Disease Diagnosis using Machine Learning

Domain Background

Machine Learning (ML), a branch of Artificial Intelligence (AI), learns from the data using various algorithms and is a self-improving process in terms of performance as make adjustments during the learning process [1]. ML has been successfully applied to practically in every domain like robotics, education, travel to health care [2]. In health care domain, the ML approaches are mainly used with the purpose of disease diagnosis [3].

The analysis of the clinical data can lead to timely diagnosis of the disease which will help to start cure for the patient in time as well [3]. Traditional approach of diagnosing disease is generally costly and time consuming. And, the potential of time and cost proficient machine learning biased disease diagnosis approaches are proven by the researchers [4]. ML techniques have not only been able to diagnose the common diseases but are also equally capable of diagnosing the rare diseases [2], [5].

Problem Statement

Although ML offers systematic and sophisticated algorithms of multi-dimensional clinical data, the accuracy of the ML in diagnosing the diseases is still a concern [6]. And the improvement in the performance of ML to diagnose disease is a hot topic in this domain. As different ML approach perform differently for different healthcare dataset, we are also in need to find the way to apply many state of art algorithms to same dataset in reasonable time, so that the search of best ML method can be pursued to diagnose a particular disease.

Solution Statement

The use of libraries like AutoGluon can help to find the best performing ML approach out of many approaches in diagnosing the disease for a given dataset in reasonable time and with optimal lines of code. This will decrease the probability of inaccurate diagnosis, which is a significantly important consideration while dealing with the health of the people.

Datasets and Inputs

For this project, we have chosen a healthcare dataset related to the diabetes. The dataset is Pima Indian Diabetes Dataset which is frequently used to evaluate the performance of developed ML techniques [7]. We downloaded the dataset from [7]. This data set has 8 attributes and one class variable. Class variable has possible value of 0 or 1, 1 being interpreted as tested positive for diabetes. The dataset has 768 instances, out of which 268 being those tested positive for diabetes

Here, we will be applying classification algorithms from the scikit-learn library [8] and AutoGluon library [9] and checking the capacity of the algorithms to diagnose the diabetes disease. Scikit-learn is the most successful and robust library for machine learning in Python. This library is primarily written in Python and is based on the modules like NumPy [10], SciPy [11] and Matplotlib [12]. And the open source AutoML library AutoGluon-Tabular can train highly accurate different machine learning models with a single line of code [9]. The ML algorithms from

scikit-learn library and Auto-Gluon library are implemented with AWS SageMaker [13]. The Amazon SageMaker is capable of building, training, and deploying state of art Machine Learning models with full managed infrastructure tools and workflows[14].

Evaluation Metrics and Benchmark Model

Disease diagnosis is a classification task. And, Classification ML Algorithms are evaluated using Classification Accuracy Measures like Accuracy, Precision, Recall and F1-score. Let us consider a value of 1 (having diabetes) be positive and value of 0 in class variable be negative in the considered dataset. Let True Positive (TP) be the correctly classified number of positive classes from a ML model. Similarly False Positive (FP) be the number of incorrectly classified as positive classes, True Negative (TN) be the correctly classified number of negative classes and False Negative (FN) be the number of classes incorrectly classified as Negative classes. Various classification accuracy measures are computed based on TP, FP, TN and, FN [15]. The Accuracy can be computed as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

the Precision as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

the Recall as

$$\text{Recall} = \frac{\text{Tp}}{\text{TP} + \text{FN}} \tag{3}$$

and the F1-Score as

$$\text{F1} - \text{Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precison} + \text{Recall}} \tag{4}$$

Classification accuracy measures in Eq. (1), (2), (3) and (4) have been used to evaluate the performance of applied classifier algorithms. The author in [7] mentions that the baseline accuracy for diagnosing diabetes is about 65 percent when Pima Indian Diabetes Dataset is used for the classification problem and the best diagnosis accuracy obtained so far is about 77 percent. We will compare our results with these benchmark results.

Project Design

The project initiates by downloading the Pima Indian Diabetes Dataset [7]. Then, we will create a notebook instance in the Amazon SageMaker. When the notebook instance is successfully created, we will open the Jupyter Notebook and upload the downloaded dataset to the instance. We will make sure to use the with most cost-effective instance as our data set is not significantly large. After that we will, import the and download (if needed) the required module and libraries.

Next step is to perform Exploratory Data Analysis (EDA). Then, we will apply different state of art classification ML algorithms to Diabetes Dataset using scikit-learn and AutoGluon libraries and compute the performance of each approach in diagnosing the diabetes. Finally, we discuss the results obtained and summarize the findings of the project.

Reference

[1] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[2] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 09, no. 01, pp. 1–16, 2017, doi: 10.4236/JILSA.2017.91001.

[3] Neural Designer, "Machine learning use cases | Neural Designer." https://www.neuraldesigner.com/solutions (accessed Jan. 13, 2022).

[4] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006, doi: 10.1146/ANNUREV.BIOENG.8.061505.095802.

[5] J. Schaefer, M. Lehne, J. Schepers, F. Prasser, and S. Thun, "The use of machine learning in rare diseases: a scoping review," doi: 10.1186/s13023-020-01424-6.

[6] T. Freeman, "Deep learning for disease diagnosis confounded by image labels – Physics World." https://physicsworld.com/a/deep-learning-for-disease-diagnosis-confounded-by-image-labels/ (accessed Jan. 13, 2022).

[7] J. Brownlee, "10 Standard Datasets for Practicing Applied Machine Learning." https://machinelearningmastery.com/standard-machine-learning-datasets/ (accessed Jan. 12, 2022).

[8] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, Accessed: Jan. 12, 2022. [Online]. Available: http://scikit-learn.sourceforge.net.

[9] N. Erickson *et al.*, "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data," Mar. 2020, Accessed: Jan. 12, 2022. [Online]. Available: https://arxiv.org/abs/2003.06505v1.

[10] C. R. Harris *et al.*, "Array programming with NumPy," *Nat. 2020 5857825*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[11] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/S41592-019-0686-2.

[12] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. \& Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[13] SageMaker, "Amazon SageMaker – Machine Learning – Amazon Web Services." https://aws.amazon.com/sagemaker/ (accessed Jan. 13, 2022).

[14] AWS, "Amazon SageMaker Amazon Sagemaker API Reference Amazon SageMaker: Amazon Sagemaker API Reference."

[15] P. Galdi and R. Tagliaferri, "Data Mining: Accuracy and Error Measures for Classification and Prediction Neonatal MRI View project Computational methods for omics data View project Data Mining: Accuracy and Error Measures for Classification and Prediction," doi: 10.1016/B978-0-12-809633-8.20474-3.