TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 5113-5117

Multivariate Data Classification Using Machine Learning

[1]Nirupama KS, [2]Akram Pasha, [3]Greeshma Reddy, [4]Sree Chandana, [5]Roopa
[2]Professor, [1,2,3,4,5]C & IT, Reva University, Bangalore, India, Bangalore, India

Article Info
Volume 83
Page Number: 5113-5117
Publication Issue:
May-June 2020

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

Abstract
Nowadays, data analytics has become one of the major business tools to gain insights into the data and make many important business decisions. The applications of data analytics are innumerable spawning major applications in various domains including healthcare. Machine Learning (ML) is one of the major tools that drive any data analytics application. Therefore, in this paper, an effort is made to classify the Liver Disease (LD) data set having multiple dimensions of attributes. The data set comprises the 583 observations taken from the liver disease patients and employed Minmax feature scaling to normalize the data. The data set was split into training and testing set in the ratio of 80% and 20% respectively. The training set was trained on Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), k-Nearest Neighbor (k-NN) classifiers to investigate the best classification model giving maximum accuracy. Amid all the ML classifiers involved, k-NN provides 80% of maximum classification accuracy.

Keywords: *Classification; Machine Learning.*

1. Introduction

Multivariate data is used for statistical analysis which allows us to concentrate and examine multiple variables at once. Multivariate data is mostly used for descriptive purposes. Multivariate models accept more than two variables which helps to examine severe complex occurrences and provides data patterns with much accuracy to represent the world. Data needs to be analyzed for their desperate purposes, ranging from the optimization of communications to cybersecurity and cyber warfare. It focuses mainly on designing systems thereby allowing them to learn and make predictions based on some experience which is data in the case of machines. At present, the terms AI, ML and Deep Learning are very familiar throughout industries. The ML advances have been widely employed in many applications that depend on the data collected in some specific form. These approaches employed statistical modeling and targets to provide productive learning-based results for data-demanding issues [1]. Generally, the ML classifier's performance reduces when the number of attributes given to them gradually increases. Most of the data sets acquired from the various field sectors have multiple dimensions of attributes leading to the obstruction of any ML classifiers throughout the classification process of the data set.

Nowadays, the advancement of computer technology has notably improved the data information in various organizations such as the university, bank, hospital, e-commerce sites, and many others. Data generated by the medical institute is very important for the knowledge extraction purpose that helps in fast diagnosing the disease and providing better treatment to patients.

The classification model aims to gather some conclusions from observed outcomes. In supervised learning algorithms, it builds a model of a data set that has both inputs and the desired outputs which work on labeled data. Classification is a kind of supervised learning. Machine learning is classified into two types. They are Linear and Nonlinear Models. Linear models consist of SVM and LR whereas Nonlinear models consist of k-NN, Naive Bayes, DT, and Random forest. In this paper, four ML classifiers have been considered: SVM, DT, LR [3], k-NN. The current research study employs the above ML classifiers for diagnosing LD.

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 5113-5117

The resulted areas of this article are contributed as follows: Session 1 contains the Introduction, Session 2 contains the Related Work, Session 3 explains the Materials and Methods with a framework, Session 4 describes Experimentation and Discussion of results, Session 5 gives the conclusion of research work with future outlook.

2. Related Work

The literature review carried out aims on delivering the most related work performed on the various sectors of classification models using ML. This session mainly deals with the studies published lately. Every single study is analyzed thoroughly to yield out the challenges and approaches used for classification purposes. The main objective of the studies published is either to recognize or record the disease by identifying the fundamental elements that are causing the disease.

In the work of [17], the various types of the liver data set were collected from two different regions and then employed on the performance of the classification techniques in terms of precision, accuracy. ML classifiers such as SVM, Back Propagation Neural Network algorithm, C4.5 were used to productively classify liver and non-liver disease data sets [5]. Liver cancer, hepatitis, and cirrhosis are the three major liver diseases that were predicted utilizing key features using the FT tree classifier and Naive Bayes. Based on the classification accuracy measure the ML classifiers were compared, concluding Naive Bayes as a preferred classifier with the best accuracy [6]. Reducing the subset of features from the Parkinson's disease data set with the help of the dimensionality reduction approach and employed on ML classifiers [1]. FT tree, Naive Bayes and K-star used to predict early diagnosis for liver disease disorder [7]. Naive Bayesian classifiers give higher performance when compared to C4.5 and SVM for the CDC chronic fatigue syndrome data set [8]. Intelligent diagnosis of liver disease by soft computing techniques [9]. Predict chronic kidney disease using data mining algorithms like k-NN, SVM with the help of a tool MAT LAB [10]. In the work of [11], standard classification algorithms were considered for figuring their performance in metrics terms like precision and accuracy to classify liver patient's data sets. The research helps in identifying the patients from healthy individuals by predicting liver disease using the ML model and helps them in diagnosing [12].

3. Methods and Materials

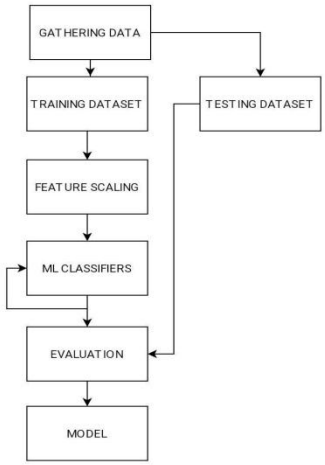

Figure 1: Proposed Framework

The fig.1 outlines the framework employed in this study. The major elements included in this framework are discussed in the following part of this section. We defined the stages of the framework like Gathering data, Data preprocessing, Classification models, Training and Testing the models and Evaluation.

Gathering Data

The idea of collecting data will always rely upon the project title. The Internet helps us to access some free data sets under different sectors. The most preferred repositories for collecting data to make project models were UCI Machine learning Repository and Kaggle. The data set opted for the research is composed of 583 observations taken over 10 attributes called features from 167 non-liver patient records and 416 liver patient records. 142 female patient records and 441 male patient records are included in this data set. The 10 attributes of the data set are like Patient Age, Patient Gender, Direct Bilirubin (*DB)*, Total Bilirubin (*TB),* Total Proteins (*TP)*, Albumin and Globulin ratio (*A/G ratio)*, Albumin (*ALB)*, SGPT Alanine Aminotransferase, SHOT Aspartate Aminotransferase and alkphos.

Data Preprocessing

Data Preprocessing is a technique used to convert raw data into clean data that is generally whenever the data is collected from any sector, it will be in a raw format which is not feasible for analysis. Data preprocessing is done to improve the quality of data in a data warehouse by increasing efficiency, removing noise, removing inconsistent data, etc. Various steps have been performed on raw formatted data to get clean data which is feasible for analysis; This technique is called data preprocessing. The model is trained once the data is clean and formatted. For achieving better results from the applied models in Deep Learning and ML using different ML classifiers, data preprocessing is very important.

The techniques used in the conversion of raw data into clean data are:

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 5113-5117

1. **Conversion of data:** As we know numeric features can be handled by ML models, hence ordinal data and categorical must be converted into numeric features by a one-hot encoding method.
2. **Removing the missing values:** When we come across missing data in the data set, depending on our needs we can remove the row or column from the data set. If missing value occurrence is more, then this method is not the appropriate one to be used.
3. **Padding the missing values:** Manually the missing data can be filled into the data set once the missing data is encountered. Highest frequency, Median, Mean is the most frequently used to fill the missing data.
4. **Elimination of Outliers:** The blunder mistakes found in the data set vary extremely from observations in a data set.
5. **Feature scaling:** Every data set we gather has features that can either be dependent or independent. Each feature has two properties called 'unit' and 'magnitude'. Unit is something used to measure the feature and magnitude refers to the value of the feature. Features having large values will lead to a bad analysis of the data. So, a pre-processing technique called feature scaling is employed to scale down the values without modifying the data set. The other need for feature scaling is to increase the performance of the classification models.

Using the above-mentioned pre-processing techniques, one hot encoding method was employed on the data set in order to convert the categorical data to numerical data without modifying the original data set. If null values are encountered in the data set two methods are applicable that is finding the average method by using mean, median, or mode and another way to clear the null values is to remove the entire column.

Classification Models

SVM: It is a discriminative classifier that is formally designed by a separative hyperplane. The examples are represented as points in a space that are mapped, and those points of different categories are separated by maintaining a gap as wide as possible between them. SVM is a subset of training data that is used to represent decision boundaries. SVM is applicable for the data that are linearly separable whereas for non-linearly data kernel functions are used. Steps to implement SVM models are loading the data, exploring the data, splitting data, generating the model and model evaluation.

LR: It produces results in binary format which is used to predict the outcome of categorical dependent variables. So, the outcome should be discrete or categorical such as 0 or 1, Yes or No, True or False, High or Low. The LR equation is derived from the straight-line equation. It helps in solving the classification problems. The curve obtained in LR is called S-curve. LR is used in weather prediction, classification problems, determines illness, etc. The steps involved in LR are collecting data, analyzing the data, data wrangling, train and test, accuracy check.

DT: The branches of the tree represent the possible decision occurrence or reaction. DT classifier falls under the category of supervised learning. They can be used to solve both regression and classification models. In DT each internal node implies a test on an attribute, the outcome of the test is represented by each branch.

k-NN: It is a supervised ML classifier. The intuition behind the k-NN is it is given to some training data and new data; we would assign a new data based on the class of the training data it is nearer to. It is the simplest of all ML classifiers. There is no explicit training required and it can be used for classification and regression purposes.

The steps for k-NN as follows:

Firstly determine 'K', followed by estimate the distance between new data and training samples. Later sort the distance followed by collecting the classes of the top three and
choosing the majority one.

Training and testing the data:

The data set will be split into two parts called the Training data set and the Testing data set. Firstly, we train our classifier with the help of a training data set and we check the performance of the classifier using a testing data set.

Training set: It is the actual data set that is used to train the model for performing various types of actions. ML classifiers are used for training.

Test set: A test set is a group of tests that belong to specific tasks or features or have some other reason to be run together.

In the data set, the model is implemented with the training data set and it is validated using validation tests and the test data set helps in the removal of data points present in the training data set. With the help of step 4, the model is employed with any of the ML classifiers.

Evaluation

The essential part of model growth is model evaluation. It guides us to find the best model that symbolizes our data set and concludes how the model is efficient and helpful in the future.

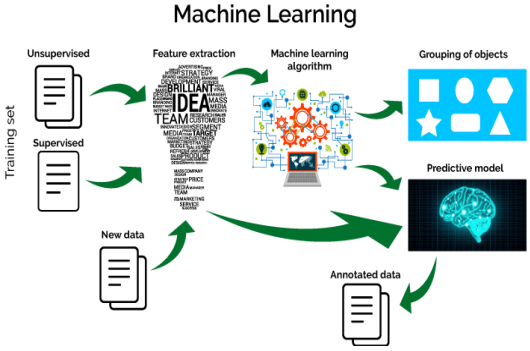

Figure 2: Evaluation of ML model [13]

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 5113-5117

4. Experimentation and discussion of Results

Experimental setup

The complete work was developed on a computing platform with the specifications of 1.5GB RAM, 1TB ROM, and i5 processor. The software packages used in this work are Pandas, NumPy, Scikit Learn, Matplotlib and Seaborn.
All these packages were employed on a web-based IDE for Python 3.7 version.

Discussion of Results

The LD data set having multiple dimensions of attributes. The data set comprises the 583 observations taken from the liver disease patients comprising 11 features.

Out[3]:

	Age	Gender	TB	DB	Alkaline	Alamine	asprate	TP	ALB	ALB RATIO	selection field
0	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...
578	False	False	False	False	False	False	False	False	False	False	False
579	False	False	False	False	False	False	False	False	False	False	False
580	False	False	False	False	False	False	False	False	False	False	False
581	False	False	False	False	False	False	False	False	False	False	False
582	False	False	False	False	False	False	False	False	False	False	False

583 rows × 11 columns

Figure 3: Loading of LD data set

Once after the data set is loaded the data is split into two sets namely the training set and testing set. We have considered 80% of the training data set and 20% testing data set which includes 466 training observations and 117 testing observations. The following step is where the Minmax feature scaling method is employed for better performance of classifiers.

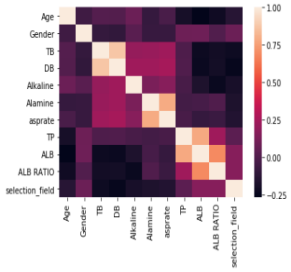

Figure 4: Correlation of attributes

Generally, feature scaling is one of the most important preprocessing techniques, mainly suits classifiers like k-NN, etc. That means to normalize the data without manipulating the original data. Once the preprocessing step is done, we choose the best model and train the model followed by testing the model and finally deploy the predictions.

Table 1: Comparison of classifier

ALGORITHM	ACCURACY	PRECISION
SVM	74.00%	50.00%
LR	72.00%	50.00%
k-NN	80.00%	68.00%
DT	70.00%	43.00%

s

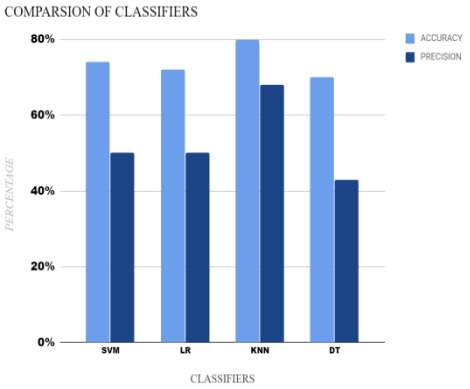

Figure 5: Accuracy Performance of different ML Classifiers

In fig 5, the performance of different ML classifiers is being compared. The results obtained are as follows: LR has an accuracy of 72%, DT has an accuracy of about 70%, SVM has 74% and k-NN has an accuracy of 80%.
By analyzing the results, k-NN with 80% accuracy gives the overall best classification result than the other ML classifiers employed above.

5. Conclusion and Future Outlook

In this study, Supervised ML classifiers were chosen for estimating their classification performance on-premises of accuracy. It made us understand how ML helps in pre-processing multivariate data and helps to overcome future problems that are present in the real world. Finally, the methods were tested on the LD data set for comparison of their performance and found k-NN has achieved higher accuracy.

The future outlook of this study is that the future of the machine is as vast as the limits of the human mind. We can always keep learning and teaching computers how to learn. Meanwhile, wondering how most complex ML classifiers have been running in the back of our mind so effortlessly all the time. The scope of ML keeps increasing. In the future, we will see ML-based houses, cars and many more things which we cannot even imagine. ML is also improving day by day like many companies are ready to adopt this technique.

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 5113-5117

References

[1] Pasha A, Latha PH. Bio-inspired dimensionality reduction for Parkinson's disease (PD) classification. Health Information Science and Systems. 2020 Dec;8(1):1-22.

[2] Dey A. Machine learning algorithms: a review. International Journal of Computer Science and Information Technologies. 2016;7(3):1174-9.

[3] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002 Oct 1;35(5-6):352-9.

[4] Faisal MI, Bashir S, Khan ZS, Khan FH. An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer. In2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST) 2018 Dec 21 (pp. 1-4). IEEE.

[5] AS D, Venkateswaran CJ. Estimating the surveillance of liver disorder using classification algorithms. International Journal of Computer Applications. 2012 Nov 10:0975-8887.

[6] Dhamodharan S. Liver disease prediction using bayesian classification. In4th National Conference on Advanced computing, applications & Technologies 2014 May (pp. 1-3).

[7] Rajeswari P, Reena GS. Analysis of liver disorder using data mining algorithms. Global journal of computer science and technology. 2010 Nov 25.

[8] Huang LC, Hsu SY, Lin E. A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data. Journal of Translational Medicine. 2009 Dec;7(1):81.

[9] Durai, Vasan, Suyan Ramesh, and Dinesh Kalthireddy. "Liver disease prediction using machine learning." (2019).

[10] Huang LC, Hsu SY, Lin E. A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data. Journal of Translational Medicine. 2009 Dec;7(1):81.

[11] Ramana BV, Babu MS, Venkateswarlu NB. A critical study of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems. 2011 May 2;3(2):101-14.

[12] Jacob J, Mathew JC, Mathew J, Isaac E. Diagnosis of liver disease using machine learning techniques. Int Res J Eng Technol. 2018 Apr;5(04).

[13] Ayush pant [2019], Machine learning project workflow [online image] retrieved April 13, 2020, from https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94.

[14] Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: A data perspective. ACM Computing Surveys (CSUR). 2017 Dec 6;50(6):1-45.

[15] Rathor S, Jadon RS. Domain Classification of Textual Conversation Using Machine Learning Approach. In2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2018 Jul 10 (pp. 1-7). IEEE.

[16] Sasikala BS, Biju VG, Prashanth CM. Kappa and accuracy evaluations of machine learning classifiers. In2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) 2017 May 19 (pp. 20-23). IEEE.

[17] Ramana, Bendi Venkata, M. Surendra Prasad Babu, and N. B. Venkateswarlu. "A critical comparative study of liver patients from the USA and INDIA: an exploratory analysis." International Journal of Computer Science Issues (IJCSI) 9.3 (2012): 506.

[18] Shu T, Zhang B, Tang YY. Effective heart disease detection based on quantitative computerized traditional chinese medicine using representation based classifiers. Evidence-Based Complementary and Alternative Medicine. 2017;2017.

[19] Auxilia LA. Accuracy Prediction Using Machine Learning Techniques for Indian Patient Liver Disease. In2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) 2018 May 11 (pp. 45-50). IEEE.

[20] Adil SH, Ebrahim M, Raza K, Ali SS, Hashmani MA. Liver Patient Classification using Logistic Regression. In2018 4th International Conference on Computer and Information Sciences (ICCOINS) 2018 Aug 13 (pp. 1-5). IEEE.

Published by: The Mattingley Publishing Co., Inc.