Article can be accessed online at http://www.publishingindia.com

Survey on Big Data and Machine Intelligence Tools

Shyam Mohan*, Shanmugapriya P.**

Abstract

Data is growing at an exponential phase today that posing challenges in analyzing, handling and sharing. The task of choosing the correct machine learning tools for such huge datasets is a difficult task. Each tool have their own limitations. Traditional tools fail to perform real time processing of huge datasets. This paper is intended for the individuals those who are interested to know about machine intelligence tools and how they are related to perform big data analytics. We have given the overview of each tools that are available with their latest versions and releases. To begin with, we have started with the introduction to big data, Hadoop and machine intelligence techniques. Then we go to the machine intelligence tools and understand the application areas where they can be implemented. We discuss the key features of each tool and provide a comparative study of all the tools. So, this paper aims to help the users to choose or take decisions easily in choosing the tools.

Keywords: Big Data, Hadoop, Machine Learning

Introduction

We are living in the era of data where the data is growing in terms of magnitude faster than ever before. A study made by International Data Corporation Digital Universe [1] states that the amount of data will reach 44 zetta-bytes (4.4×1022 bytes) by 2020 that is more than ten times larger than the data in 2013. Many industries today are still using traditional techniques for processing huge and complex data. For example, Ancestry.com stores billions of records about 10 peta-bytes of data [2]. With the advent of high performance machines becoming more popular for capturing, finding and analyzing huge datasets, Machine Learning techniques have become the host of many industries like finance, e-commerce, and entertainment etc.[3]. The main aim of machine learning is to learn from the past or present and make predictions or decisions for the future. The main part of machine learning is the datasets that are available for making predictions. The success of machine learning depends on how effectively it can process the big data. Hadoop provides working with big data [4]. Choosing the appropriate tools for a particular task or environment is a difficult task because there is no single tool or framework that can perform all the trade - offs.

The goal of this paper is to provide a comprehensive review of the current open source tools for machine learning assuming that the reader some basic knowledge of machine learning. It is intended for people who are interested to learn machine learning and big data.

Sharing of data is done in many ways. There are free models that provide free data sharing and are open like Berkeley Software Distribution (BSD) [5]. While performing data analysis over collected data, some of the data may be incomplete or may contain outdated data as the data is collected from a wide variety of resources. However, there are some policies laid down for big data for data sharing, distribution and data efficiency etc. Hadoop is an open source tool for Big Data. Apart from Hadoop, there are many open source tools for machine learning. Example: WEKA – Massive Online Analysis that performs online streaming analysis.[6]. MADLib that is designed to run and scale in the database which uses a collection of SQL based algorithms that include clustering, classification etc. to perform validation. [7].

* Assistant Professor, Department of CSE, Sri Chandrasekharendra Saraswathi Viswa University, Enathur, Kanchipuram, Tamil Nadu, India. Email: browseshyam100ster@gmail.com

** Sri Chandrasekharendra Saraswathi Viswa University, Enathur, Kanchipuram, Tamil Nadu, India. Email: priya_prakasam@yahoo.co.in

GraphLab is used in graph analysis when used along with Hadoop. Today it is a commercial product available with Github that uses C++ and Python Libraries for performing graph analysis.[8].

Cloud and Big Data

Many related works have been done in Cloud and Big Data that can be found in [9][10]. Some of the services offered by Cloud are:

1. Software as a Service (SaaS) – Any software that is accessible through the Cloud.[11][12].
2. Platform as a Service (PaaS) – customize or create software applications in cloud.[13].
3. Infrastructure as a Service (IaaS) – Provides storage and processing that relies on SaaS and PaaS. [14].

Some of the Big Data Cloud services are:

1. Google Cloud - https://cloud.google.com/products.
2. Amazon Cloud Services - http://aws.amazon.com.
3. IBM Cloud Services - www.ibm.com/cloud.

Lack of effective mechanisms for deployment, performing analytics, etc. are barriers involved in transferring terabytes or peta-bytes of data [15].

The goal of this paper is to provide a comprehensive review of the current open source tools for machine learning assuming that the reader some basic knowledge of machine learning. It is intended for people who are interested to learn machine learning and big data.

Big data – An User's perspective

"Big Data " has become a buzzword today [16]. So from users perspective, we define what is big data as follows:

Volume, that refers to the huge size of the data.

Velocity, the speed at which the data can be processed.

Variety, different forms of data, like sensor data, satellite data, etc.

The problem of solving big data collections today is solved by distributed storage systems that are capable of fault-tolerance. It is also done by machine learning or parallel algorithms [17]. Hadoop provides an effective way to handle big data.

Data Processing Engines

In 2004, Google introduced MapReduce for processing big data [18] that was later regarded as Hadoop. Subsequent years of development of Hadoop paved way for machine learning techniques. Generally, Big data is processed in two methods viz., batch processing or stream processing. Bulk Synchronous Parallel (BSP) model is used for performing iterative tasks. Example: Apache Hama [19] etc. Many algorithms uses various color images with embedded watermarking techniques for resolving the visual distortions.

Table 1: Hadoop Data Processing Tools

SI.No.	*Name*	*Version*	*Release Date*
1.	Apache Hadoop	2.7.2	26.1.2016
2.	Spark	1.6.1	9.3.2016
3.	Flink	1.0.0	8.3.2016
4.	H2O	NA	1.3.2016

Machine Learning Techniques

Finding a right tool for capturing and performing analysis for huge datasets for effective data processing is a difficult task.[20]. Generally, machine learning techniques provide effective way to process huge datasets where traditional tools fail and they are helpful in finding out hidden details in big data by extracting values from the huge datasets without human intervention that can be applied even to growing datasets. The tools mentioned in this paper provides various tools used in data processing.

Table 2: Types of Machine Learning Techniques

Type	*Attributes*
Supervised Learning	It is based on labeled training data.
Unsupervised Learning	It is based on unlabeled data. It is used for finding unseen relationships in the data independent of class label. It is used in clustering.
Semi-supervised Learning	It is combination of labeled and unlabeled data.

Table 3: Machine Learning Toolkits

Sl.No.	*Name*	*Latest Version*	*Release Date*
1.	Mahout	0.11.2	12.3.2016
2.	MLlib	1.6.1	9.3.2016
3.	Scikit Learn	0.17.1	NA
4.	Shogun	4.1.0	9.2.2016
5.	AccordFramework	3.0	NA
6.	Cloudera Oryx	1.1.0	6.7.2015
7.	Weka	3.7	27.3.2016
8.	ConvNetJS	NA	NA
9.	NVIDIA Cuda	7.5	14.12.2015

A. Mahout

The latest version of Mahout 0.12.0 has three major components, Scala+Spark, H2O and Flink.[21]. It has distributed algebraic optimizer, IScala, etc. and supports Stochastic Singular Value Decomposition and Principal Component Analysis, Distributed Cholesky QR and regularized Alternating Least Squares, Collaborative Filtering, Naive Bayes Classification. Algorithms based on Mahout are shown in table 4.

B. MLlib Algorithms

It is a scalable machine learning library built in Apache Spark that interoperates with NumPy in Python. It can be used in Hadoop data source like HDFS, HBase, etc., that makes it easier for performing workflows. It is built for high-quality algorithms that are 100 times faster than MapReduce. See table 5 for more details.

C. Apache Spark

The latest version of Spark 1.6.1 is a fast and general-purpose cluster computing system that provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, Graph X for graph processing, and Spark Streaming.[22]. MLlib contains the following algorithms that rely on machine learning.

D. Scikit Learn

Scikit learn version 0.17 is an open source, commercial software that is used for data mining and analysis. It is built on NumPy, SciPy and matplotlib. Scikit Learn version 0.18 is under development.[23]. Further details can be found in table 6.

E. Shogun

It provides efficient machine learning method for a wide variety of datasets. Open-source ML software allows effective binding of libraries like Lib Linear, Vowpal Wabbit, etc. for data processing. [24]. It uses C++, Python, etc., and runs on Linux, Mac OS and Windows. The latest version of Shogun 4.1.0 was released on February 9, 2016. Shogun includes machine learning algorithms like Classification, Regression, Dimensionality Reduction, Clustering, Metric, Multi-Task, Structured Output, Online Learning, Feature Hashing, Ensemble Methods, Optimization, SVM, Multiple Kernel Learning and Krylov Methods.

F. Accord.Net Framework

Accord.NET is a framework built in visual studio used for scientific computing. It is written in C#. The framework consists of libraries available through NuGet Packages that fit for a wide range of applications such as statistical data processing, machine learning, computer vision etc. [25]. It even supports image processing. It consist of more than 40 different statistical distributions that can be in Hidden Markov models and more than 38 kernel functions to support Kernel Support Vector Machines, Kernel Principal Components and Kernel Discriminant Analysis. The library functions and their applications are given in table 7.

Table 4: Mahout Machine Learning Algorithms

Type and description	*Single Machine*	*Map Reduce*	*Spark*	*H2O*	*Flink*
Mahout Core Library and Scala					
Distributed ALS,etc.			X	X	**X**
Collaborative Filtering					

Type and description	Single Machine	Map Reduce	Spark	H2O	Flink
User-Based Collaborative Filtering	**X**		X		
Item-Based Collaborative Filtering	X	X	X		
Matrix Factorization with ALS	X	X			
Matrix Factorization with ALS on Implicit Feedback	X	X			
Weighted Matrix Factorization, SVD++	X				
Classification					
Logistic Regression - trained via SGD	**X**				
Naive Bayes / Complementary Naive Bayes		X	X		
Random Forest		X			
Hidden Markov Models	X				
Multilayer Perceptron	X				
Clustering					
k-Means Clustering	**X**	**X**			
Fuzzy k-Means	X	X			
Streaming k-Means	X	X			
Spectral Clustering		X			
Dimensionality Reduction using Mahout Math-Scala Core Library					
Singular Value Decomposition	**X**	**X**	**X**	**X**	**X**
Stochastic SVD	**X**	**X**	**X**	**X**	**X**
PCA (via Stochastic SVD)	**X**	**X**	**X**	**X**	X
QR Decomposition	X	X	X	X	X
Latent Dirichlet Allocation	X	X			

Table 5: MLlib Algorithms

SI.No.	Algorithm
1.	Linear Support Machine(SVM)
2.	Classification and Regression Tree.
3.	Random Forest
4.	Clustering:K-Means,Gaussian Mixtures(GMM)
5.	Latent Dirichlet Allocation (LDA).
6.	Singular value Decomposition(SVD) and QR decomposition
7.	Principal component Analysis(PCA)
8.	Linear regression with L1, L2, and elastic-net regularization
9.	Isotonic regression
10.	multinomial/binomial naive Bayes
11.	frequent itemset mining via FP-growth and association rules
12.	sequential pattern mining via PrefixSpan
13.	summary statistics and hypothesis testing
14.	feature transformations
15.	model evaluation and hyper-parameter tuning

Table 6: Algorithms Based on Scikit Learn

Sl.No.	*Type*	*Algorithm*	*Description*	*Example*
1.	**Classification**	SVM, Random Forest, Nearest Neighbours.	Identify the objects and their category.	Image detection.
2.	**Regression**	SVR, ridge regression, Lasso.	Predicting continuous object valued attribute.	Stock prices.
3.	Clustering	K-Means, spectral clustering, mean-shift	Grouping similar objects into clusters.	Customer segmentation.
4.	**Dimensionality reduction**	PCA, feature selection, non-negative matrix factorization.	Reduces the number of random variables.	Visualization.
5.	**Model selection**	grid search, validation, metrics.	Parameter choosing , comparing and validating.	Parameter tuning.
6.	**Preprocessing**	Feature extraction.	Feature extraction and normalization.	Transforming input data for machine learning.

G. Cloudera Oryx

Innovations in cloud, use cases and new techniques is done by Cloudera. Oryx is a cloud based lambda architecture that is built on Apache Spark and Kafka (specialized in machine learning).Some of its key features include: collaborative filtering, classification, regression, clustering, random decision forest, K-Means etc. [26]. Some of the projects in Cloudera labs are: Apache HTrace, YCSB, Apache Phoenix, Ibis, Impyla, etc. For further reading, refer to [27].

H. WEKA

One of the most popular and mostly used machine learning software is WEKA (Waikato Environment for Knowledge Analysis) written in Java that was developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. [28][29].

Some of the related tools used in WEKA are given below:

1. Environment for Developing KDD-Applications Supported by Index-Structures (ELKI) - It focuses on cluster analysis (unsupervised methods).
2. KNIME - It is a combination of machine learning and data mining software that is implemented in Java.
3. Neural Designer - Data mining software which is written in C++.
4. Orange - Open source project written in C++ and Python.
5. Rapid Miner - Integrates Weka that is written in Java.

I. Conv Net JS

It is a deep learning model that is based on neural networks which runs on Javascript library for which no GPU is needed.[30].

It includes:

1. Common Neural Network modules (fully connected layers, non-linearities).
2. Classification (SVM/Softmax) and Regression (L2) cost functions.
3. Ability to specify and train Convolutional Networks that process images.

J. Nvidia CUDA

Compute Unified Device Architecture is popularly known as CUDA. The toolkit latest version is 7.5. Most of the code in CUDA is written in C ,C++ and Fortran.[31] [32]. It is a parallel computing platform that is built for running parallel tasks is CUDA® invented by NVIDIA by increasing the computing performance by graphics processing unit (GPU). Some of the applications where CUDA is used:

1. Identifying hidden plaque in arteries.
2. Analyzing air traffic flow.

Molecular visualization called as Nanoscale Molecular Dynamics (NMD).

CUDA is provided with the GPU Wizard that provides runtime analysis and reports potential speedups and performance tuning. NSight Eclipse Edition supports multiple toolkit versions to support cross compilation. It is supported by IBM Power 8 platform. CUDA supports programming frameworks such as Open ACC and Open CL.[33].

A comparison of popular machine learning toolboxes is shown in table 8.

Table 7: Accord .NET Framework Applications and Libraries

Type	*Library function*	*Application and examples*
Scientific Computing	.Math	Numerical optimization algorithms, constrained and unconstrained problems.
	.Statistics	Linear regression, Logistic regression, Hidden Markov Models, (Hidden) Conditional Random Fields, PCA, Partial Least Squares, Discriminant Analysis, Kernel methods.
	.MachineLearning	SVM, Decision trees, Naïve Bayes, K-Means , Grid Search.
	.Neuro	Deep learning, Levenberg-Marquardt (LM), Parallel Resilient Back propagation,
Signal and Image Processing	.Imaging	SURF, FAST, image matching, image stitching, image transformations,
	.Audio	Signal processing, transformations of audio signals.
	.Vision	Real-time face detection and tracking, image tracking using dynamic template matching.
Support Libraries	.Controls	Used in drawing histograms, scatter-plots, etc.
	.Controls. Imaging	Displaying images.
	.Controls. Audio	Display audio related information.
	.Controls.Vision	Tracking head, face and hand movements and other computer vision related tasks.

Table 8: Comparison of Machine Learning Algorithms

Type	*Application / Purpose*	*Language*
Shogun	ML Package built on large scale learning; Kernel Methods and provides Interfaces to various languages.	C++
Scikit-learn	General Purpose using simple API's like numpy ,etc.	Python
Weka	General Purpose ML Package	Java
Shogun	Provides efficient machine learning method for a wide variety of datasets.	C++,Python
Accord. NET	Framework used for scientific computing.	C#
MLlib	Supports many machine learning algorithms.	
ConvNetJS	Deep learning model based on neural networks in which no GPU is needed.	Javascript
Mahout	Runs on Scala, Spark, H2O and Flink	Java

Table 9: Comparison of Machine Learning Algorithms based on Their Type, Classification and Application

Type	*SVM*	*Naive Bayes*	*Random Forest*	*Hidden Markov Models*	*Multilayer Perceptron*	*k-Means*	*Fuzzy k-Means*	*Streaming k-Means*	*Spectral Clustering*	*SVD*	*PCA*	*QR Dec.*	*LDA*
Mahout MapReduce		X	X			X	X	X	X	X	X	X	X
ConvNetJS	X			X									
MLlib				X	X			X	X				
Scikit Learn		X		X	X					X		X	X
Accord.Net				X	X				X				
Shogun				X	X								
Weka			X		X	X	X						
NVIDIA Cuda						X	X						
Cloudera Oryx			X	X		X			X				

Conclusion

This paper provides a survey on various tools available for peforming analytics. This paper is intended for the users who want to know about the machine intelligence tools available for performing analytics. Table 9 shows the summary of all the tools listed above so that the users can choose the tool of their own based on their working classification. We thank all the members who have extended their support to complete this survey paper.

References

International Data Corporation. Digital Universe Study. (2014). Retrieved from http://www.emc.com/leadership/digital-universe/index.htm.

Ancestry.com Fact Sheet. http://corporate.ancestry.com/press/company-facts/.

Landset, S. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(24).

Apache Hadoop. Retrieved from https://hadoop.apache.org/.

Feller J., & Fitzgerald, B. (2002). *Understanding open source software development.* Addison-Wesley, London, Retrieved from http://dl.acm.org/citation.cfm?id=513726.

MOA (Massive Online Analysis). Retrieved from http://moa.cs.waikato.ac.nz/.

Hellerstein, J. M., Schoppmann, F., Wang, D. Z., Fratkin, E, Welton, C., Feng, X., Li, K., & Kumar, A. (2012). The MADlib Analytics Library or MAD Skills. *The SQL.In: VLDB Endowment*, (pp. 1700-171).

Dato Core. Retrieved from https://github.com/dato-code/Dato-Core.

O'Driscoll, A., Daugelaite, J., & Sleator, R. D. (2013). 'Big data', Hadoop and cloud computing in genomics. *Journal of Biomedical Informatics*, 46(5), 774-781

Bellini, P., di Claudio, M., Nesi, P., & Rauch, N. (2013). Tassonomy and review of Big data solutions navigation. *In Big Data Computing. Chapman and Hall/CRC, Boca Raton*, (pp. 57).

Howell-Barber, H., Lawler, J. P., Joseph, A., & Narula, S. (2013). A study of cloud computing Software-as-a-Service (SaaS). *Financial Firms. Cloud Computing,* Special Issue.

Foster, I., Yong, Z., Raicu, I., & Shiyong, L. (2008). Cloud computing and grid computing 360-degree compared. *Grid Computing Environments Workshop*, 2008. GCE'08, Austin, Texas., Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4738445.

Lawton, G. (2008). Developing software online with platform-as-a-service technology. *Computer*, June, 41(6), 13-15.

Bhardwaj S, Jain L, & Jain, S. (2010). Cloud computing: A study of infrastructure as a service

(IAAS). *International Journal of Engineering and Information Technology*, 2(1), 60-63.

Schutt, R., & O'Neil, C. (2013). *Doing Data Science: Straight Talk from the Frontline.* O'Reilly Media, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2544025.

Laney, D. (2001). *3D data management: Controlling data volume, velocity and variety*. META Group.

Bekkerman, R., Bilenko, M., & Langford, J. (2011). *Scaling up machine learning: Parallel and distributed approaches*. Cambridge: Cambridge University Press.

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. *In Proceedings of the 6th Symposium on Operating Systems Design and Implementation*.

Apache Hama. Retrieved from https://hama.apache.org/.

http://www.skytree.net/machine-learning/why-do-machine-learning-big-data/

http://mahout.apache.org/users/basics/algorithms.html

http://spark.apache.org/mllib/

http://scikit-learn.org/stable/#

http://www.shogun-toolbox.org/page/features/

http://accord-framework.net/intro.html.

http://www.cloudera.com/developers/cloudera-labs.html

http://oryx.io/.

http://wiki.pentaho.com/display/DATAMINING/Data+Mining+Algorithms+and+Tools+in+Weka

https://en.wikipedia.org/wiki/Weka_(machine_learning).

http://cs.stanford.edu/people/karpathy/convnetjs/index.html

http://www.nvidia.com/object/cuda_home_new.html#sthash.0Vo1PF8C.dpuf.

NVIDIA CUDA TOOLKIT 7.5" Release Notes for Windows, Linux and Mac OS, RN – 06722-001_v7.5, September (2015). Retrieved from

https://en.wikipedia.org/wiki/CUDA