IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 24-29

Analogous Investigation of Most Commonly Used Machine Learning Tools

Aruna Pavate[1], Divya Kumawat[2], Suvarna Pansambal[3], Apeksha Waghmare[4], Anita Chaudhari[5]

[1, 2,3,4](Computer Department, Atharva college of Engineering/Mumbai,India)
[5](InformationTechnology,St. John College of Engineering / Mumbai, India)

Abstract: *Machine learning is the field most commonly used in many Industries including academics, gaming, designing expert system, and government institutions, Image Recognition, Medical Diagnosis, Financial Services, Marketing & Sales, Transportation Services, Other Biometrics, Safety and Security etc. There are plenty of tools supports to design and experiment soIt is necessary to know the details of the tools available and how we can utilize these tools. This paper provides a general idea of the most commonly used machine learning tools from a practitioner's perspective. Essential services available with these tools are described as well as a comparative examination of these tools represented.*

I. Introduction

Machine learning (ML) is a subsection of an Artificial Intelligence system that allows to access to the right data, machine can learn by themselves without being explicitly being programmed to solve a specific problem [1]. What is the best tool or language for machine learning? The tools used for machine learning depends on the own requirements and preferences. By taking advantage of statistical and mathematical tools, ML system extracts knowledge as well as capable of executing independently. Statistical and machine learning tools provides different functionality to represent, analyze and model the data like feature selection, regression, dimensionality reduction methods that helps to identify features that impact model. There are tools that typically avoid the programming aspect and offer user-friendly GUI (Graphical User Interface) so that anyone with negligible facts of algorithms can simply practice them to build high quality machine learning models. Now a day's industries are using ML systems to improve business decisions, improve productivity, forecast weather, and detect disease, sentiment analysis and many more. With the exponential progress of technology, it is necessary to use better tools to understand the current data as well as that will be generated in future.

Understanding Machine Learning

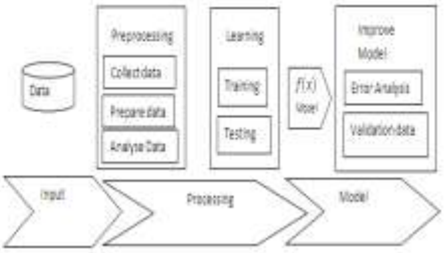

Figure1: Steps in developing a machine learning application

Machine Learning Process:

1. Preprocessing:

Preprocessing include collection of data, preparation of data and analysis of data. Data samples could be collected by scraping a web and extracting data, or get the information from an RSS feed or an API or can be gathered from any device or using survey or from public repositories to save your time and efforts etc. There are lots of options available to collect data. After collecting data, it is necessary to make sure that the data is in useable format. Here sometimes need to do algorithm-specific formatting as some algorithm need features in a special format or some algorithm deals with target variables [2].

2. Learning

Learning include training and testing the data. These are the two steps where the machine learning takes place. In these two steps, the knowledge is extracted by feeding the clean data to the system. In case of unsupervised learning there is no training step because such learning doesn’t have target values. In such type of learning the different types of metrics used to evaluate the success. Testing phase included to evaluate the performance of the system. In case of supervised learning known values are used to evaluate the algorithm.

3. Improve Model

The most valued part of machine learning is predictive modeling. This stage includes development of model that are trained on collected data and make predictions on new data. The performance of the model improved by considering

3.1 Improvement in data

Data can be improved by getting more quality data, inventing more data, cleaning previous samples, resampling data, reorganizing problem, rescaling data, transforming data, selecting feature, reducing feature and projecting data into essential dimensional space[3].

3.2 Improve performance with algorithms

To improve the model, it is necessary to identify the algorithm and the performance of the algorithm. To evaluate the performance of the system, use the best metric captures the requirement of the problem, resample the data, investigate on how to configure each algorithm well [3].

3.3 Improving model performance with algorithm tuning

Model performance can be improved with considering different tuning tactics like visualization and diagnostics, evaluating the performance of standard parameters, using random search algorithm, tuning using direct search procedure or stochastic optimization [3].

3.4 Improve the performance with Hybrid systems

The next area for model improvement is to combine the output of multiple well- performed models. The performance of machine learning system improved by combining the predictions from multiple models directly or can combine models trained on different data representations [3].

Challenges:

The following some of the issues [4][6] are step-up while developing machine learning models

- Which algorithm is suitable for which types of problems and representations?
- What algorithms available to learn general target functions from training samples?
- Are they considered samples being enough for training data?
- What setting is enough to meet the desired hypotheses function?
- What is the unsurpassed approach for taking a useful next training experience?
- How does the planed strategy help to change the complexity of the learning?
- How to tackle with Multi-language smell problem(across the language barrier)?
- What is the finest method to decrease the learning task to one or more function approximation problems?
- What are the limits of current machine learning technology?
- What if an algorithm’s diagnosis is wrong?

Sebastian Schelter et al discussed other different challenges on machine learning model management like over conceptual, data- processing, engineering ascending in the management of the equivalent models and given direction for future research. [5]

II. Machine Learning Toolkits

As machine learning extending into more areas, the numbers of tools and frameworks available to developers and data scientists have multiplied almost all the tools. Many companies like Microsoft, AWS, Google, IBM offer machine learning APIs via their individual cloud platforms, helping developers to build models by conceptualizing some of the difficulty of their algorithms. In this section some of the most commonly used machine learning tools discussed

1.Shogun

Shogun is the primogenital tool for machine learning. This tool was developed by Sonnenberg and Gunnar Raetsch in 1999 using C++ language. This tool offers algorithms and data structures for all the models of machine [9][10][11].

2.Apache Mahout

Apache Mahout is an open source tool which is entirely free. Apache Mahout is developed on Apache Hadoop platform. Mahout offers tools to find some specific patterns from the large data sets and the data is warehoused on Hadoop Distributed File system (HDFS)[].

3. Scikit-Learn

Scikit-Learn is based on most popular language which commonly used for maths, science and statistics. This tool helps in analysing and mining the data. This tool is an open source and built on matplotlib, SciPy, NumPy. Scikit-Learn supports to design application Spam detection, Image recognition, Predicting a continuous-valued attribute associated with an object.Feature extraction and normalization. Comparing, validating and choosing parameters and models. [7][8].

4. Apache Spark MLlib

Apache Spark MLlib supports scalable machine learning library along with distributed machine learning framework. MLlib includes dimensionality, decision trees, regressing, clustering, collaborative filtering, higher level pipeline APIs. Apache spark is an open source project[13].

5. TensorFlow

TensorFlow is an project of Google's which include machine learning library. Now it is an open system as it replaced google's previous technology. TensorFlow is the most popular library, helps us It is now an open source framework as earlier it was developed to build machine learning into its own system. Tensor flow is the most flexible library as it lets you write your own libraries. It uses C++ and Python languages [15][16].

6. Accord.NET

The Accord.NET Framework written in c# helps us to combine audio and image processing libraries. This framework helps to build computer vision, signal processing, computer audition, statistics applications [12]

7. Keras

Kera is one of the competitor for building most types of Machine learning and AI applications. Keras run on top of Theano. Keras support both CPU's and GPU's for developing most types of applications [17].

8. Apache Mahout(TM):

Apache Mahout(TM) is a distributed linear algebra framework and mathematically expressive Scala DSL designed to let mathematicians, statisticians, and data scientists quickly implement their own algorithm. [21].

9. Spark

Spark run on standalone machine on different platforms such as Hadoop YARN, Kubernetes, EC2. Sparks library support high quality algorithms that gives better results than one-pass approximations used on MapReduce [13].

10. H2O

H2O supports distributed in-memory machine learning platform. It is an open source. H2O's supports the most widely used machine learning and statistical algorithms [22].

11. cloudera Oryx

Oryx 2 architecture is built on Apache Spark and Apache Kafka. It supports building end to end applications[23].

12. Weka :

Weka is an open source which provides collection of machine leaning algorithm. Weka possible us to access other data mining packages[18]

13. ConvNetJS

ConvNetJS is developed using Java script library basically for training Deep Neural netwok models. ConvNet entirely work on browser without installation, no compilers , no installations , no GPU[19].

14. Caffe

Caffe supports based on deep learning with speed expression and modularity[20]

III. Evaluation of machine Learning Toolkit

Table no 1 : Comparative Analysis of Most commonly used Machine Learning Tools

Sr. No.	Tool	Features	Initial Release	stable Release	Operating System Supported	Written In	Algorithms Supported	
1	scikit-learn v0.19.1	classification, regression clustering	June 2007; 10 years ago	0.19.1 / 22 October 2017; 4 months ago	Linux, macOS, Windows	Python, Cython, C, C++	SVM, nearest neighbors, random forest, SVR, ridge regression, Lasso, k-Means, spectral clustering, mean-shift, PCA, feature selection, non-negative matrix factorization., grid search, cross validation, metrics.	open Source
2	Shogun	Dimensionality reduction algorithms, classification, regression clustering .	1999,	6.0.0 / April 25, 2017	Linux/Unix, MacOS and Windows	Python, Octave, R, Java/Scala, Lua, C#, Ruby, etc	PCA, Kernel PCA, SGD-QN, Vowpal Wabbit, k-means and GMM Kernel Ridge Regression, Support Vector Regression Hidden Markov Models K-Nearest Neighbors	open-source
3	Accord.net	classification, regression clustering, Distributions, Hypothesis Tests, Kernel methods , Imaging, Audio & Signal, vision	2008	Updated Feb 20, 2018	Cross-platform	C#	Support Vector Machines, Decision Trees, Naive Bayesian models, K-means, RANSAC Linear and Logistic regression, Hidden Markov Models, (Hidden) Conditional Random Fields, Principal Component Analysis, Partial Least Squares, Discriminate Analysis, Kernel methods	Open souce
4	Apache Mahout	collaborative filtering, clustering and classification.		0.13.0 / 17 April 2017	Cross-platform	Java, Scala	Logistic Regression - trained via SGD ,Naive Bayes / Complementary Naive Bayes Hidden Markov Models Canopy Clustering k-Means Clustering Fuzzy k-Means Streaming k-Means Spectral Clustering Singular Value Decomposition	Open Source
5	H2O	classification regression Clustering Dimensionality reduction Structured prediction Anomaly detection	2011; 7 years ago	Tutte (3.10.2.2) / January 12, 2017; 13 months ago	Linux, macOS, and Microsoft Windows.	Java, Python, and R.	Cox Proportional Hazards (CoxPH) Deep Learning (Neural Networks) Distributed Random Forest (DRF) Generalized Linear Model (GLM) Gradient Boosting Machine (GBM) Naïve Bayes Classifier Stacked Ensembles XGBoost Aggregator Generalized Low Rank Models (GLRM) K-Means Clustering Principal Component Analysis (PCA)Word2vec	Open source
6	Cloudera Oryx	collaborative filtering, classification, regression and clustering.	2014	2018	Apache Spark and Apache Kafka	Java 8	ALS, random decision forests, k-means	Open source
7	Apache Spark MLlib	classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as underlying optimization primitives	May 26, 2014	v2.3.1 / June 8, 2018	Microsoft Windows, macOS, Linux	Scala, Java, Python, R	SVMs, logistic regression, linear regression ,naive Bayes decision trees ,Random Forests and Gradient-Boosted Trees Clustering k-means Dimensionality reduction singular value decomposition (SVD) principal component analysis (PCA) BFGS (L-BFGS)	Open source

8	Tensor Flow	classification regression Clustering Dimensionality reduction	November 9, 2015	1.10.0 / August 8, 2018	Linux, macOS, Windows, Android, website	Python, C++, CUDA	K-means clustering Random Forests Support Vector Machines Gaussian Mixture Model clustering Linear/logistic regression,Deep Neural Networks, Recurrent Neural Networks	Open Source
9	Keras	prediction, feature extraction, and fine-tuning.	27 March 2015; 3 years ago	2.2.0 / 7 June 2018; 2	Python	Python	ANN,RNN,DNN,CNN	Open source Cross platform
10	cloudera Oryx	Collaborative filtering, classification, regression and clustering.	Options 07-08-2014	10-02-2017	Ubuntu	Python ,Perl,python-psycopg2	ALS, random decision forests, k-means	Open source
11	Weka :	Data preparation, classification, regression, clustering,association rules mining, visualization.		3.8.3 (stable) / September 4, 2018; 1 day ago	Windows, OS X, Linux	Java	Naive Bayes: bayes.NaiveBayes. Decision Tree (specifically the C4.5 variety): trees.J48. k-Nearest Neighbors (also called KNN: lazy.IBk. Support Vector Machines Neural Network	Open source
12	ConvNetJS	classification, regression, clustering	Sep 1, 2014		Windows, OS X, Linux	Javascript,	Common Neural Network modules (fully connected layers, non-linearities) Classification (SVM/Softmax) and Regression (L2) cost functions Convolutional Networks Reinforcement Learning module, Deep Q Learning.	Open source
13	Caffe	Image classification and image segmentation		1.0[1] / 18 April 2017; 16 months ago	Linux, macOS, Windows	C++	CNN, RCNN, LSTM and fully connected neural network designs	Open Source

IV. Conclusion

This paper presents t the study of the most commonly used machine learning tools & the features supported by such algorithms. The most used machine learning tools Shogun ,Apache Mahout ,Scikit-Learn ,Apache Spark MLlib, TensorFlow, Accord.NET, Kera, spark,weka, tensorflow etc have analyzed & performed a comparative analysis of different machine learning tools.

References

[1]. S. Marsland, Machine learning: an algorithmic perspective. CRC press, 2015

[2]. Peter Harrington "Machine Learning In Action", DreamTech Press

[3]. https://machinelearningmastery.com/machine-learning-performance-improvement-cheat-sheet/

[4]. Tom M. Mitchell "Machine Learning" McGraw Hill

[5]. Sebastian Schelter, Felix Biessmann, Tim Januschowski, David Salinas, Stephan Seufert, Gyuri

[6]. Szarvas," On Challenges in Machine Learning Model Management", Bulletin of the IEEE ComputerSociety Technical Committee on Data Engineering 2018

[7]. https://www.netguru.com/blog/7-challenges-for-machine-learning-projects

[8]. Fabian Pedregosa; GaëlVaroquaux; Alexandre Gramfort; Vincent Michel; Bertrand Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vincent Dubourg; Jake Vanderplas; Alexandre Passos; David Cournapeau; Matthieu Perrot; Édouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830.

[9]. http://scikit-learn.org/stable/

[10]. http://www.shogun-toolbox.org

[11]. S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. De Bona, A. Binder, C. Gehl and V. Franc: The SHOGUN Machine Learning Toolbox, Journal of Machine Learning Research, 11:1799−1802, June 11, 2010.

[12]. M. Gashler. Waffles: A Machine Learning Toolkit. Journal of Machine Learning Research, 12 (July):2383–2387, 2011.

[13]. http://accord-framework.net/

[14]. http://spark.apache.org/docs/1.2.1/

[15]. http://oryx.io/docs/endusers.html

[16]. https://www.tensorflow.org/

[17]. https://terrytangyuan.github.io/2016/08/06/tensorflow-not-just-deep-learning/

[18]. https://keras.io/applications
[19]. http://wekaclassalgos.sourceforge.net/
[20]. https://cs.stanford.edu/people/karpathy/convnetjs/intro.html
[21]. http://caffe.berkeleyvision.org/
[22]. https://mahout.apache.org/
[23]. https://www.h2o.ai/
[24]. http://oryx.io/