2018 4th International Conference for Convergence in Technology (I2CT)
SDMIT Ujire, Mangalore, India. Oct 27-28, 2018

Iris Flower Species Identification Using Machine Learning Approach

[1]Joylin Priya Pinto
Department of Computer Science and Engineering
NMAM Institute of Technology,
Nitte
[1]priyahirgan@gmail.com

[2]Soumya Kelur
Department of Computer Science and Engineering
NMAM Institute of Technology,
Nitte
[2]pkelursoumya16@gmail.com

[3]Jyothi Shetty
Department of Computer Science and Engineering
NMAM Institute of Technology,
Nitte
[3]jyothi_shetty@nitte.edu.in

***Abstract* -Classification is one of the most important approach of machine learning. Main task of machine learning is data analysis. Various algorithms are available for classification like decision tree, Navie Bayes, Back propagation, Neural Network, Artificial Neural, Multi-layer perception, Multi class classification, Support vector Machine, K-nearest neighbor etc. In this paper, three methods are explained in detail. Implementation is done using iris dataset. Scikit tool is used for the implementation purpose. This paper mainly applies classification and regression algorithms on IRIS dataset, by discovering and analyzing the patterns, using sepal and petal size of the flower. We have found that SVM classifier gives best accuracy compared to KNN and logistic regression models.**

Index Terms - Classification, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Machine Learning.

I. INTRODUCTION

The machine learning is the subpart of computer science. Machine learning centers around the advancement of computer programs that can show themselves to grow and change at the point when exposed on new unseen data. It is an exploration field which has the intersection of both predictive and statistical analysis. There are two fundamental categories of machine learning. They are supervised and unsupervised learning and here in this paper, we focus on supervised learning approach, which is the process of inferring a function from labeled training data. The training dataset consists of training samples. In supervised learning each training sample consists of a pair of an input value with the desired output value. Supervised learning can be based on classification and regression. If the output value is categorical, then that is termed as classification or else if the output value is a real value, then that is regression.

In this paper, we have presented the methods for the identification of Iris flower species. The Iris dataset or Fisher's Iris dataset is a multivariate data set presented by biologist and statistician Ronald Fisher in 1936. It is basically published at UCI Machine Learning Repository. Dataset is also called as Anderson's Iris dataset because he collected data or information to evaluate the morphological variety of Iris dataset of three related species. Two of the three species were collected in the Gaspe peninsula. Here the Iris dataset consists of 50 samples from each three species: setosa, vesicolor and virginica. This dataset consists of four characteristics of the three species of the Iris flower, which are sepal length, sepal width, petal length, and petal width. The aim of this analysis is to obtain good accuracy scores for the unseen data.

Classification works on the basis of training and testing sets. While doing the training process training dataset is loaded into the models of machine learning and further labels are assigned. From the samples four features are identified: sepal length, petal length, sepal width and petal width in centimeters. Fisher discovered linear discriminant model to differentiate the species from each other by integrating these four attributes.

Paper presents the machine learning approach with Scikit-learn tool. A number of classification algorithms are used for Iris flower species recognition. But here, we are concentrating on only three classification algorithms such as Support Vector Machine, K-Nearest Neighbor and Logistic Regression classifiers, to perform iris flower classification.

II. LITERATURE REVIEW

Many methods have been introduced for iris dataset. Each and every method uses different strategy. This review consists of some prominent solution for iris species.

Deeptam Dutta et. al.[1] proposed a method on training Artificial Neural Networks. In this paper, IRIS bloom classification is done by utilizing Neural System. The issue concerns the recognizable proof of IRIS bloom species on the premise of bloom quality estimations. Characterization of IRIS informational collection would find designs from analyzing petal and sepal size of the IRIS blossom along with how the forecast was produced using breaking down the example to frame the class of IRIS bloom. By utilizing this example and order, in future upcoming years the obscure information can be anticipated all the more unequivocally. Author also analyzed that artificial neural systems have been effectively connected to issues in design arrangement, work approximations, advancement, and affiliated recollections. In this work, Multilayer nourish forward systems are prepared utilizing back propagation learning calculation.

Poojitha A et. al.[2] reviewed collection of Iris flower using neural networks. Machine learning is subpart of the computer science. Existing iris bloom dataset is preloaded in MATLAB and is utilized for bunching into three unique species. The dataset is grouped utilizing the k implies calculation and neural system bunching instrument in MATLAB. Neural system bunching apparatus is mainly utilized for grouping huge information sets with no supervision. It is likewise utilized for design

acknowledgment, highlight extraction, vector quantization, picture division, work approximation, and information mining. Results/ Discoveries: The outcomes incorporate the grouped iris dataset into three species with no supervision.

Viashali Arya et. al.[3] focused on efficient neural fuzzy approach for classification. In this paper, the proposed technique is connected on Iris informational indexes and groups the dataset into four classes. For this situation, the system could choose the great highlights and remove a little yet satisfactory arrangement of standards for the grouping assignment.

Shashidhar T et. al.[4] proposed identification of iris flower using classification. In this work, they have made predictions on unvisible data which is not used to train the model. They have shown machine learning models which predict the accurate feature of the species. They have done the work on the machine learning model by training the data sets and they have also discovered model for prediction using the species.

Patrick S. et. al.[5] focused on statistical analysis of IRIS flower dataset. In this paper, they have analyzed two different methods. First, plotting dataset to determine different patterns in the classification. Secondly, they developed an application in java to extract statistical information.

III. IMPLEMENTATION DETAILS

The objective of our methodology is to choose the best classification model which performs well on iris flower species identification. Learning models are created based on three machine learning algorithms: Support Vector Machine (SVM), Logistic Regression and K- Nearest Neighbor (KNN) classifier. For developing these methods four iris dataset features are used in the train and test datasets. All these algorithms are implemented using scikit-learn tool kit based on Python.

In this analysis we are comparing the accuracy of three models and trying to find out which classification model holds good. To improve model accuracy, we have also used cross validation technique which evaluates the predictive models by dividing the original sample data into a training set in order to train the model and a test set to evaluate it.

A. Dataset

For implementation purpose we have used Iris dataset which is a multivariate dataset that quantifies the structural variation of three species of iris flower. Classification is done on the basis of flower species which are Iris-setosa, Iris-versicolor, Iris-verginica. The dataset consists of 50 samples from each of three species that, totals to 150 samples. From each sample four features were measured that is, sepal length, petal length, sepal width, petal width – all in centimeters. The fifth attribute will be the species of the observed flower.

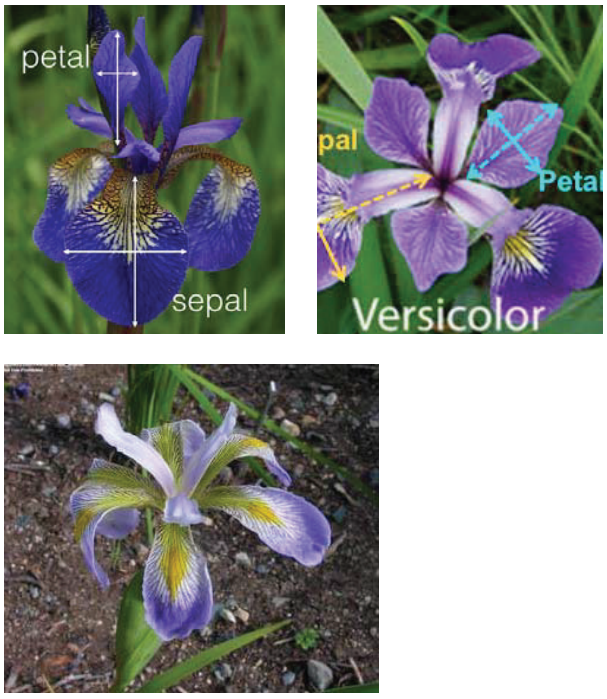
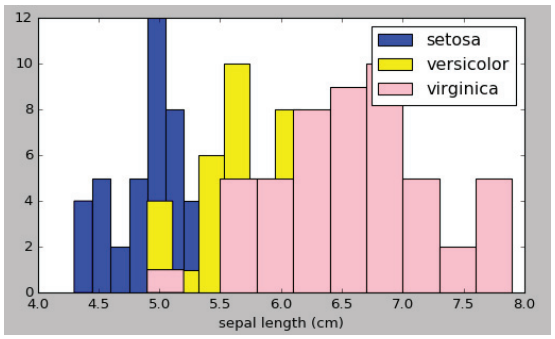
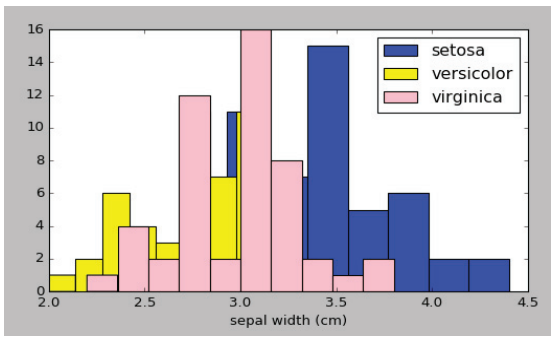


Fig 1. Iris Flower Species

Fig 2 shows how each iris feature is distributed among three flower species. We have plotted the histogram of the targets with respect to each feature of the data set. We can clearly see the feature ‘petal width’ can distinguish better than sepal length, sepal width and petal length.





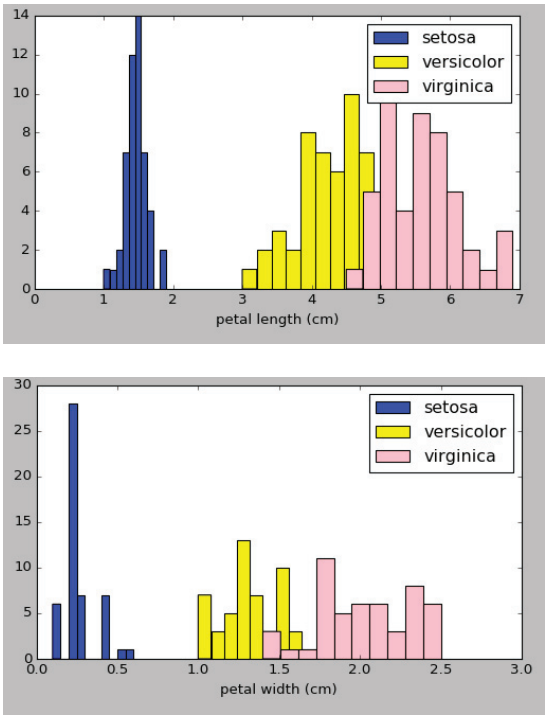


Fig 2. Iris Feature Distribution

B. Loading

First we have collected dataset of Iris flower from UCI Machine Learning Repository. There are total 150 samples belong to three various species of Iris flower that is setosa, versicolor, virginica. In the next step, collected dataset is loaded into the machine learning model. Here, we have used scikit learn tool and load_iris( ) python function to import the Iris dataset from scikit-learn datasets. This function is used to run and save return value in an object called "Iris". Then the attributes are assigned. In the Iris flower dataset attributes are data, feature names, target etc. and the target name for the Iris are classes that is setosa, virginca, versicolor. Feature names are nothing but sepal length, petal length, sepal width, petal width. Then dataset is divided into train and test data. Test data size is set to 40% while remaining 60% is kept for training purpose from the original dataset. Also random_state is set to 0. We should set random state to some random number. Because, if we run the model without specifying random_state we will get a different result every time when we execute, which makes difficulty in the accuracy analysis.

C. Model selection

In this step we have imported the model, based on which species prediction should be done. From scikit-learn tool, three classifiers are imported as follows:

1) Support Vector Machine

SVM is a promising new strategy for the arrangement of both linear and non-linear information. It utilizes the non linear mapping to change the first preparing information into a higher measurement. Inside this new parameter, it looks for the isolating hyperplane[9]. The hyperplane utilizes set of vectors. Support vector machine is a supervised machine learning algorithm which is used for both classification and regression problems. In this algorithm, we have plotted data item in n dimensional space (n is number of features considered in IRIS dataset) and then classification has been done.

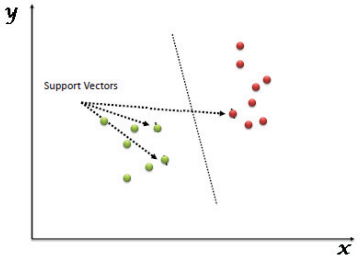


Fig 3. SVM Hyperplane

We have implemented learning model using SVM algorithm. SVC( ) function of sklearn is used to import the model. Predictions are made on the input test set. After comparing the actual response values with predicted response values, we got the SVM model accuracy as 96%.

2) Logistic regression

Logistic Regression is a statistical method for dataset analysis in which, there are one or more independent variables that determine the output[5]. The main objective of logistic regression is, it is a way used to split the data to get an accurate prediction of the class which uses the present information. Here in this Iris flower dataset, we are examining the Iris dataset. This classifier finds the method to split the given data based on length, width of Iris flower. One of the famous classification algorithms is logistic regression to analyze the target feature. It is a Nonlinear function which uses the sigmoid function as hypothesis which is given by p=1/(1+e-y). Here categorical and binary data are taken as the target variable. It is based on binary outcome i.e. 1/0 or Yes/No or True/False. Logistic Regression works well with large dataset.

LogisticRegression( ) function from sklearn.linear_model is used here to import the model. After examining the Iris dataset on the basis of logistic regression we got the model accuracy as 91%.

3) K- Nearest Neighbor Classifier

K-Nearest Neighbor is used for both classification and regression problems. KNN belongs to supervised learning. This algorithm is easy to understand. KNN model is entirely based on the training dataset. Here whenever we require prediction for unseen data, this algorithm will search for k-most similar instances. K-nearest neighbor classifier is powerful because, in order to perform classification it will measure the distance between two instances to find the similarity. Then based on the similarity, it will classify the incoming data.

For implementation of KNN classifier, we have used KneighborsClassifier(n_neighbors=3) function from sklearn.neighbors package. Based on the predictions, we got the model accuracy as 93%.

D. Cross validation

Cross validation is a technique to reserve a particular sample of a dataset using which model is not trained. Each training subset is validated with this portion of sample data before finalizing. While validating, first we have to reserve the sample of dataset, and next we have to train this model

by using left over part of the dataset. And again we have to use reserve samples of the test set or validation set which is helpful for the effectiveness of model performance.

Cross Validation steps:

i) First we should reserve sample data

ii) Train the model using remained portion of the datasets

iii) To check model's performance, now we should reserve sample of the test set. If the model gives good result with this validation data, then that model can be considered as a good model.

The main advantage of cross validation is to get more accurate estimate result from the sample accuracy and it is more efficient and helpful for the effectiveness of model performance.

The Fig 4 shows model accuracy for SVM, KNN and Logistic Regression classification methods applied on the iris dataset. Here, we found the accuracy with and without cross validation process. We analyzed that, model accuracy is increased compared to without cross validation observations.

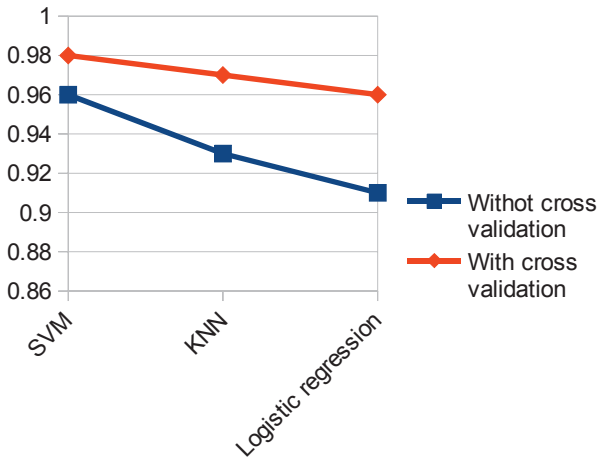


Fig. 4. Model Accuracy Score

E. Confusion Matrix

A Confusion matrix is a table to describe classification model performance. It is based on test data for which, output label is already known. Confusion matrix gives clear idea about model is giving proper outcome or not. Also it is easy to identify the model errors. We have calculated model accuracy score using confusion matrix.

Table 1 illustrates sample iris dataset.

TABLE I. ILLUSTRATION OF SAMPLE IRIS DATASET

SEPAL LENGTH	SEPAL WIDTH	PETAL LENGTH	PETAL WIDTH	SPECIES
2.0	5.6	3.0	0.5	SETOSA
3.2	5.3	3.7	1.3	VERSICOLOR
7.1	3.5	3.9	1.5	VERSICOLOR
6.8	3.3	4.9	1.3	VERSICOLOR
4.2	5.2	5.1	2.7	VIRGINICA
5.6	4.5	5.6	8.0	VIRGINICA

Here, first four columns are the Iris dataset attributes and fifth column specifies the target, which shows the classified label for the above sample data.

IV. CONCLUSION

The paper describes the various methods and algorithms used in the analysis of iris dataset. SVM, KNN and Logistic Regression methods are used to get good accuracy result and we have also applied the cross validation technique to improve the accuracy. We have compared the accuracy with and without cross validation technique. With the help of three methods, we have found that SVM classification method is more effective than the other methods. This paper also mentions the important functions of scikit– learn software tool that are applied to learn machine learning.

REFERENCES

[1] Diptam Dutta, Argha Roy, Kaustav Choudhury, "Training Aritificial Neural Network Using Particle Swarm Optimization Algorithm", International Journal on Computer Science And Engineering(IJCSE), Volume 3, Issue 3, March 2013.

[2] Poojitha V, Shilpi Jain, "A Collecation of IRIS Flower Using Neural Network CLusterimg tool in MATLAB", International Journal on Computer Science And Engineering(IJCSE).

[3] Vaishali Arya, R K Rathy, "An Efficient Neural-Fuzzy Approach For Classification of Dataset", International Conference on Reliability, Optimization and Information Technology Feb 2014.

[4] Cho, SungBae.and Dehuri, Satchidananda (2009) "A comprehensive survey on functional link neural network and an adaptive PSOBP learning for CFLNN, Neural Comput & Application"DOI10.1007/s00521-00902885

[5] Patric s hoye "stastical analysis of iris flower dataset" University of

[6] Massachusetts At Lowell

[7] Card, S., Mackinlay, J., and Shneiderman, B. Information Visualization.Readings in Information Visualization: Using Vision to Think, pp.1-34; 1999, Morgan Kaufmann Publishers, Inc., USA.

[8] Bache, K.& Lichman, M. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[9] Bishop, C. 2006. Pattern Recognition and Machine Learning. New York:Springer, pp.424-428.

[10] Fisher, R.A. 1936. UCI Machine Learning Repository:Iris Data Set. Available at:http://archive.ics.uci.edu/ml/datasets/Iris. Consulted 10 AUG 2013.

[11] Mjolsness, E. & Decoste, D. 2001. Machine learning for science: state of the art and future prospects.Science, 293 (5537), pp. 2051-2055.

[12] Pedregosa, F.& Varoquaux, G. 2.11., Scikit-learn: machine learning in Python— Scipy lecture notes, Available at:http://scipy lectures.github.io/advanced/scikit-learn/.Consulted 22 AUG 2013

[13] D. M. J. Tax, M. Van Breukelen, R. P. W. Duin, and J. Kittler. Combining multiple classifiers by averagingor by multiplying?Pattern Recognition, 33(9):1475–1485, 2000.

[14] J. Canny. A computational approach to edge detection.Pattern Analysis and Machine Intelligence, IEEE Transactions on, PAMI-8(6):679-698, 1986.

[15] X. He, S. An and P. Shi. Statistical texture analysis-Based approach for fake iris detection using support vector machines advances in biometrics. Springer Berlin /Heidelberg, 540-546, 2007