International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S, August 2019

Research of Machine Learning Algorithms using K-Fold Cross Validation

Nagadevi Darapureddy, Nagaprakash Karatapu, Tirumala Krishna Battula

ABSTRACT--- In machine learning, Classification is one of the most important research area. Classification allocates the given input to a known category. In this paper different machine algorithms like Logistic regression (LR), Decision tree (DT), Support vector machine (SVM), K nearest neighbors (KNN) were implemented on UCI breast cancer dataset with preprocessing. The models were trained and tested with k-fold cross validation data. Accuracy and run time execution of each classifier are implemented in python.

Keywords - Logistic regression (LR), Decision tree(DT), Support vector machine (SVM), K nearest neighbors (KNN), K-fold cross validation.

I. INTRODUCTION

Machine learning at an esteemed level is commonly the course of teaching an algorithm on how to gradually improve upon a given task. In research machine learning may be observed as the theoretical and mathematical modeling on work process. Practically it is a study of building applications that display iterative improvement. Predominantly out of many ways three most recognized classes are unsupervised learning, supervised learning and reinforcement learning.

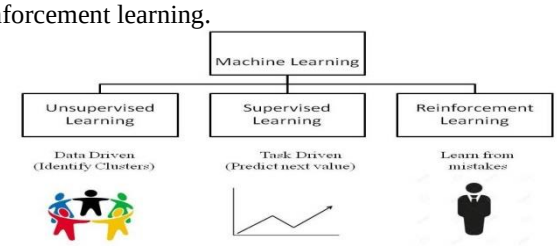

Figure 1: Types of machine learning

Supervised Learning utilizes training data and feedback from individuals to determine the relation of provided inputs to outputs. For instance, to predict the given image is apple or not. Supervised learning practiced when data is labeled. This algorithm forecasts new data. There are two categories of supervised learning:

- Classification
- Regression

Classification categorizes the data into given number of classes. It can be binary classifier with only two classes ex. Classification of spam and non spam email or Multi class classifier with more than two classes ex. Classification of types of students.

When the output is a continuous value, the task is a regression. For example, a financial interpreter may require calculating the amount of a stock based on a range of features like equity, previous stock performances, and macroeconomics index. The system will be trained to estimate the amount of the stocks with the lowest possible error.

II. METHODOLOGY

The algorithms in this paper implemented on the dataset of mammograms with classification as the task.

A. Dataset: Breast Cancer Wisconsin (diagnostic)

Features are enumerating from a digitized image of a fine needle aspirate (FNA) of a breast mass. They characterize the cell nuclei exist in the image. The attributes in this data set are ID number and diagnosis (M=malignant, B=benign) [7], from each cell nucleus ten real-valued features compute are radius, texture, perimeter, area, smoothness, compactness, concavity, concavity points, symmetry, and fractal dimension. Below figure 2 shows the parameters of the dataset and figure 3 shows the count of malignant which is represented with 1and benign with 0.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_m
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280

5 rows × 33 columns

Figure 2: Parameters of the dataset

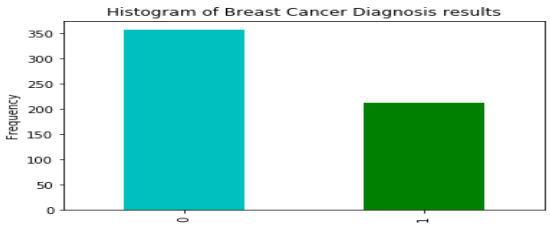

Figure 3: Parameters of the dataset

Manuscript published on 30 August 2019.
* Correspondence Author (s)
Nagadevi Darapureddy, Department of ECE, Chaitanya Bharathi Institute of Technology, Hyderabad, A.P, India. (E-mail: devi.darapu@gmail.com)
Dr. Nagaprakash Karatapu, Department of ECE, Gudlavalleru Engineering College, Gudlavalleru, A.P, India. (E-mail: drprakashece@gmail.com)
Dr. Tirumala Krishna Battula, Department of ECE, J N T U Kakinada, Kakinada, A.P, India. (E-mail: tkbattula@gmail.com)

B. Data Preprocessing and features selection

In data preprocessing, data is to be cleaned first as it contains irrelevant data. The parameters of dataset as shown in fig.2 has irrelevant data i.e. column Id. Next data transformation is done by attribute selection and then data reduction is done by attribute subset selection. The dimensionality of features are reduced by observing the correlation between features. Figure 4 shows the heat map which shows the correlation between features. It can be seen that texture_mean, perimeter_mean, smoothness_mean, compactness_mean, symmetry_mean area are highly correlated and are selected as features for classification.

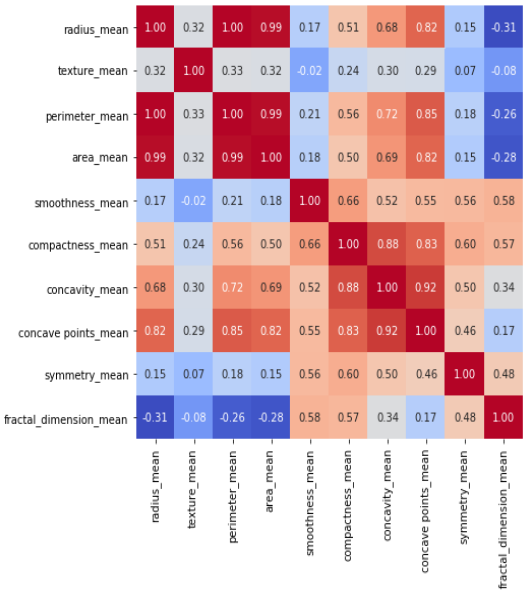

Figure 4: Heat map of features.

C. Task

The task performed is classification to predict the stage of breast cancer whether it is Malignant or Benign. The Classification used only when the output variable is a category. The categories used are malignant and benign which is two class or binary class classification.

D. Machine learning algorithms

To solve different kinds of problems different machine learning algorithms are available. The choice of the algorithm depends on the objective. In supervised learning, Classification is used to identify fraud detection, Image classification, in diagnostics, customer retention and Regression is used to forecast the weather, market forecasting, advertising popularity prediction, estimating life expectancy, population growth prediction.

TABLE 1:DIFFERENT ALGORITHMS WITH THEIR TASKS

Algorithm Name	Description	Type
Logistic Regression	Output variable is binary rather than continuous	Classification
Decision Tree	Splits the feature values of data into branches at decision nodes	Classification , Regression
Support vector machine	Optimally divides the classes by finding a hyper plane	Classification , Regression
K Nearest neighbours	Based on similarity measures from its neighbours	Classification , Regression

General Algorithm of predictive analysis on breast cancer dataset:

1. Acquiring the dataset.
2. Performing exploratory data analysis (EDA) to understand structure, nature, interrelationships of the data.
3. Preprocessing of the data involves activities such as normalizing the feature, assigning numerical values to categorical data etc.
4. Build a model
5. Hyper parameter tuning
6. Finalize the model

E. Cross validation

Cross validation is used to test the effectiveness of machine learning model; it's a model validation technique. It is used to limit the problems like over fitting, under fitting and get an intuition how the model will generalize to an independent dataset. This is done by dividing the whole data into two sets training and test set. In this paper K-fold cross validation method is used with k value as 5.So the whole data is divided into 5 folds and iterate 5 times.

III. CLASSIFIERS

Different classifiers are implemented on the data to find the accuracy.

A. Logistic regression

It is used to find a relationship between features and probability of particular outcome. E.g. When we want to predict whether it rains or not for given climatic data as features, the response variable has two values, Yes or No. This is called as Binomial logistic regression. When the output response has two or more possible values, this type of problem is referred as Multinomial logistic.

Algorithm:

- Initialize w and b randomly
- Using sigmoid function calculate the predicted output values.
- Calculate loss using a loss function
- The values of w and b are updated such that loss steadily decreases to an acceptable minimum value.

B. Decision Tree

Decision tree uses tree representation in which each node corresponds to a attribute or feature and the branch from each node represents the outcome of that node.

Algorithm:

- Compute entropy for every attribute or feature
- Take average information entropy and gain for the current attribute.
- Pick the attribute which has highest gain.
- Repeat until we get the tree as we decided.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

C. Support vector machine

It is used to find hyper planes that classify data points. Two types of SVM classifiers. In linear classifier model, finding the hyper plane is on maximizing the distance from hyper plane to the nearest data point of either class which is called maximum margin hyper plane. Non linear classifiers are created by applying kernel trick to maximum margin hyper planes.

Algorithm for separable case

It should separate two classes A and B defined by

- f(x)= a.x+b is positive if and only if x ∈ A
- f(x)<0 if and only if x ∈ B

D. K Nearest Neighbours

For a new instance, predictions are made by searching the entire training set for the most k similar neighbors and summarizing the output variable for those k cases.

Algorithm:

- Initialization of parameter k= number of nearest neighbors
- Calculate the distance between new instance and all the training samples.
- Sort the distance and determine nearest neighbor based on kth minimum distance.
- Assign the category based on the nearest neighbors majority category.

IV. RESULTS

The cross validation score for k=5 with different classifier are given in table 2.Accuracy is calculated with average of five iterations which is shown in table 3.

TABLE 2. CROSS VALIDATION SCORE FOR K=5 FOR DIFFERENT ALGORITHMS

Model	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Logistic Regression	85.00%	88.12%	90.41%	89.01%	89.44%
Decision Tree	90.00%	91.87%	91.25%	89.95%	90.44%
Support vector machine	90.00%	91.25%	90.00%	89.33%	90.20%
K Nearest Neighbours	88.75%	90.00%	90.41%	89.33%	89.69%

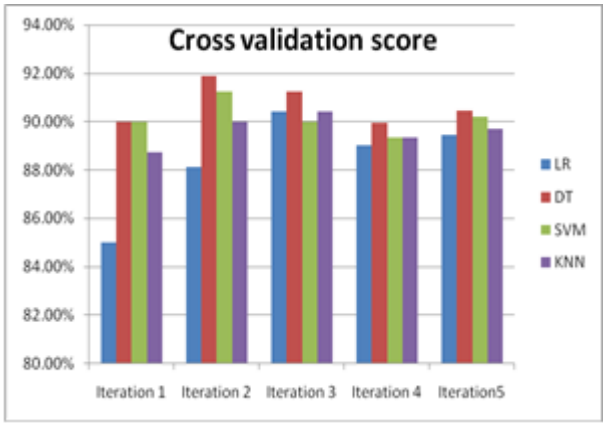

Figure 5: Cross validation score for k=5.

TABLE 3: ACCURACY AND RUN TIME FOR DIFFERENT ALGORITHMS

Model	Accuracy	Run time(in sec)
Logistic Regression	88.39%	0.06
Decision Tree	90.70%	0.07
Support vector machine	90.15%	0.17
K Nearest Neighbours	89.63%	0.02

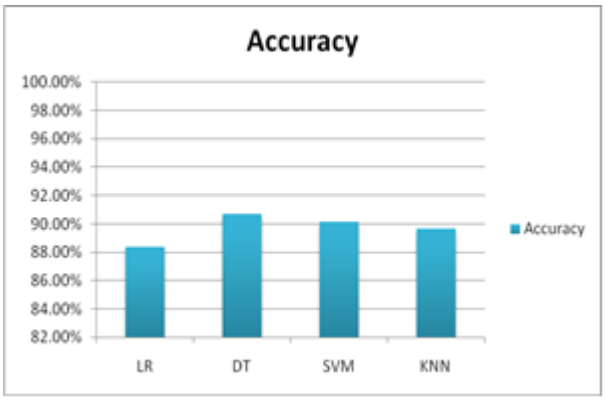

Figure 6: Accuracy for different Algorithms.

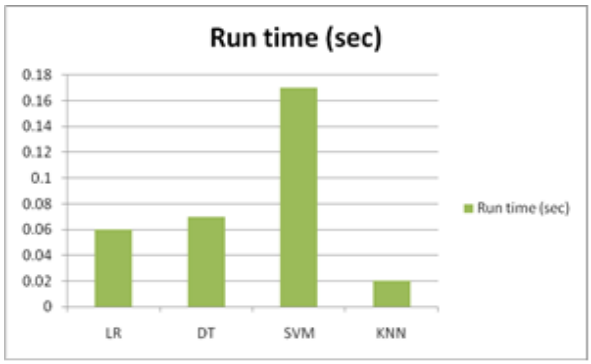

Figure 7: Run time for different Algorithms.

V. CONCLUSIONS AND FUTURE SCOPE

From the table 3, it can be observed that the accuracy for classifying the dataset using decision tree classifier is more when compared with other classifiers using k-fold cross validation. Other cross validation techniques like repeated k-fold, leave one out (LOO) can be implemented. Further the remaining learning algorithms like Gradient Descent, Nesterov accelerated gradient, Adagrad, Adam, Adadelta, momentum-based gradient descent can be applied to deep networks to compare the accuracy.

REFERENCES

1. "Patch-based system for Classification of Breast Histology images using deep learning" Kaushiki Roy, DebapriyaBanik, Debotosh Bhattacharjee, MitaNasipuri science direct article 2018.
2. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2016.html

3. R.L. Siegel, K.D. Miller, A. Jemal, Cancer Statistics, 2017, CA Cancer J. Clin. 65(1) (2015).
4. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction", *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8-17, 2015.
5. An overview of gradient descent optimization algorithms article by Ruder, Sebastian.
6. T. J. Cleophas, A. H. Zwinderman, "Machine Learning in Medicine", pp. 1-271, 2013.
7. M. Lichman, "UCI Machine Learning Repositry", 2013, [online] Available: https://archive.ics.uci.edu/.
8. J. Anitha, J.D. Peter, A wavelet-based morphological mass detection and classification in mammograms, in International Conference on MachineVision & Image Processing, 2013.
9. L. Arbach, D.L. Bennett, J.M. Reinhardt, et al., Classification of mammographic masses: comparison between backpropagation neural network (BNN) and human readers, Proc. SPIE 5032 (2003) 1441–1444, vol. 3.
10. N.V.S.S.R. Lakshmi, C. Manoharan, An automated system for classification of microcalcification in mammogram based on Jacobi moments, Int. J. Comput.Theory Eng. 3 (3) (2011) 431–434.
11. X. Liu, J. Liu, D. Zhou, et al., A benign and malignant mass classification algorithm based on an improved level set segmentation and texture feature analysis, in International Conference on Bioinformatics & Biomedical Engineering, IEEE, 2010.
12. J.Seok, B.Hyun, J.Kasa-Vubu, and A.Girard, ‘‘Automated classification system for bone age X-ray images,’’ inProc. IEEE Int. Conf. Syst., Man, Cybern. (SMC), Oct. 2012, pp. 208–213.
13. Ch. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
14. T. Mitchell, Machine Learning, MIT Press and McGraw-Hill, 1997.
15. L. Bottou, "Stochastic gradient descent tricks" in Neural Networks: Tricks Trade, Berlin, Germany:Springer, pp. 421-436, 2012.