Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 10, No. 1, October : 2023

AN ANALYTICAL RESEARCH ON COMMAND LINE INTERFACE FOR DATA PREPROCESSING IN MACHINE LEARNING

Pranav Kumar Assistant Professor, Government Engineering College, Arwal, Bihar, India Email id-pranavkmr79@gmail.com
Md. Talib Ahmad Assistant Professor, Katihar Engineering College, Katihar, Bihar, India Email id-talib@keck.ac.in

Abstract-
Machine learning is a way for getting a machine to pick up knowledge and act without being explicitly programmed. Affective computing, Adaptive websites, Bioinformatics, Brain-machine interfaces, Cheminformatics, Classifying DNA sequences, Computational anatomy, and Computer vision, including object identification are just a few of the fields where it is becoming widely used due to its many advantages. Since so many users rely heavily on the cloud, it is becoming important to deploy machine learning methods there to make the most efficient use of cloud resources and services. To assist the machine learning approach, several algorithms and tools have been developed. In this study, the author examines the most widely used machine learning technologies.

Keywords—
Command Line Interface, Data Preprocessing, Machine Learning, Classifying DNA Sequences.

INTRODUCTION

Programming systems for automated learning and improvement over time is called machine learning. The machine can make wise judgements based on knowledge and experience by using machine learning techniques. Here, learning entails comprehending the intricate patterns that the system goes through. The set of all feasible decisions for the given inputs cannot be fully described due to its complexity. On the basis of statistics, probability theory, optimisation, reinforcement learning, and control theory, many machine learning algorithms are created to find knowledge from particular data and experience in order to solve this problem. Language processing, forecasting, pattern recognition, gaming, data mining, expert systems, and robotics all use existing machine learning techniques. The Turing Test to evaluate the performance of the ENIAC computer in 1950, the checkers game programme in 1952, ELIZA, a simulation of a psychotherapist in the 1960s, expert systems, statistical AI in the 1990s, and Bigdata to identify complex patterns from vast amounts of data are all considered milestones in machine learning. Machine learning will be a key component of knowledge discovery in the next years from the vast quantity of data that is now accessible in a variety of application domains. Machine learning techniques are used in the cloud to forecast resource requirements. Machine learning advancements are advantageous to businesses and sectors. Machine learning is used in the design of quadcopters to investigate all potential solutions and fresh concepts for efficient flying and payload. The computer generates novel concepts that designers have never considered when given restrictions like a decent payload. On a project with Airbus, Autodesk is using machine learning to rebuild and reimagine a new aeroplane cabin divider. Machine learning also left a mark in robotics. 54.6 percent of developers choose machine learning over all other options when creating robotics apps, and 24.6% of developers employ it in their papers. Computer learning, imitation learning, self-supervised learning, assistive and medical technologies, and multi-agent learning are the primary machine learning-related disciplines in robotics. Inverse optimal control, Bayesian models, support vector machines, and kinematics are some of the machine learning principles relevant to robotics. IoT (Internet of Things) relies on the machine learning technique of modifying the result and behaviour depending on information or observation. The finest use of IoT and machine learning integration is in

traffic routing, which investigates several routes to the target. IoT and ML algorithms are used by numerous businesses and the government for things like traffic, health care, and more. Such an application is highly appreciated given the billions of data collected from the devices. Since the majority of businesses now use the cloud, machine learning also plays a significant role in the forecasting of resource supply.

RELATED WORK

The command line is considered a relic of computing past. The command line, on the other hand, is essential as a developer tool. Command Line Interfaces are frequently used to manage local processes and source control (CLIs). Another type of CLI has gained popularity in recent years, and it goes beyond the local system. For developers, the command line has evolved into a powerful tool for interacting with cloud services. CLIs are available for cloud computing services, continuous integration products, and some APIs. Pre-processors, as the name implies, are programs that process our source code before compilation. Between writing a program and executing it, there are several steps to take. Learning algorithms have a preference for specific types of data, on which they excel. They're also known for making reckless predictions based on un scaled or un standardized data. Algorithms like XG Boost, in particular, require dummy encoded data, but decision trees appear to be unconcerned. Pre-processing, in simple terms, is the alteration of data prior to feeding it to the algorithm. The scikit-learn package in Python comes with pre-built functionality called sk learn Preprocessing. There are various data pre-processing steps such as:

Importing the libraries → importing the data set →
checking out the missing values → encoding the categorical
values → splitting the data → feature scaling

Previously, the category variables were encoded using one of the encoding techniques known as Label Encoding. Each category is given a value ranging from 1 to N(where N is the number of categories for the feature). One major flaw with this approach is that there is no relationship or order between these classes, though the algorithm may be considered such. Various techniques such as Robust scaling and Maximum Absolute Scaling are used for feature scaling pre-processing steps.

PURPOSE FOR DATA PREPROCESSING IN MACHINE LEARNING

Several researchers made novel, creative suggestions utilising machine learning methods. Machine learning has recently gained significant traction in sectors such as business, healthcare, and transportation. Knowing about the different machine learning algorithms and tools used in those fields is more important as machine learning approaches are employed in fields like robotics, IoT, and cloud computing. Support Vector Machine (SVM), Neural Networks (NN), and Linear Regression (LR) were three machine learning approaches used by Akindele et al. to construct and test cloud client prediction models for the TPCW benchmark web application. The client was given a more reliable scaling decision option as a result. According to the study, SVV offered the best prediction model. To recognise the presence of a human in space using data from various home automation devices, a system is built utilising supervised learning techniques. The model asserted the existence of somebody being at home by using feature extraction and a labelling phase that used heuristics. The outcomes demonstrated potential directions for raising system accuracy. Machine learning techniques are being used even in the healthcare industry. In study, the authors employed machine learning techniques to calculate the bare minimum resources required to guarantee that there are as few bottlenecks in the patient flow as possible, which not only increases patient happiness but also helps hospitals financially. Healthcare institutions have been collecting more data recently, which has created new potential to use machine learning approaches to solve this issue. This effort improved the accuracy of resource utilisation predictions. Most frequently, visitors have trouble identifying the most pertinent online sites

or information based on their interests. Research groups concur that machine learning techniques must be used on the cloud.

ISSUE WITH PROBLEM

(i) Existing System and its disadvantages- The processing and input of various sorts of inputs, as well as the model and its parameters, all affect how effectively a machine learning model performs. Preprocessing categorical data becomes crucial because the majority of machine learning models only take numerical variables. 'Strings' or 'categories' are two typical ways to describe categorical variables, which have a finite number of possible values. Robust Scaling's primary drawback is that it ignores the median and only focuses on the regions with the most data. One of the drawbacks of Maximum Absolute Scalar is the presence of extremely large outliers. The Multi-Season Holt-Winter model is used in time series analysis by the current method for forecasting crime rates in Chicago. The error rate for this model is 2.11%, however. Nevertheless, this model can be fully optimised due to its 91% accuracy rate. Two types of categorical data exist:

- Ordinal Data: There is a natural order to the categories.
- Nominal Data: There is no intrinsic order to the categories. Ordinal data should maintain during encoding the information on the presentation of the categories. The inclusion or exclusion of a characteristic must be taken into account while encoding nominal data. In such a circumstance, order does not exist. Categorical variables should be transformed to integers in order for the model to comprehend and derive valuable information. One of the preprocessing procedures in data preparation is encoding the categorical variable. Prior until now, label encoding was employed as an encoding method. The label encoding procedure may be used to encode ordinal data since it is straightforward and takes order into account. The label encoding should also represent the sequence. Label encoding, however, has a drawback in that it takes into account column hierarchy, which may be deceptive to nominal characteristics contained in the dataset.

(ii) Proposed System and its advantages- The goal of this research is to suggest a system that addresses the previously highlighted shortcomings of the current system. In this study, the one-hot encoding approach was used to encode the categorical variable. By transforming categorical input data into machine and deep learning methods, one-hot encoding raises the accuracy of model predictions and classification. For feature scaling, normalization and standardization were utilized. A scaling method called normalization moves and rescales data such that they fall between 0 and 1. The fundamental benefit of normalization is that it significantly increases model accuracy. Standardization, which centers data around the mean with a unit standard deviation, is another scaling strategy. As a result, the attribute's mean is reduced to zero, and the distribution that results has a unit standard deviation. Standardization has the benefit of ensuring that each characteristic has the same effect on the distance measure. Additionally, a prophet model has been used to forecast Chicago crime rates. Face book Prophet is an additive regression-based time series forecasting model. Both data from numerous seasons and data with strong seasonal impacts can be used successfully. The Prophet is unaffected by outliers, missing data, and quick changes in time series.

TOOLS FOR MACHINE LEARNING

The perspectives of libraries, graphical user interfaces (GUI), command line interfaces (CLI), application programming interfaces (API), local machine learning (LML), and remote machine learning (RML) tools, may be used to discuss machine learning tools.

(i) Command Line Interface (ML-CLI)- With command line programmes, command line parameterization, and an emphasis on input and output, ML-CLI offers a command line interface. These are some of its features: Machine learning initiatives are accessible to people who are not programmers. The capacity to offer programming options for certain machine learning project

subtasks. The capacity to save command-line parameters. Examples of machine learning tools for a command line interface are waffles and WEKA. It is a group of cross-platform command-line tools for machine learning researchers. It automates all of the C++'s functionality and scripting. Waffles supports the following operations: classification, clustering, data transformations, reduced dimensionality evaluation of the data, model training, and visualisation. It offers a "Wizard" tool that leads the user through a number of forms to build a command that will carry out the desired action.

(ii) Graphical User Interface (ML-GUI)- Tools for machine learning include a graphical user interface with windows, point-and-click functionality, and a visualization-focused approach. A graphical user interface has the following advantages: Enables less technical users to collaborate with machine learning. Put an emphasis on the process and how to make the most of machine learning tools. An interface that forces the user to follow a structured procedure; a stronger emphasis on information shown graphically, such as visualisation. Examples of machine learning tools featuring a graphical user interface are KNIME, RapidMiner, and Orange. An open source platform for data analytics, reporting, and integration is called KNIME (Konstanz Information Miner). Through the use of its modular data pipelining architecture, it integrates numerous machine learning and data mining components. The construction of nodes for data preparation (ETL: Extraction, Transformation, Loading), modelling, data analysis, and visualisation is made possible through a graphical user interface. KNIME has been employed in financial data analysis, corporate intelligence, CRM customer data analysis, and medicinal research.

(iii) Machine Learning Library (ML-L)- In order to construct a project, the ML library includes tools like configuration data, documentation, help data, message templates, pre-written code and subroutines, classes, values, or type specifications. It could give users modelling methods to complete their tasks. Scikit-Learn, JSAT, and ACCORD Framework are three machine learning libraries. The Python programming language has an open source machine language library called Scikit. Support vector machines, random forests, gradient boosting, k-means, and DBSCAN are among the classification, regression, and clustering techniques included in it. The Java library for machine learning is called JSAT (Java Statistical Analysis Tool). This library provides numerous algorithms, including data transformations, predictive, tree-based, vector quantization-based, and meta algorithms. As a result, it is suitable for research and particular purposes. In accordance.C#-based Net platform is used. It provides a comprehensive platform for creating professional-grade signal processing, statistics, computer vision, and computer audition applications, even for commercial usage.

(iv) Application Programming Interface (ML-API)- We have the freedom to choose which components to utilise and exactly how to use them within our own programmes thanks to ML-API. Application programming interfaces include the following features: Incorporating machine learning into other software projects. developing our own machine learning tools by developing our own workflows, automating tasks, and integrating fresh approaches into libraries and current techniques. Pylearn2, Deeplearning4j, and LIBSVM are a few examples of application programming interfaces for Python, Java, and C, respectively. When necessary, researchers can add features. a toolkit for machine learning that makes scientific experimentation simple. There should be reference implementations of all models and techniques used in the LISA lab in Pylearn2. When it makes sense, Pylearn2 may wrap other libraries, such as Scikit-Learn. Pylearn2 wants to provide researchers a lot of flexibility and enable them to perform practically anything. Dataset interface is offered for vector, pictures, video, etc. It enables the reuse of Pylearn2's component parts. Additionally, it facilitates serialising learnt models across several platforms. The distributed, open-source deep-learning library for Java and Scala is called Deeplearning4j. DL4J is integrated with Hadoop and Spark and made to be utilised on distributed GPUs and CPUs in corporate contexts.

(v) Local and Remote Tools (ML-LT and ML-RT)- Whether a machine learning tool is local or remote is one parameter to compare. In contrast to remote tools, which can only be used on third party

servers, local tools may be downloaded, installed, and used locally. This distinction may be helpful in understanding and selecting a machine learning technology. These are the characteristics of local tools: Control over run setup and parameterization. Fit for in-memory data and algorithms. Integrate into our own systems to satisfy your demands. Shogun Library for C++ and GoLearn for Go are just a few of examples of local tools. SHOGUN is for unified large-scale learning for a variety of feature kinds and educational environments. It provides a broad selection of machine learning models, including linear discriminant analysis, hidden Markov models, support vector machines, and more. The C++-based SHOGUN programme offers interfaces for MATLAB, TM R, Octave, Python, and a standalone command line interface.

CONCLUSION

Giving a system training to respond to diverse scenarios based on data patterns is what machine learning is all about. It includes a variety of data preparation techniques, including classification, regression, clustering, visualization, and prediction, among others. Our system's architecture performs data cleansing, transformation, and reduction during data preparation. Unprocessed datasets are accepted by our application and are later cleaned up. When all of the preparation is finished, the users are shown the cleaned data. Time is saved with this procedure since hand cleaning is not necessary. After purification, a machine learning model that will produce accurate plots is available for the user to pick. Users that need to organize sizable datasets and visualize the analysis of pre-processed data may find this beneficial. The accuracy and comparison of machine learning algorithms will be possible in the future, all through a user-friendly interface. An research into how Face book Prophets functions revealed that the programmed accurately anticipated and correctly projected the pattern of crimes occurring in Chicago for the following two years. This article discusses the tools that are used to put machine learning approaches into practice. Academics, students, and researchers interested in machine learning techniques will find this beneficial. To get the best results, a specialized tool has to be chosen and used based on the researcher's use case.

REFERENCES

[1] Cristian Felix, Anshul VikramPandey, and Enrico Bertini, "textile: AnInteractive Visualization Tool for Seamless Exploratory Analysis of Structured Data and Unstructured Text", IEEE-2018.
[2] Data, Huawen Liu, Xuelong Li, Jiuyong Li, and Hicham Zhang,"Efficient Outlier Detection for High-Dimensional", IEEE- 2019.
[3] M. Bostock, V. Ogievetsky, and J. Heer, "Data Driven documents," IEEE- 2011.
[4] F. Beck, S. Koch, and D. Weiskopf, "Visual Analysis and Dissemination of Scientific Literature Collections withSurVis", IEEE-2016.
[5] Parke Godfrey, JarekGryz and Piotr Lasek,"Interactive visualization of large datasets", IEEE-2016.
[6] Dileep Kumar Ashley and Raju Hadler, "Data Cleaning: An Abstraction-based approach",IEEE-2015.
[7] Mehmet Adil Yalçın;Niklas Almqvist; Benjamin B. Bederson, "Keshif: Rapid andExpressive Tabular Data Exploration forNovices",IEEE-2018.
[8] Elham Hormozi, Hadi Hormozy , Mohammad Kazem Akbari and Morteza Sargolzai Javan, “Using of Machine Learning into Cloud Environment (A Survey), Seventh International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, DOI 10.1109/3PGCIC.2012.69,2012, 363-368.
[9] David Burford, “Cloud computing: a brief introduction,”,LAD NTERPRIZES, February 20, 2010.
[10] Gurmeet Singh, “Scope of machine learning in cloud computing,”, Under the guidance of Dr. Diganta Goswami, 8-11-2010.

[11] A. Serengul guven smith1 and ann blandford," MLTutor: An Application of Machine Learning Algorithms for an Adaptive Web-based Information System", International Journal of Artificial Intelligence in Education, 2002.
[12] Akindele A. Bankole and Samuel A. Ajila," Predicting cloud resource provisioning using machine learning techniques", 2013 26th IEEE Canadian Conference on Electrical and Computer Engineering, 2013.
[13] https://redshift.autodesk.com/machine-learning/
[14] http://techemergence.com/machine-learning-in-robotics/
[15] http://www.cs.wustl.edu/~jain/cse570-15/ftp/iot_ml/
[16] Rui Madeira and LuisNunes, "A machine learning approach for indirect human presence detection using IOT devices", 2016 Eleventh International Conference on Digital Information Management (ICDIM), 2016.
[17] Daniel Vieira and Jaakko Hollmen,"Resource Frequency Prediction in Healthcare: Machine Learning
Approach", IEEE 29th International symposium on Computer-Based Medical Systems(CBMS), 2016.
[18] https://github.com/EdwardRaff/JSAT
[19] Tiwari, Abhishek; Sekhar, Arvind K.T. (October 2007). "Workflow based framework for life science informatics". Computational Biology and Chemistry. 31 (5-6): 305–319.
[20] Jan N. van Rijn, Venkatesh Umaashankar, Simon Fischer, Bernd Bischl, Luis Torgo, Bo Gao, Patrick Winter, Bernd Wiswedel, Michael R. Berthold and Joaquin Vanschoren, "A RapidMiner extension for Open Machine Learning".
[21] Janez Demsar et.al, "Orange: Data Mining Toolbox in Python", Journal of Machine Learning Research 14 (2013), 2349-2353.
[22] Mike Gashler, "Waffles: A Machine Learning Toolkit", Journal of Machine Learning Research, 12 (2011) , 2383-2387.
[23] 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.
[24] Pylearn2 Documentation Release dev, LISA lab, University of Montreal, 2015.