Choosing Best ML Algorithm For Given Data

Final Project - (CS-6100-100 - Advanced Storage, Retrieval, Processing of Big Data)

Nanda Kishore Mandadapu
Western Michigan University
win:[574822478]

Rishitha Reddy Kasireddy
Western Michigan University
win:[012419648]

Sai Jyothi Bhumireddy
Western Michigan University
win:[361648962]

***Abstract*— This work explains how the application made is useful to select the best algorithm for the data. Best algorithm here means the algorithm with the most accuracy. A particular model can be brought up from a bunch of algorithms that have been set up in the program for the given data. This report explains how the application developed here makes choosing the appropriate algorithm for a given data easy.**

I. Problem Statement

The big tech industries all over the world are able to generate huge amounts of data and have plenty of ways to properly organize it and store it. Though there are plenty of problems in organizing, storing, cleaning, and making information out of that data, one of the major problems is choosing an appropriate algorithm to bring the best model out of the particular data, which can generate accurate data with low latency. As there are different kinds of data that suit different kinds of algorithms , speaking about the kind of data, the nature and statistical measures of the data like variances , deviations , outliers , also does matter to select the best algorithm to produce the accurate and low latency model.

II. Introduction

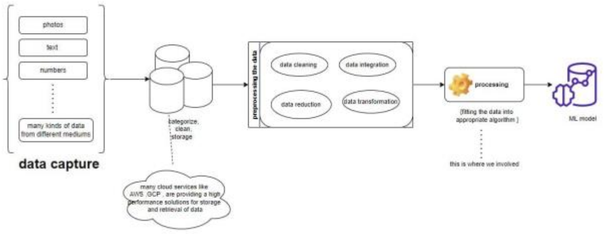
According to the resources every day nearly 2.5 quintillion bytes of data is being generated. Data is collected in numerous ways through our activities on a daily basis. If a person searches for a product online, the same product advertisement happens to be on his/her social account feed like Instagram, Facebook etc. on the same day. How is this happening?!

For instance, if we are using Instagram or Facebook the data is collected from different kinds of sensors from our devices like microphone, camera. Minute details like our scrolling time, text messages, screen time are also being captured. Likewise, Sam’s club by Walmart collects the data that includes a user’s cursor activity time for advertisement and promotion of the products.

The collected data is used in different ways by different industries like the health industry uses the data for predicting diseases and to make appropriate medicines for them, the sales department uses the data for promotions, the weather sector uses data for weather forecasting etc.

The collected data is used to build machine learning models. But, here the question is how are they deciding the appropriate algorithm to build these ML models with the available data? Different Organizations are spending a lot of resources like man power, cloud technologies, Finances for utilization of this data.

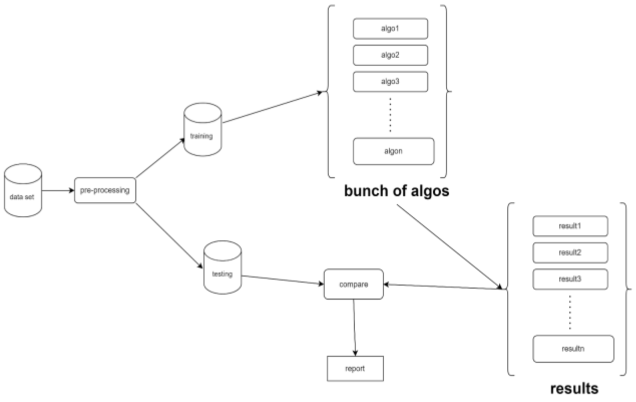
This project focuses on finding an appropriate algorithm that best fits the data. As shown in the below fig(1), an accurate ML model can be produced with the chosen algorithm.


photos
text
numbers
many kinds of data from different mediums
data capture
categorize, clean, storage
many cloud services like AWS ,GCP , are providing a high performance solutions for storage and retrieval of data
preprocessing the data
data cleaning
data integration
data reduction
data transformation
processing
(fitting the data into appropriate algorithm)
this is where we involved
ML model

fig(1) : Big Data Lifecycle

III. Proposed Architecture

For this, the proposed solution is an application called “Test & Trial” abbreviated as TAT. This application produces a performance report of the data with calculated values of various statistical measures like accuracy, sensitivity, specificity, precision and F score, on different classification algorithms. For now the application is limited only to classification algorithms. Assuming the data passed through the application to be already preprocessed as per the requirement, the process is continued as follows.



fig(2) : Proposed Architecture

IV.SELECTING THE PIECE OF DATA

The main focus of TAT is to help the developer choose the best performing algorithm as per the requirement. So, it would be more meaningful to use the useful sample data instead of all the data and hence only a piece of data is selected. This piece of data is used to demonstrate this cancer dataset from kaggle, selecting a certain data in between means to reduce the dimension of data in both rows and columns to reduce the time complexity and increase the efficiency of application. using the NumPy, Pandas we perform the basic cleansing on data i.e. deleting the duplicates , null values and outliers but when considering only a piece of data there will be some constraints i.e. if the selected data set has two clusters and the piece of dataset is taken from one of cluster therefore the results obtained after processing this dataset will be more lenient towards the cluster from which the small data set is considered so it has to be performed well , here “Principal component analysis(PCA)” can be used to deal with reducing the columns. constraints by reducing the dimensions of the dataset.

V . BIG DATA-NESS

1) Business Case Evaluation

The Business Case Evaluation stage requires that a business case be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks. Here, breast cancer prediction is the case in consideration.

2) Data Identification

This stage is dedicated to identifying the datasets required for the analysis of this project and their sources. The data set considered in this project is “Breast Cancer Wisconsin (Diagnostic) Data Set” from Kaggle.

3) Data Acquisition & Filtering

The acquired data from the previous stage is then subjected to automated filtering using “NumPy” and “pandas” packages for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.

4) Data Extraction

In this stage “Principal component analysis(PCA)” is used for extracting disparate data and reducing the data dimensions thereby transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.

5) Data Validation & Cleansing

This stage is dedicated to establishing often complex validation rules and removing any known invalid data. From the original dataset some of the known invalid values like ID and Name of the patients are removed.

6) Data Analysis

In this stage extensive statistical data analysis is done to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables. The excel report is generated which consists of all the calculated values of statistical measures for all the classification algorithms in this project.

7) Data Visualization

In this stage , the Model Comparison function is used to graphically communicate the analysis results for effective interpretation .

8) Utilization of Analysis Results

Using the results from above two stages i.e. “Data Analysis” and “Data Visualization” an appropriate algorithm and suitable statistical measure is chosen which will be the best fit for the dataset.

VI . Data Sets and Technologies Used

Data Sets Used:

To demonstrate this particular application, health care data which has information of patients tested for the breast cancer with the parameters like number of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age to predict the chances of getting the breast cancer has been used.

Technologies Used:

The main language used in developing this project is python. Libraries and packages used here are all pretty much predefined in python. For the frontend, interface styles and templates used are the ones which are already in tkinter which is predefined in python. For the backend, predefined modules and libraries in python such as numpy for cleaning the data, pandas for structuring the data, sklearn for making the models and matplotlib for data visualization are used.

VII. Conclusions & Future Work

Our aim for this project was to develop an application that helps in deciding the most appropriate algorithm to build the machine learning models and we have managed to produce it to the maximum extent. In this application, a set of selected algorithms namely K Neighbors Classifier, Linear SVC, SVC, Random Forest, Decision Tree, Bagging Classifier, AdaBoost Classifier, MLP Classifier, Gaussian NB, SGD Classifier, Quadratic Discriminant Analysis have been tested on the dataset to build a suitable ML model, out of which Random Forest algorithm is believed/chosen to be the most appropriate one.

Furthermore, the application capabilities can be extended so that it can be used for choosing the best algorithm not only in classification analysis as in the current case, but also in the image analysis, time-series analysis, etc depending upon the statistical measure requirement.

References

[1] Choosing Best Algorithm Combinations for Speech Processing Tasks in Machine Learning Using MARF - Serguei A. Mokhov

[2] Selecting Machine Learning Algorithms Using Regression Models- Tri Doan; Jugal Kalita

[3] Selection of relevant features and examples in machine Learning - Avrim L. Bluma**, Pat Langley

[4] https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/

[5] https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

[6] Textbook-Big Data Fundamentals: Concepts, Drivers & Techniques (The Prentice Hall Service Technology Series from Thomas Erl)

WORK ALLOCATION

Below is an area-of-expertise breakdown and estimated effort of our team members. All members learned from each other and expanded their knowledge base outside of their area of comfort.

Group member	Primary area	Estimated effort
Nanda Kishore	Application development	34%
Rishitha Reddy	Data analysis , documentation , UI	33%
Sai Jyothi	Data Acquisition, data pre processing , research	33%