Volume 9, Issue 5
October 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Comparative Analysis of Machine Learning-Based Techniques/Models for Classification and Regression Using Epidemic Datasets

[1]Rameshwer Singh and [2]Dr. Rajeshwar Singh

[1]Research Scholar, I.K. Gujral Punjab Technical University, Kapurthala, Punjab, INDIA

[2]Group Director, Doaba Khalsa Trust Group of Institutions, SBS Nagar, Punjab, INDIA

E-mail: [1]rameshwer.banga@gmail.com, [2]directordgc3@gmail.com

ABSTRACT

Machine learning techniques are playing a vital role in the present era or research and applications related to computer science and engineering. These techniques are very helpful in making such systems than can make predictions, generate important information and, help in enhancing the capabilities of the present systems. In this paper, we have presented a comparative analysis of popular classifiers and regression models based on machine learning techniques. We have divided our paper into three parts. The first part presents the popular machine learning-based techniques and models for classification and regression analysis. In the second part, we have presented different metrics to measure the efficiency of different classification and regression models. In the third part, we have presented the comparative analysis of eight classification techniques SVM (Linear SVC), Perceptron, Passive Aggressive Classifier, Logistic Regression, Ridge Classifier, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Ada Boost Classifier and six regression-based models that are Linear Regression, SVM, Bayesian Ridge, KNN, Decision Tree, Passive Aggressive regression. We have used COVID-19 related tweets for classification and real data sets of COVID-19 for regression analysis. As per our experimentation results, we have recommended SVM (Linear SVC) and perceptron as best machine learning-based classification techniques and Linear Regression as the most optimal performer among all regression models used in the experimentation.

Keywords: *Machine Learning Techniques/Models, Metrics, Comparative Analysis, Predictions*

1. INTRODUCTION

In the present era of artificial intelligence, machine learning techniques have a significant role in generation important information with an automated machine learning-based process. Applicability of machine learning techniques is broadly divided into three categories i.e. classification, regression analysis, and clustering [1,2]. Classification techniques are used for the automated classification of the object. When we need to make predictions related to continuous variables we shall use regression algorithms and, to group similar sort of objects in sets we need to use clustering techniques [3].

In many, if the studies it's been proved that machine learning techniques are very useful in making predictions systems, automated decision making systems, etc [1,2]. These kinds of prediction systems can predict continuous variables like a future sales forecast, future trends in the stock market, etc. Further, based on static variables, machine learning techniques can be used to classify the objects into different class types. For instance, based on the symptoms algorithms can classify that person seems to have disease A, disease B or no disease, etc.

2. BACKGROUND

Authors of paper [4], presented a prediction system to predict heart disease based on the symptoms described in a pre-formatted tweet on Twitter. This is a kind of classification problem that classifies whether the person is having the heart-related disease of not. In paper [5], influenza-related tweet classification is performed using machine learning algorithm linear regression.

In paper [6], authors performed a classification of more than 45000 tweets based on the epidemic Ebola. This paper presents the significance of machine learning techniques in the prediction of the early hint about the outbreak.

As per the paper [7], machine learning techniques help predict the disease-related information like predictions based on a predefined variable which comes under the classification techniques and other type is a prediction based on a continuous variable which comes under the category of the regression model. In this paper, we are going to present the comparative analysis of popular machine learning techniques for classification and regression.

3. POPULAR MACHINE LEARNING TECHNIQUES/MODELS

3.1 LINEAR REGRESSION

Linear regression is one of the popular machine learning techniques that can be used to predict using features of a continuous variable. The linear regression model is trained based on polynomial features [3]. In this technique, we need to use the weighted parameter that will be required for each training feature.

Volume 9, Issue 5
October 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Coefficient Θ is used with every feature that will produce (h(Θ)) [1,2,3,8].

$$h_\Theta = \Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \cdots \quad (1)$$

In the liner function squared error is used as an error function. While training the model Θ is to be updated to minimize the gap between the actual and predicted value to gain the best possible results.

3.2 LOGISTIC REGRESSION

Although the name of the model is logistic regression its is basically used for classification. It is also known as maximum entropthy[3]. It uses a Sigmoid function to produce a linear output to perform classification. Below is the YΘ which is the same as the hΘ in linear regression. It will set the Y to 0 when g(Y) will be 0.5 and the result will be 1 when h(Θ) is greater than 0.5 [1,2,3,8].

$$Y_\Theta = \Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \cdots \quad (2)$$

$$h(\Theta) = g(Y) \quad (3)$$

$$g(Y) = \frac{1}{1 + e^{-Y}} \quad (4)$$

This technique uses the sigmoid function (non-linear). The loss function of these techniques is also referred to as cross-entropy.

$$J(\theta) = \frac{1}{m}\sum cost(\beta', \beta) \quad (5)$$

$$cost(\beta', \beta) = -\log(1 - \beta') \text{ if } \beta = 0 \quad (6)$$

$$cost(\beta', \beta) = -\log(\beta') \text{ if } \beta = 1 \quad (7)$$

The value cost will be infinity i.e. –log (0) when the wrong value is being predicted.

3.3 K-NEAREST NEIGHBOR

Classification and regression both can be performed by using the k-nearest neighbor. It works based on the predefined sample sets and its distance to the new point. Based on this it predicts the label. It used the local approximation method to know about the neighbors and based on this it predicts the value [1,2,3,8].

3.4 DECISION TREE

The decision tree is s supervised machine learning technique that can be used to solve both classification and regression problems. This technique uses inference rules to learn from the sample sets predicts the traded value [3]. Every node in the decision tree tests a condition on the feature and moves further till it reaches the leaf node. Decision trees use Gini-index for classification [3].

$$giniindex\ (gi) = 1 - \sum x_t^2 \quad (8)$$

Attribute with the highest Gini-index will be selected for the next iteration.

3.5 NAIVE BAYES

In general, Naïve Bayes is a classification algorithm. Supervised learning is used. It works based on probabilities [1,2]. Multinomial Naïve Bayes variant specifically used for text classification [3]. Bernoulli Naive Bayes is another type of Naïve Bayes algorithm. It may have multiple features each one is having the binary value.

3.6 SUPPORT VECTOR MACHINE

Support Vector Machine(SVM) is one of the most famous machine learning techniques. It can be used for both classification and regression. Scikit Learn python library for machine learning techniques have different set function for classification and regression. SVR for regression and SVC for classification. Both sets of functions are based on SVM [3]. In SVM labeled data is provided for training and there is some tuning parameter for SVM like kernel, regulation, margin, and gamma. These parameters can be tuned to gain better results.

3.7 PERCEPTRON

Perceptron is also a supervised machine algorithm. It is a kind of linear classifier. It updates itself based on the errors only. It one of the simplest and better performer techniques [3]. Similarly, there are some other popular techniques like a passive-aggressive classifier. It is the same as the perceptron. The only difference is that it contains the tuning/regulation parameter. Ridge classifier converts the binary targeted output to -1 and 1 and then classifies the data. Another interesting machine learning technique is Ada Boost Classifier, in this technique classifier initially fits based on original data then makes the multiple copies of the same datasets and tries to adjust itself based on the wrong classification [3].

4. POPULAR METRICS TO EVALUATE THE ACCURACY OF MACHINE LEARNING ALGORITHMS

There are different kinds of applications of machine learning techniques like classification, clustering, and regression. In this section of the paper, we are going to discuss the various metrics that are used for evaluating the classification and regression analysis. We will also present some example functions used for the same in one of the famous python library i.e. SciKit learn [3].

Table 1 Metrics for machine learning-based classification and regression techniques

Sr. No	Metric	Description	Type	SciKit Function
1.	Accuracy	Prediction accuracy score based on labeled predicted and actual label of data.	Classification	accuracy_score
2.	F1 Score	It is the harmonic mean based on precision and recall.	Classification	f1_score
3.	Precision	This score tells how much true positive correct predictions	Classification	precision_score
4.	Recall	The score that describes several correct true positive predictions out of all the predictions	Classification	recall_score
5.	Variance	It is the measure of the difference between values observed and predicted value average.	Regression	explained_variance_score
6.	R2 Score	It is used to measure regression score known as coefficients of determination	Regression	r2_score
7.	Mean Absolute Error	It is a loss function to find an absolute error	Regression	mean_absolute_error
8.	Mean Squared Error	It is the square of the difference between the actual value and predicted values	Regression	mean_squared_error

5. EXPERIMENTATION & COMPARATIVE ANALYSIS OF CLASSIFICATION AND REGRESSION TECHNIQUES

To compare the different machine learning classification and regression techniques/models. We have experimented with the datasets of two types. Classification algorithms are tested based on tweets related to COVID-19 on twitter. More than 65000 tweets are collected and all the tweets are then labeled with the help of polarity score evaluated by TextBlob library of python [10]. After that different machine learning algorithms i.e. SVM (Linear SVC), Perceptron, Passive Aggressive Classifier, Logistic Regression, Ridge Classifier, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Ada Boost Classifier. We have tested the classification of tweets using unigram, bigram, and trigram. Below is the best accuracy score achieved by each of the classifiers. If we try to rank all the algorithms based on average accuracy scored based on all three types i.e. unigram, bigram, and trigram. We find an average accuracy score of 0.981113 achieved by both linear SVC and Perceptron [9]. Below is the classifiers as per our experimentation results.

Table 2 Maximum accuracy of various machine learning-based classification techniques

Machine Learning Technique	Maximum Accuracy Score
Linear SVC	0.985835
Perceptron	0.983122
Passive Aggressive Classifier	0.985232
Logistic Regression	0.980108
Ridge Classifier	0.977999
Multinomial NB	0.948764
Bernoulli NB	0.943942
Ada Boost Classifier	0.785112

Table 3 F1, Precision and Recall score of best performer classifiers

Machine Learning Techniques/Models	Class	Max F1 Score	Max Precision Score	Max Recall Score
Linear SVC	-1	0.99	0.99	0.96
	0	0.99	0.98	1
	1	0.99	0.98	0.99
Perceptron	-1	0.97	0.99	0.96
	0	0.99	0.98	1
	1	0.98	0.99	0.98

Similarly, we have experimented with the COVID-19 datasets for making predictions about the confirmed cases, death cases, and recovered cases word wide. We have used Linear Regression, SVM, Bayesian Ridge, KNN, Decision Tree Regressor, Passive Aggressive Regressor machine learning-based techniques/models for regression analysis. As per the experimentation results on world cases, we have found that the R2 Score and Variance score of linear regression is better than other while prediction about the confirmed cases and death cases and SVM performed better than linear regression and others only in predicting the recovered case. As per our experimentation, we recommend Linear Regression as the most optimal regression algorithms in such king of epidemic predictions.

6. CONCLUSION

In this research paper, we have presented the popular machine learning techniques/models that are used for classification and predictions. We have presented the different metrics that can be used to evaluate the performance of classification and regression techniques. further, we have presented the experimentation results based on a comparative analysis of eight classification techniques and six regression techniques. As per our experimentation results, we have found Linear Regression as the most optimal performer in case of regression analysis, and in case of classification both i.e SVM (Linear SVC) and Perceptron performed better than other in case of classification.

REFERENCES

[1] Singh, Rameshwer, Rajeshwar Singh, and Ajay Bhatia."Sentiment analysis using machine learning techniques to predict outbreaks and epidemics."

[2] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." Computing, Communications, and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.

[3] SCIKIT-LEARN [Online]. Available on [https://scikit-learn.org/stable/ as accessed on 10/05/2020]

[4] Nair, Lekha R., Sujala D. Shetty, and Siddhanth D. Shetty. "Applying spark based machine learning model on streaming big data for health status prediction." Computers & Electrical Engineering 65 (2018): 393-399.

[5] Santos, José Carlos, and Sérgio Matos. "Analysing Twitter and web queries for flu trend prediction." Theoretical Biology and Medical Modelling 11.S1 (2014): S6.

[6] Odlum, Michelle, and Sunmoo Yoon. "What can we learn about the Ebola outbreak from tweets?." American journal of infection control 43.6 (2015): 563-571.

[7] Singh, Rameshwer and Rajeshwar Singh. "A Survey and Analysis of Machine Learning Based Techniques/Models to Predict Epidemic Outbreak and Pattern using Twitter Data and Official Datasets" ACJ(DOI:10.01011.ACJ.2020.V9I4.00068749.00978): 1894-1909

[8] Jason Brownlee, Time Series Forecasting as Supervised Learning [online]. Available [https://machinelearningmastery.com/time-series-forecasting-supervised-learning/ as accessed on 31/03/2020]

[9] Singh, Rameshwer and Rajeshwar Singh. "Machine Learning Based Twitter Data Mining to Analyse Sentiments of Tweets Allied to COVID-19 Epidemic & Its Patterns", Vol. 9, No. 2, April 2020 International Journal of Information Technology and Electrical Engineering: 11-22

[10] TextBlob: Simplified Text Processing [Online]. Available [https://textblob.readthedocs.io/en/dev/ as accessed on 01/03/2020.

AUTHORS PROFILE

RAMESHWER SINGH received his degree of MCA from IKG-PTU. He is pursuing a Ph.D. in App-Sci-Computer Applications from IKG-PTU, INDIA. His areas of interest are Machine Learning, Data Mining, NLP. He is specialized in JAVA, Python, Android, and PHP programming. He is having more than 11 years of experience as a faculty of Computer Science.

DR. RAJESHWAR SINGH is presently working as Group Director at Doaba Khalsa Trust Group of Institutions, SBS Nagar, Punjab, India. His current research interests include Swarm Intelligence-based Optimization (Energy, Security, Routing), Wireless Sensor Network, intellectual information technology, Mobile Ad Hoc Networks