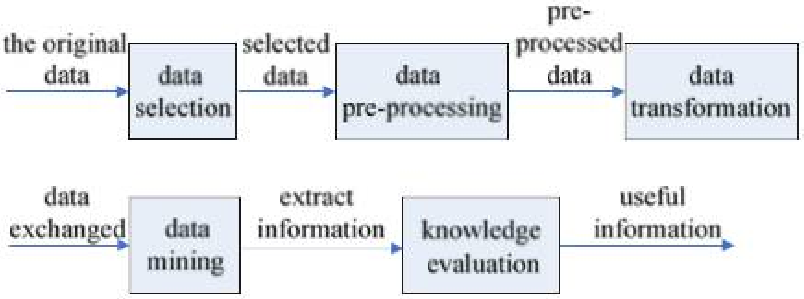


Figure 2: The discovery process of knowledge

Data selection extracts relevant data from database based on the needs of users. Data pre-processing gives data further processing, checks data's integrity and consistency, processes the noisy data, fills up the loss of data by statistical methods and forms explored database.

Applications of machine learning

According to [2], machine learning technology has been widely used in marketing, finance, telecommunications, and network analysis. In the field of marketing, this learning technology is more widely used in tasks classification and other related activities. In the field of finance, this technology is more used for forecasting purposes. In the field of network, machine learning has been used to relate tasks and in telecommunications sector, it has been widely used in the tasks of classifications and prediction.

[2] mentioned that we are in an era of industrial growth that combines computers, sensors, data repositories, high bandwidth networks, mobile device, autonomous machines, and data analytics that drive industrial innovation and growth. More and more industrial data is being collected and stored by these industrial systems. For this reason,

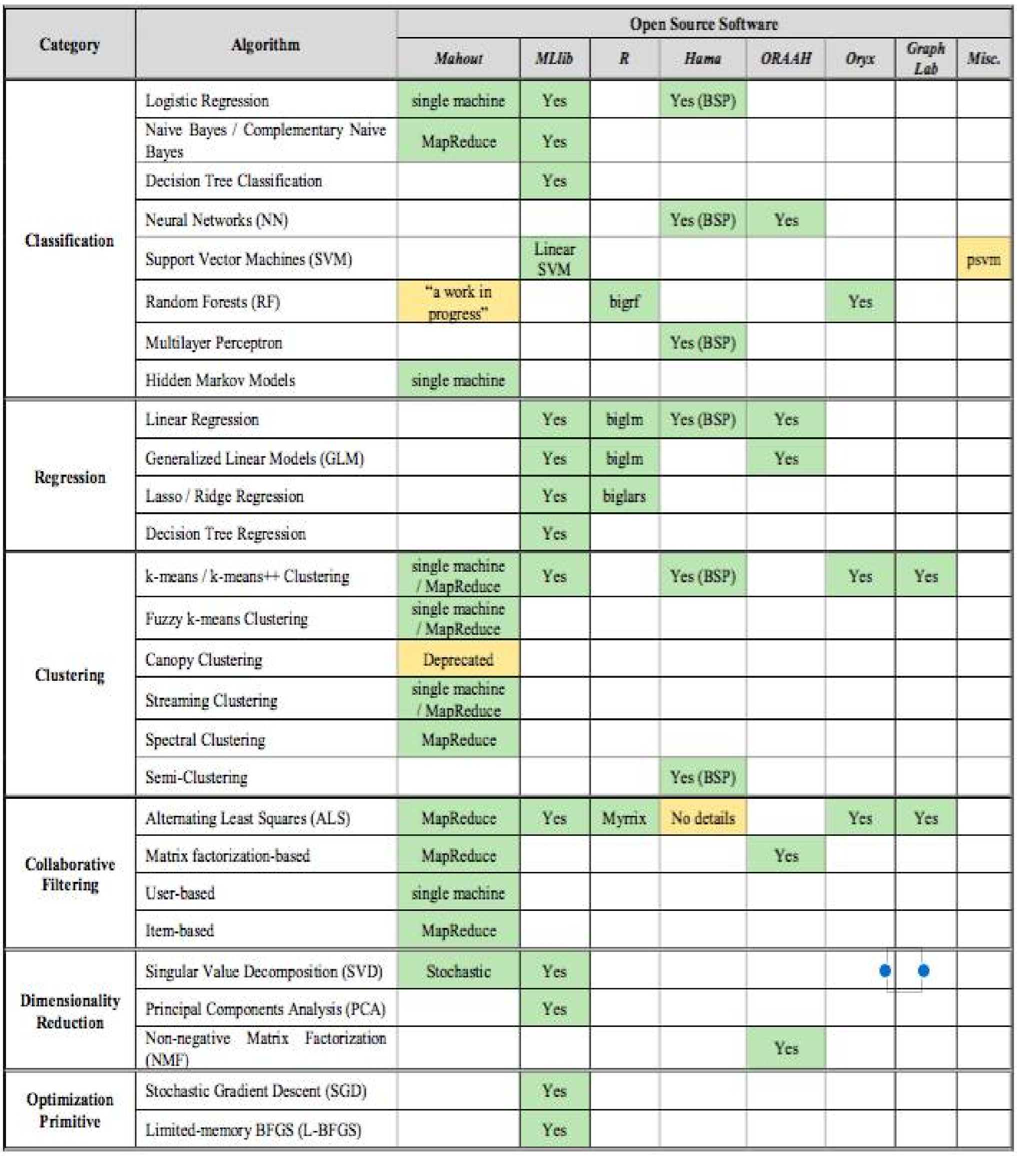
Industrial Analytics require more powerful and intelligent machine learning tools, strategies, and environments to appropriately extract knowledge from the large volumes of industrial data to unleash its great potential value.

They [2] started their research on predictive machine learning analytics for Big Data by conducting a literature survey of machine learning libraries and tools for big data analytics. They did initial studies on how to forecast substation faults and power loading. Furthermore, their results indicated that it is feasible to forecast substations fault events and power load using Naïve Bayes algorithm in MapReduce paradigm or machine learning tools specific for Big Data. They collected more industrial data for their study with most of it based on two cases and more industrial analytics domains in ABB. More statistical and machine learning algorithms will be developed, utilized and verified to mine more values from their industrial data. Table 1 shows the overview of open source machine learning tools for big data.

One of the major challenges faced by [2] was the implementation of complex machine learning algorithms, such as Neural networks, in MapReduce paradigm. Mahout made a proposal to implement the Neural Network with back-propagation learning on Hadoop, but had been never implemented so far. A second challenge faced was that before building machine-learning models, we usually need to conduct basic statistics to examine the dataset for better understanding. Open source tools, investigated by them, lacked powerful statistic functionalities for Big Data. They wrote programs to calculate median, mode, correlation and quartiles in the Scala programming language using Spark shell script.

According to [3], Big Data machine learning and graph analytics have been widely used in industry, academia and government. Continuous advancements in these areas has been crucial in many businesses success, scientific discoveries, as well as cyber security. In this paper, author presents the current projects and propose some next generation computing systems for big data machine learning and graph analytics need, innovative designs, in both hardware and software, that provide a good match between big data algorithm and the underlying computing and storage resource.

[3] divided the big data computing systems into two major categories, i.e., Batch processing, and streaming processing.



Category	Algorithm	Open Source Software							
		Mahout	*MLlib*	*R*	*Hama*	*ORAAH*	*Oryx*	*Graph Lab*	*Misc.*
Classification	Logistic Regression	single machine	Yes		Yes (BSP)				
	Naive Bayes / Complementary Naive Bayes	MapReduce	Yes						
	Decision Tree Classification		Yes						
	Neural Networks (NN)				Yes (BSP)	Yes			
	Support Vector Machines (SVM)		Linear SVM						psvm
	Random Forests (RF)	"a work in progress"		bigrf			Yes		
	Multilayer Perceptron				Yes (BSP)				
	Hidden Markov Models	single machine							
Regression	Linear Regression		Yes	biglm	Yes (BSP)	Yes			
	Generalized Linear Models (GLM)		Yes	biglm		Yes			
	Lasso / Ridge Regression		Yes	biglars					
	Decision Tree Regression		Yes						
Clustering	k-means / k-means++ Clustering	single machine / MapReduce	Yes		Yes (BSP)		Yes	Yes	
	Fuzzy k-means Clustering	single machine / MapReduce							
	Canopy Clustering	Deprecated							
	Streaming Clustering	single machine / MapReduce							
	Spectral Clustering	MapReduce							
	Semi-Clustering				Yes (BSP)				
Collaborative Filtering	Alternating Least Squares (ALS)	MapReduce	Yes	Myrrix	No details		Yes	Yes	
	Matrix factorization-based	MapReduce				Yes			
	User-based	single machine							
	Item-based	MapReduce							
Dimensionality Reduction	Singular Value Decomposition (SVD)	Stochastic	Yes						
	Principal Components Analysis (PCA)		Yes						
	Non-negative Matrix Factorization (NMF)					Yes			
Optimization Primitive	Stochastic Gradient Descent (SGD)		Yes						
	Limited-memory BFGS (L-BFGS)		Yes						

Table 1: Overview of Open Source Machine learning tools for Big Data

Batch processing can analyze large volumes of on-disk data but the processing time is much more compared to streaming processing, e.g. in days and weeks. *Streaming processing* can analyze in memory within a short time period, like milliseconds. While, Batch processing focuses on large data sets, streaming processing only deals with small datasets.

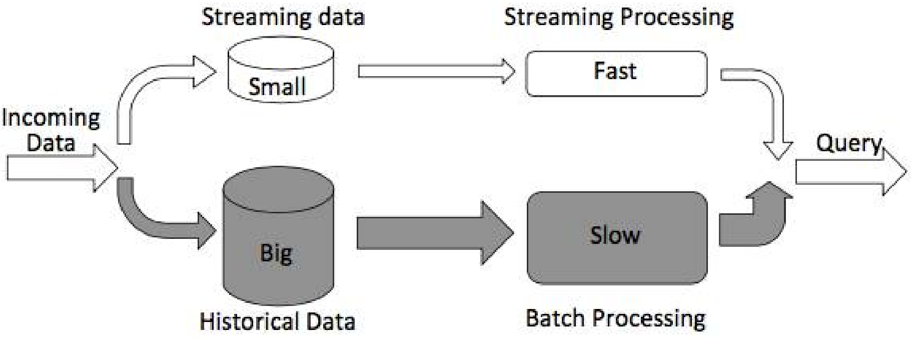


Figure 3: Streaming data and Batch processing

[4] In this paper, authors studied the problem of collaborative machine learning over distributed training data and proposed to use the data locality property of big data processing framework, such as MapReduce, to achieve privacy preservation. In general, this problem is decomposed into subtasks such that each sub task is only related to one share of the training data. After this decomposition, local mappers are able to work independently to get local training results, which are then summarized by a secure protocol on reducer. Figure 4 below shows the structure of system discussed above.

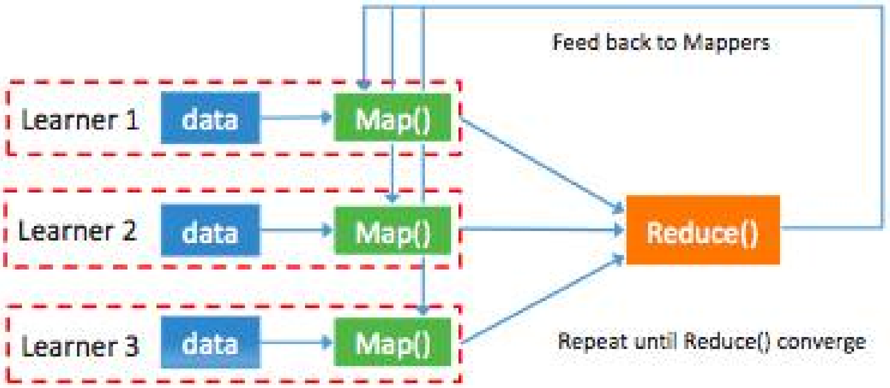


Figure 4: system structure

[5] They introduced and described the novel architecture for performing machine learning on Big Data. Their architecture provides reliable storage on HDFS (Hadoop Distributed File System) and HBase. This architecture is developed using the batch-processing and stream-processing models and this architecture provide machine learning tools and algorithms. In their paper, [5] proposed two useful cases and developed two systems for this task. The first of these is a recommender system for web advertising, which provides real-time advertisement as soon as website is loaded. Whereas, the other used case describes the study of social gamer behavior, in order to read their history of events and predict their future behavior.

Figure 5 shows the purposed architecture supporting machine learning over big data by [5], including a storage module built over a distributed filesystem, batch- and stream-processing modules, a dashboard for displaying results and visualizations and a REST API to bundle the systems supported by this architecture as a service.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

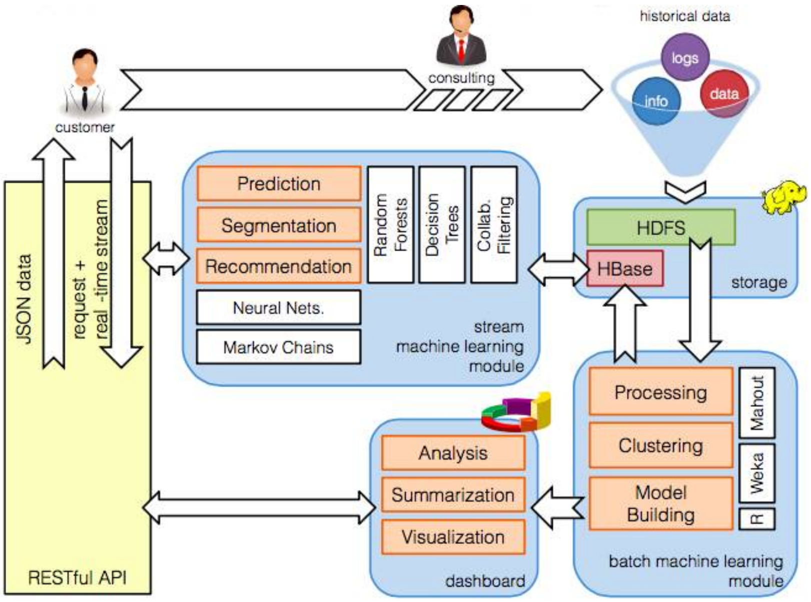


Figure 5: Architecture supporting machine learning over big data, including a storage module built over a distributed filesystem, batch- and stream-processing modules, a dashboard for displaying results and visualizations and a REST API to bundle the systems supported by this architecture as a service

As mentioned earlier, [5] proposed two use cases and developed architecture for both cases. Figure 6 and 7 below shows the architecture of both cases used.

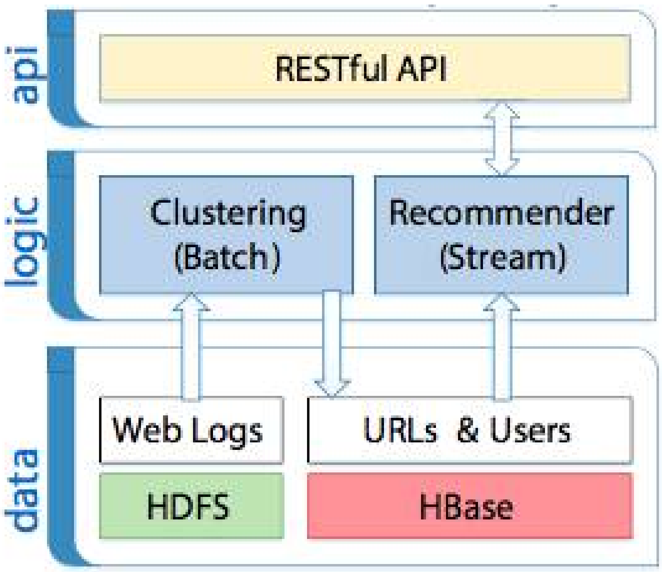


Figure 6: Architecture of the web recommender system. Raw logs are stored in HDFS and a clustering batch process extracts category of both webs and users in a non-time-critical fashion. Then, the stream processor can provide recommendations in real-time as they are requested.

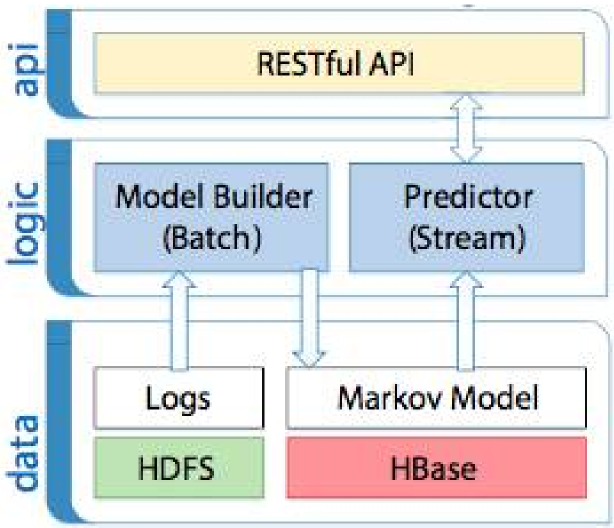


Figure 7: Architecture of the user behavior prediction system. Event logs are stored in HDFS and a job periodically processes them to update a probabilistic model stored in HBase. Then, the stream processor can use the model to predict the future behavior of users given their current behavior

[6] In their paper, the authors considered some of the relevant ideas and have carried out a set of systematic experiments on e-mail categorization. These have been conducted with four machine learning algorithms that are applied to different parts of e-mail.

Professional and un-professional e-mails have become important part of our life to communicate with each other. Whereas, spam emails are becoming a major problem for corporations and common people with a colossal upsurge in leakage of private and confidential information being leaked online. There are many disadvantages associated with spam, such as, wastage of useful storage space and communication bandwidth, time waste in tackling with the spam e-mails, confidentiality issues, viruses, etc. Hence, to deal with this soaring problem, many solutions have been proposed. But the most preponderant solutions ones are such machine learning techniques as naïve bayes, rule learning, decision trees, support vector machines or combination of different learners. The common concept used by all these approaches is using a classifier to filter out the spam and these classifiers are based on learning from trained data, instead of constructing by hand.

Naïve-Bayes: Naïve-Bayes classifier differentiates the e-mails into spam or non-spam, based on the words present in the e-mail.

$$C_{NB} = \arg\max_{c_i \in T} P(c_i) \prod_k P(w_k \mid c_i),$$

Where, T is the set of target class and P(wk|ci) is the probability that word w_k occurs in the email and e-mail belongs to class Ci.

Term frequency-inverse document frequency: Weight vectors represent a set of messages, which are used in Vector space models. The weight of a term in a message can be computed by tf . idf. The tf (term frequency) indicates the number of times that a term t appears in an e-mail.

$$w_{ij} = tf_{ij} \cdot \log(\frac{N}{df_i}),$$

DOI: 10.15680/IJIRCCE.2017. 0508001

K-nearest neighbor: It has a very good performance on text categorization tasks. Firstly, the emails are indexed and then converted into document vector representation. When classifying e-mail, comparison between its document vector and each training set has to be computed.

Support Vector machine: Support vector machine is widely used because of its high performance and its capability to handle high dimensional data. An e-mail is categorized either as spam or non-spam by carrying a simple dot product between the features of an e-mail and the SVM model weight vector,

$$y = w \cdot x - b \,,$$

Where w is weight vector corresponding to feature vector x and b is bias parameter in SVM model.

Thus, if the performance of the above approaches is compared, then Naïve-bais, TF-IDF performs better with header information of e-mail. But, SVM requires both, body as well Header for the high accuracy. The performance of K-nearest neighbor is worst in comparison to the others. Among all, the classification with the header was the most accurate compared to all the other parts of e-mail.

III. Conclusion and Future Work

According to Jainendra Singh, “Machine learning is a subfield of artificial intelligence concerned with techniques that allow computers to improve their outputs based on previous experiences” [7]. In this paper, author gives introduction of machine learning and some fields where machine learning can be used. Although researcher got some good output in the field of machine learning, still some many issues need to be resolved. This field is closely related to data mining and uses techniques from statistics (density estimation etc), pattern recognition and different areas. Nowadays, many big tech companies, including Facebook, Google, Uber, Amazon implemented different machine learning algorithms in their products.

References

1. Hua, Wang, M. A. Cuiqin, and Zhou Lijuan. "A brief review of machine learning and its application." Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on. IEEE, 2009.
2. Zheng, Jiang, and Aldo Dagnino. "An initial study of predictive machine learning analytics on large volumes of historical data for power system applications." Big Data (Big Data), 2014 IEEE International Conference on. IEEE, 2014.
3. Huang, He Helen, and Hang Liu. "Big data machine learning and graph analytics: Current state and future challenges." Big Data (Big Data), 2014 IEEE International Conference on. IEEE, 2014.
4. Xu, Kaihe, et al. "Privacy-Preserving Machine Learning Algorithms for Big Data Systems." Distributed Computing Systems (ICDCS), 2015 IEEE 35th International Conference on. IEEE, 2015.
5. Baldominos, Alejandro, et al. "A scalable machine learning online service for big data real-time analysis." Computational Intelligence in Big Data (CIBD), 2014 IEEE Symposium on. IEEE, 2014.
6. Lai, Chih-Chin, and Ming-Chi Tsai. "An empirical performance comparison of machine learning methods for spam e-mail categorization." Hybrid Intelligent Systems, 2004. HIS'04. Fourth International Conference on. IEEE, 2004.
7. Singh, Jainendra. "Real time BIG data analytic: Security concern and challenges with Machine Learning algorithm." IT in Business, Industry and Government (CSIBIG), 2014 Conference on. IEEE, 2014.

BIOGRAPHY

Navdeep Singh Jaggi is currently pursuing Ph.D. (Engineering) and working as a sessional lecturer in the Faculty of Science and Technology at Federation University Australia. He was awarded with a Tuition Fee-Waver Scholarship for the year 2016-2019 by the Faculty of Science and technology, Federation University Australia. He has a master’s degree in Mining Engineering from Federation University Australia and Bachelor’s degree in Geoscience Engineering from University of Petroleum & Energy Studies, India. His research interests range in versatile fields including CAD and Numerical modelling, Sentiment Analysis, Mining Engineering, Soil Mechanics, Rock Mechanics and engineering, Big Data. He is also a student member of Engineers Australia, AusIMM and Golden Key International Honour Society.