IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Machine Learning Tools and Toolkits in the Exploration of Big Data

Afreen Khan[1], SwalehaZubair[2]*

[1,2]Department of Computer Science, Aligarh Muslim University, Aligarh, India

*Corresponding Author: swalehazubair@yahoo.com, Tel.:+91-9410059635

Available online at: www.ijcseonline.org

Accepted: 28/Dec/2018, Published: 31/Dec/2018

***Abstract*-**Machine learning (ML) is the best way to make progress towards human level artificial intelligence, which allows software applications to become more accurate in predicting results. It is the most promising technique that has profound realization in reorganizing practices pertaining to various fields viz. healthcare, education world industry, retail and manufacturing sectors, traffic and urban planning etc. The compilation and storage followed by specific training of the stored data are some of the salient features of the machine learning process that has tremendous scope in discovering novel output in various relevant fields. There are plenty of tools in ML that may help in the training of data without being explicitly programmed. Tools are categorized into- framework, platform, library, and interface. For the successful development and effective execution of ML, one can categorically manipulate various related tools. Working through such tools advances the process as applied to the various applications. In the present study, we intend to exploit recommendation engines for the development of tools that can handle the huge quantity of data. The usage of the overwhelming quantity of multimodal data and streamlining the same for its personalized usage are some of the unique features of the study. We also focus on the evaluation of a toolkit with loads of data and furthering several ML tools along with their features and use for the desired application in the relevant field.

Keywords-Big data, Application Programming Interface (API), Command Line Interface (CLI), Graphic User Interface (GUI), Machine learning, Tool, Toolkits, Platform, Library, Interface

I. INTRODUCTION

McKinsey, in one of its report, stated, "As ever more of the analog world gets digitized, our ability to learn from data by developing and testing algorithms will only become more important for what are now seen as traditional businesses" [1]. In the past several years, there has been a technological advancement in the data and digitization; and analytics have been restructuring the entire world, amplifying performance and facilitating the advent of recent innovations in the various fields like health, education, agriculture, finance, etc. In order to build a dynamic model, the realm of the technological world is now extended to include autonomous several other domains so as to solve them with various competent computer tools and techniques [2]. This has now become achievable and essential through innovations in the field of Artificial Intelligence (AI). With the advancement in the sphere of AI, there comes Machine Learning (ML) which has brought new and innovative waves in the present technological globe.

When building programs and algorithms become a complex task to achieve, ML aids in unravelling the challenging issues. Along with solving these, understanding the algorithms and their intricacy have increased as well. ML is used in data mining, character recognition, search engine, spam detection, to name a few. These examples echo the essential role and center stage that the ML has grabbed and is all set to hit the technology in today's data-rich world [1].

With the continuous growth and subsequent development in the key functionalities, methods and practices, uses, and the global significance of ML, it is high time and chief concern to have the knowledge of the right ML tools and techniques in order to solve the demanding tasks [3]. As data deluge is on the rise, there has been an increased sophistication of ML whichaids in analyzing the huge amount of Big Data. The present traditional practicesprovide little for applying statistical ML algorithms in dealing with Big Data [4].In the present times, ML have replaced the conventional statistical methods and has also transformed the approach of data extraction, its processing and interpretation by encompassing mechanized groups of generic methods [1].

Around the globe, researchers are laying emphasis on building effective ML tools and techniques to examine the various problem domains. Thus, this paper will include the

budding realm of ML in the age of AI and Big Data Analytics. The paper is organized as follows: The subsequent section involves the state of Machine Learning as an indispensable technique. Section 3 discusses how to evaluate an ML toolkit. Section 4 presents the various ML tools. Section 5 introduces the comparison of ML tools followed by a conclusion in Section 6.

II. MACHINE LEARNING- AN INDISPENSABLE TECHNIQUE

Due to the pervasiveness of data and the huge scalability of cloud computing power, there has been a massive advancement in the use of AI and ML, which has developed in significance with its competence to filter through large datasets, explore and analyze them, interpret them with the aim of discovering effective patterns and lastly, constructing useful predictions based on the result attained.

As Intel CEO Brian Krzanich in one of his interview stated that "Data is the new Oil" then accordingly, it can be further stated that Machine Learning, a subset of AI, is fuelled by data. It is built on a modelling scheme of not only analyzing data but also, it has the ability to learn- get trained and improve- get better by using different algorithms that provide new and innovative insights [5].

When the talk is about Big Data (BD), ML is the best choice for unravelling the robust issues. In a report, Gartner defined Big Data as, "high volume, high velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization [6]." ML and BD are related in such a way where ML is centered on diverse algorithms that know how to learn from data with no dependency on rules-based programming while BD is the kind of data that is loaded into such a system where analytical practices are carried consequently leading to the improvement in the precision of the predictions that ML model is trained for [7].Big Data analytics aids in acquiring insights and hence therefore, better decisions are taken while modifications in modelling are applied [8].

In a report, published in 2018, Gartner mentioned that "almost half of CIOs are preparing to implement AI in their respective organizations [9]." The main purpose of this is to obtain insights from the collected data so as to understand and acquire knowledge of the respective model used and thereby analysing through their behaviour in order to build improved decisions [5]. On the other hand, the level of success, how enhanced results, how effective predictions is only achieved when the data is used righteously in a way as how better understanding one gain from it [5]. ML and statistical model differ in a way where the aim of ML is to learn the structure of the data where one does not have a knowledge behind the theory of what the structure appears to be and completely understood theoretical distributions are then adjusted to the data whereas in statistical model, every model has a theory behind it which is scientifically proven but the rule is that the data should meet some robust assumptions [10]. Furthermore, the test for an ML model is a validation error that is performed on a new data,however, a theoretical test is never carried out which is used to prove a null hypothesis [10]. ML uses an iterative methodology to learn and discover from data until a strong pattern is obtained and this acquisition of knowledge is automated easily.

The strategic rule of an ML model is the ability to independently learn and progress as new data is loaded into the system and to convert this data into actionable knowledge. The chief notion of the entire ML concept is the capability of a model to automatically apply sophisticated mathematical computations to Big Data repeatedly until a most probable solution is achieved [10].

III. HOW TO EVALUATE AN ML TOOLKIT

The first thing to consider when ML need to be adopted is the tool on which the work will be performed. Each step in the ML process can be automated if the right tools are used. The key importance of selecting the correct tool is reflected later when right predictions and improved results are achieved, thereby it is as essential as working with the finest algorithms [12]. The toolkit provided differs considerably, therefore it is necessary to maintain an equilibrium between keeping up with the latest developments and rigid reliability and stability of a project [11]. These tools not only provide the facility of implementation of ML algorithms but also support at every step while the tasks are being executed and can be used throughout the ML challenge.

ML advances well only when critical decisions need to be taken which thereby is built on assumptions that are generated from the analysis of data [13]. Thus, there isno one particular criteria for deciding the best toolkit for ML. Each and every toolkit is developed to focus on the needs as observed by the developer. Presently, there are numerous ML toolkits available and how to evaluate a specific tool is an essential issue in the current times so as to deal with the problem statement in a most practicable way. Following are a range of different criteria that is usually used to assess any tool:

1. **Language:** In regard to developing the ML models and writing the ML algorithms, the toolkit's programming language in which it is written in, influence the entire modelling. The choice of language depends on the comfort level i.e. ability to use it efficiently, type of problem to be solved, the quantity of data to be processed, and

developer's expertise and past experience [14]. There are certain factors to consider before selecting a language, likewise, speed, concurrency, performance, cost effectiveness, learning curve i.e. functional or procedural, application development, and community support [15].

2. **Type:** This includes the different categories in which the toolkits are divided. The categorization done is as follows: platform, library, interface and local or remote tool.

3. **Documentation:** The documentation of a specific toolkit plays an important role in deciding which one to choose and which one to avoid. If a toolkit is documented well in terms of quality, coverage of a huge number of examples that look similar to problems one work on, then it is easier to build a solution to the particular problem.

4. **Integrated Development Environment (IDE):**The IDE used for ML is as important as the ML techniques that are used to solve the predictive modeling challenges. Certain toolkits have graphical IDE, and others include command line and editor IDE.

5. **Execution Speed:** The execution speed of ML toolkit is a speed with which the algorithms and models execute the instructions and classify the tasks via a trained model. It is faster than training but it is not as important as the training speed that trains the data in order to predict through the test data.

6. **Training Speed**: The training speed is of great importance as compared to the execution speed. It depends on the efficiency of different math libraries that the toolkit consists of, and exactly how these libraries use the existing computational resources. Also, it heavily depends on the kind of the problem to be solved, the trained data and the images.

7. **GPU (Graphical Processing Unit) Support:**GPUs are chiefly used to accomplish and increase the performance of video and graphics, and boost the network's learning speed. They speed up the ML algorithms by latency reduction, bandwidth increment, and lessening the communication cost.

IV. ML TOOLS

Even though a good set of ML algorithms exist, the recent development illustrates the capability of ML to apply sophisticated mathematical operations to data and refine them rapidly as soon as the processing begins [5]. Mostly used ML processes and algorithms aredecision trees, Bayesian networks, support vector machines, random forests, gradient boosting and bagging, self-organizing maps, Gaussian mixture models, k-means clustering, neural networks, and more to the list. As more and more computation power is required, a huge interest is rising in exploiting ML in addition to the growing access to heaps of data so as to gain action-driven advantages. To the same degree, there exist plenty of tools in ML, when machines need to be trained to work without being explicitly programmed [13]. The nucleus of the ML system consists of ML storage cluster and its computation power, which usually differs based on the learning method used, its application and the need to automate it [5].

ML tool can be a platform, library, an interface or any local or remote tool. These toolkits provide the developers to create ML models more quickly and easily without stepping into the details of the core algorithms thereby providing a well-defined and brief approach for classifying ML models by applying a set of pre-built and improved modules [16]. The tools can be divided into four branches, as depicted in Figure 1 below.

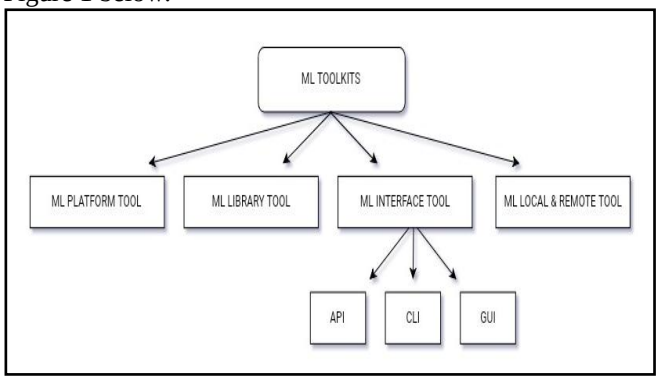

Figure 1. Classification of ML Toolkits

Significant attributes of a good ML toolkit are developer friendly, optimized for performance, language and coding choice, ease of understanding, and computation process parallelization among the various processes [2]. In general, an effective ML toolkit lessens the complex nature of ML, causing it to be user-friendly and understandable to new developers. The 5 tools defined above are further described below, and in particular, specific examples belonging to those are illustrated well in the table-format.

1. **ML Platform:** ML platform provides an environment where an ML project can be completed from the start till the end. It offers capabilities like data pre-processing that includes preparing the data so as to model it, data analysis, data modeling and evaluation and selection of an algorithm [12]. There are many factors that become relevant when the ML platform toolkit needs to be implemented. They are: type of data, its characteristics, features of automation, characteristics pertaining to ease of use and integration, algorithm and modeling techniques, supported open source resources and management features [17]. Some of the characteristics of ML platform are:

- Provide complete facility needed at every step of an ML project development.
- The ML platform interface may include API, CLI or GUI, or a combination of these while programming.
- They are used for general purpose modelling, instead of focusing on accuracy, scalability and speed.
- Features are loosely coupled, therefore it is the task of a user to assemble all the components collectively for the particular project.

2. ML Library: ML library contains capabilities for only finishing a fragment i.e. one or more steps of an ML project [12]. It is used to unravel the predictive use-cases. It includes facilities like documentation, configuration and help data, pre-written subroutines and code, and message templates [18]. An ML library also provides modeling algorithms that are suitable for particular use-cases, each having their own pros and cons. When determining which ML library to apply, several features need to be considered. They are programming language, performance and hardware features, and ML algorithm [19]. Certain characteristics of the ML library are:

- ML library interface is usually an API which involves programming.
- They are designed for a particular use-case or environment.

3. ML Interfaces: ML interface is another ML tool and is further forked into three parts, namely, ML API, ML CLI and ML GUI.

i. ML API (Application Programming Interface): Machine learning tools support an API which provides the ability to decide what components to work with and how to apply them in the ML programs [18]. The characteristics of ML API are as follows:

- It provides the capability of developing our own ML tools.
- ML API tool can be used to build our own processes, and thereby, it can be further implemented on ML projects so as to automate them in an improved way.
- It gives the flexibility to develop our own methods, merge them with existing the libraries and methods.

ii. ML CLI (Command Line Interface): Machine learning tools give an environment of CLI that focuses on input and output i.e. it structures ML tasks in terms of the required input and output to be produced [12]. In addition, it also comprises of command line parameterization and command line programs. The characteristics of ML CLI are as follows:

- It provides such an environment where non-programmers can perform their tasks through ML projects.
- It has the ability to reproduce the results by storing the commands and command line arguments.
- It supports several small programs and many program genres for certain subtasks of ML project.

iii. ML GUI (Graphical User Interface): Machine learning tools support a GUI which mainly focuses on the graphical representation of data i.e. visualization and also consists of windows, point and click [18]. The characteristics of ML GUI are as follows:

- The users that are not an expert in programming, for those, ML GUI provides an environment where they can complete their tasks easily through ML.
- The chief emphasis of ML GUI is on the process and by what means maximum information can be extracted from the ML tools and techniques.

4. A. ML Local Tool: Machine learning Local tool is a tool that can be downloaded, installed and can be used and run in the local environment. The characteristics of ML Local tool are as follows:

- It is built for main memory data and its algorithms.
- This ML tool can be incorporated into our own ML machines so as to model it according to our needs
- It provides control over the parameters so as to devise predictions on newer data thereby supporting the run configuration of the system.

4. B. ML Remote Tool: Machine learning Remote tool is a tool that runs on the server of a third-party. It is a tool that is established on a server and the operations are carried out on the local environment by calling it remotely. Thus, ML Remote tools are called Machine Learning as a Service (MLaaS) [12]. The characteristics of ML Remote tool are as follows:

- These ML tools can handle large datasets even though the data scales up rapidly.
- It provides a set up where the processes can run amongst the multiple machines, numerous cores while sharing the memory.
- Because these tools run remotely at scale, it supports less number of ML algorithms since complex modifications are needed.
- It has the ability to get incorporated within our local environments though RPCs (Remote Procedure Calls).

V. ILLUSTRATION OF TOOLKITS

The tools described above are illustrated in the following Table 1.

Table 1. Illustration of ML Toolkits

ML Toolkit	Language	Features and Use
ML PLATFORM		
WEKA ML Workbench	Java	Consists of huge number of ML algorithms, data preparation, data pre-processing, feature selection and visualization methods. It provides all the three interfaces- Java API, CLI and GUI [20].
R Platform	R	Built for data mining majorly. Chief uses are- data manipulation, matrix arithmetic, statistical computing, graphical display, efficient data management and storage capabilities [21].
Subset of the Python SciPy (Pandas and Scikit-learn)	Python	It includes modules for various tasks such as, optimization, image processing, and linear algebra and include tools like Pandas, Matplotlib, SymPy, and a developing set of various scientific computing libraries [22].
ML LIBRARY		
Scikit	Python	Supports numerous ML algorithms like, classification, clustering, regression, support vector machines, k-means, to name a few [23].
JSAT (Java Statistical Analysis Tool)	Java	Consists of many MLalgorithms that are research and need specific, such as data transformation, tree based, predictive, meta algorithms, and algorithms based on vector quantization [24].
Accord (For .NET Platform)	C#	Provides a comprehensive framework for developing applications such as computer audition, computer vision, statistical operations, signal processing [25].
ML INTERFACE		
API		
Pylearn2	Python	It has a capability of wrapping other libraries like,Scikit-learn. It supports a dataset interface for images, vector, video, etc. Also, it provides cross-platform serialization of trained models [26].
LIBSVM (A Library for Support Vector Machines)	C	It supports a simple interface in which the users can link their individual programs with it easily. It includes: various SVM formulations, probability estimates, automated model selection. It comprises of certain data mining set up too such as LIONsolver,RapidMiner [27].
SysWEKA	Java	It extends the features of WEKA. Furthermore, it supports the software interface which is used by advanced applications for the management of resources on cloud setup [28].
CLI		
Waffle	C++	It provides cross-platform command-line techniques. It supports data transformation, clustering, classification, data evaluation in reduced dimensionality, model training and visualization. Moreover, in order to accomplish a specific task, it includes a 'Wizard' tool which directs the user across a sequence of methods to build a command for this task [29].
WEKA	Java	WEKA as a CLI, supports an environment according to a particular domain. It provides tools and techniques for data handling, data visualization, cross-validation, database linkage, comparison of rule sets [30].
GUI		
KNIME (Konstanz Information Miner)	Java	It provides an integrated data pipelining model that combines different modules for ML and data mining. It consists ofin-built modular workflow method that allowsscalability by means of complex data handling techniques, allows user to construct data flows visually, parallelization on multi-core machines, blending of data and tools according to the problem statement [31].
RapidMiner	Java	It is an extension for OpenML, which is an open platform for ML. It provides ML and data mining techniques, such as ETL (Extract, Transform, Load), tools for data pre-processing, data visualization, statistical modeling and predictive analytics [32].
Orange	Python, C++	It is a ML and data mining collection of algorithms for data analysis through visual programming. It makes easier the construction of data analysis workflow and the development of different data mining methods from existing modules [33].
ML LOCAL AND REMOTE TOOLS		
ML LOCAL		
GoLearn	Go	It employs KNN (k-nearest neighbours) classification and regression, along with certain essential CSV parsing [34].
Shogun Library	C++	It consists of a stand-alone CLI.It is developed for large scale training of models for a wide variety of learning and feature settings [35].
ML REMOTE		
Apache Mahout for Hadoop	Java	It provides clustering, collaborative filtering, classification, linear algebra operations, repeated item set timing, and utilities to speed up the vector and sparse matrix calculations [36].
AWS Machine Learning	Java, C++	It trains and tests automatically a bunch of sophisticated models that are set with various parameters. It supports splitting of dataset, input normalization, and evaluation of model [37].
Microsoft Distributed ML Toolkit	C++	It supports numerous programming interfaces and is more competent in Big Data research. It includes two distributed ML algorithms that are used to make the machines learn the

(DMTK)		largest and fastest topic model and the biggest word-embedding model around the globe [38].
Microsoft Azure ML	Java	It is a strong cloud-based tool which is used in analytics that allows predictive management [39].
MLib for Spark	Usable in Java, Scala, Python, and R.	It consists of good set of ML algorithms that influence iteration and produces improved results. It supports feature transformation, development of ML pipeline, hyper-parameter tuning and model evaluation [40].

VI. CONCLUSION

Machine learning is a complicated field and the graph is being rising at an elevated speed as we are heading forward and becoming stronger technologically.As ML algorithm is quite difficult to write from scratch, Machine learning toolkit provides a tremendous amount of resources that can be used according to the problem statement to solve any challenge. In this paper, we have illustrated many tools that can be used for applying ML techniques. The best toolkit is selected on the basis of skills, background and use-case of a researcher. Also, the type of project and available resources play an important role in the selection of a tool. Therefore, when a project is started, it is required to spend a certain amount of time to assess existing toolkits so as to be confident enough that the chosen toolkit is best for the situation.

REFERENCES

[1] https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article
[2] https://dzone.com/articles/5-open-source-machine-learning-frameworks-and-tool
[3] https://www.forbes.com/sites/ciocentral/2018/02/28/gartner-magic-quadrant-whos-winning-in-the-data-machine-learning-space/
[4] J. V. N. Lakshmi and A. Sheshasaayee, "A Big Data Analytical Approach for Analyzing Temperature Dataset using Machine Learning Techniques," *Int. J. Sci. Res. Comput. Sci. Eng.*, vol. 5, no. 3, pp. 92–97, 2017.
[5] C. E. Sapp, "Preparing and Architecting for Machine Learning," 2017.
[6] https://www.gartner.com/it-glossary/big-data/
[7] https://www.quora.com/How-are-big-data-and-machine-learning-related
[8] Rakesh. S.Shirsath, VaibhavA.Desale, Amol. D.Potgantwar, "*Big Data Analytical Architecture for Real-Time Applications*", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.4, pp.1-8, 2017
[9]https://www.forbes.com/sites/ciocentral/2018/02/28/gartner-magic-quadrant-whos-winning-in-the-data-machine-learning-space/#3995d9407dab
[10] https://www.sas.com/en_us/insights/analytics/machine-learning.html
[11] https://towardsdatascience.com/gui-fying-the-machine-learning-workflow-towards-rapid-discovery-of-viable-pipelines-cab2552c909f
[12] https://machinelearningmastery.com/machine-learning-tools/
[13] https://knowm.org/machine-learning-tools-an-overview/
[14] https://blogs.opentext.com/choosing-the-right-programming-language-for-machine-learning-algorithms-with-apache-spark/amp/
[15] https://medium.com/@UdacityINDIA/machine-learning-programming-languages-why-is-the-best-and-why-56f9f370cb99
[16] https://www.analyticsindiamag.com/machine-learning-framework-10-need-know/
[17] https://searchenterpriseai.techtarget.com/feature/How-to-make-a-wise-machine-learning-platforms-comparison
[18] V. Vinothina, "MACHINE LEARNING TOOLS-AN OVERVIEW," in International Conference on Recent Trends in Engineering Science, Humanities and Management, 2017, pp. 629–637.
[19] https://www.oreilly.com/ideas/square-off-machine-learning-libraries
[20] https://machinelearningmastery.com/tour-weka-machine-learning-workbench/
[21] https://bookdown.org/rdpeng/rprogdatascience/history-and-overview-of-r.html
[22] https://en.wikipedia.org/wiki/SciPy
[23] *https://github.com/scikit-learn/scikit-learn*
[24] https://github.com/EdwardRaff/JSAT
[25] http://accord-framework.net/intro.html
[26] Pylearn2 Documentation Release dev, LISA lab, University of Montreal, 2015.
[27] *https://www.csie.ntu.edu.tw/~cjlin/libsvm/*
[28] Thomas A. Henzinger, Anmol V. Singh, Vasu Singh, Thomas Wies,DamienZufferey, "Static Scheduling in clouds"
[29] Mike Gashler, "Waffles: A Machine Learning Toolkit", Journal of Machine Learning Research, 12 (2011), 2383-2387.
[30] G.Holmes, A.Donkin, I.H Witten, "WEKA: a machine learning workbench", Proceedings of Second Australian and New Zealand conferences on Intelligent Information System, 1994.
[31] https://www.predictiveanalyticstoday.com/knime/
[32] https://rapidminer.com/products/studio/feature-list/
[33] https://orange.biolab.si/#Orange-Features
[34] http://126kr.com/article/yucgkiovd
[35] S¨orenSonnenburg et.al, "The SHOGUN Machine Learning Toolbox", Journal of Machine Learning Research 11 (2010) , 1799-1802.
[36] https://mahout.apache.org/docs/latest/index.html
[37] http://cloudacademy.com/blog/aws-machine-learning/
[38] https://www.microsoft.com/en-us/research/blog/microsoft-open-sources-distributed-machine-learning-toolkit-for-more-efficient-big-data-research/
[39] https://www.predictiveanalyticstoday.com/microsoft-azure-machine-learning/
[40] https://spark.apache.org/mllib/

