Iris Flower Classification Utilizing The Nearest Neighbor Algorithm

Israel Nolazco
Introduction to Data Mining and Analytics
Lewis University
Romeoville, Illinois
israelnolazco@lewisu.edu

***Abstract*— Machine Learning is a field of study in Data Science in which the primary goal is to create a system that can learn from data. This document will demonstrate the capabilities of the Nearest Neighbor Algorithm. This algorithm utilizes Euclidean distances to remember the training example and uses it as a reference point for the desire test set. Euclidean distance[1] refers to the distance between two points. These points can be in different dimensional spaces and are represented by different forms of coordinates. In two-dimensional space, the coordinates are given as points on the x- and y-axes. In this case, the dataset is going to be represented width and length as a two-dimensional variable. Additionally, this document will present the accuracy of the Nearest Neighbor Algorithm and present alternative insight to improve accuracy.**

Keywords

Nearest Neighbor Algorithm, Data Analytics, Euclidean distances

I. INTRODUCTION

The word Iris means rainbow [2]. Irises come in many colors, shapes, and forms. Although, there are three primary types of Iris Flower types that will be the focus in this document. Iris Setosa, Iris Versicolour, Iris Virginica which are visually similar from one to another Figures 1,2,3. Given the difficulty to identify each type of flower by just looking at it; The need for utilizing Machine Learning is highly demanded. As previously mentioned, the use of the Nearest Neighbor Algorithm is one best suited for this type of workload. All that is required to get started is to have two data sets. One represents the training point, a data set that will be represent a reference point in which all the test data sets will be compared to. If we were to represent the petal width and sepal length, as its own set of coordinates, then the Euclidean Distance can be calculated based on a two-dimensional space. However, as usual, simple mathematics is no true answer. In order to classify Iris flower to the highest possible accuracy, we must only look at the closest distance between the testing and training data sets.

Figure.1. Iris Setosa Image obtain from the following link: https://upload.wikimedia.org/wikipedia/commons/5/56/Kosaciec_szczecinkowaty_Iris_setosa.jpg

Figure 2. Iris Versicolour . Image obtain from the following link: https://upload.wikimedia.org/wikipedia/commons/4/41/Iris_versicolor_3.jpg

Figure 3. Iris Virginica. Image obatin from the following link: https://www.flickr.com/photos/33397993@N05/3352169862

II. Methods

The training data was obtained from THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS [3] in which a total of seventy-five rows measurements were classified as either Iris Setosa, Iris Versicolour, and Iris Virginica. As for the testing data [4], this dataset was obtained from the CPSC-51100: Statistical Programming under the guidance of Dr. Gina Martinez Assistant Professor, Computer and Mathematical Sciences @ Lewis University. In which a similar seventy-five-row data set was provided in order to facilitate the comparison with the training data set. Additionally, the two-dimensional Euclidean distance formula can be represented as follows [1]:

$$dist(x,y) = \sqrt{(sl_x - sl_y)^2 + (sw_x - sw_y)^2 + (pl_x - pl_y)^2 + (pw_x - pw_y)^2}$$

In this case the X and Y axis will represent the training data point and testing data points. Additionally, the "L" and "W" initial represent the length and width of either sepal and petal of the Iris flowers.

Furthermore, the programing language Python was utilized for this study. The data set was divided into four different arrays. As represented in the Euclidean distance formula. The sepal length and width of the training data points were stores in a 2-dimensional array. The similar steps were also taken for the training data set. However, in order to do an accuracy test for each point result; the labels for each data set then stored in a one-dimensional array.

Thus far, as represented in the Euclidean distance formula. Th e distance between the training to the testing are measure for e ach point. This will leave out an array with a dimension of sev enty-five by seventy-five.

III. Analysis

The seventy-five by seventy-five array must be narrow down to one by seventy-five array. To do so, the argument of the minimum is needed. The argument of the minimum is a mathematical operation that takes into consideration only the minimum point within an array. For our purposes, the minimum points will represent the highest likely hood of the representation of the type of flower the Nearest Neighbor Algorithm concludes to be.

However, given that both datasets were individually label; it is then possible to calculate the accuracy percentage based on the comparison of how many are a true match vs a false positive. This is done by first comparing the training label array to the testing label array and count how many times both contents of the array space are equal. Then, the result is divided by the total number of rows. In this case, comparing both datasets sets gives an accuracy rate of ninety-four percent.

IV. Conclusion

The result shows a very promising result for approximating the identity of either flower type. This is truly beneficial in a logistics environment. As the control dataset obtains from scientific research can potentially be the foundation in which products can be an identity without removing them from the containment. Additionally, if the data points were to include additional information such as Pedicel, is it then possible to increase the accuracy of our results. Last, although the Nearest Neighbor Algorithm shows promising results, let us not forget that the testing data set had to be within the same data range as the training dataset in order to properly use the Euclidean distance formula. Had there been a case in which the testing data set is bigger or smaller than the training data set; the adjustment for each case can potentially harm the accuracy results. The reason being is because we are adding and removing values to our dataset that will potentially create different results.

References

[1] Robinson, A. (2019). How to Calculate Euclidean Distance. [online] Sciencing.com. Available at: https://sciencing.com/how-to-calculate-euclidean-distance-12751761.html [Accessed 3 Oct. 2019].

[2] The Flower Expert. (2019). Iris Flower - Varieties and Types of Iris | TheFlowerExpert. [online] Available at: https://www.theflowerexpert.com/content/mostpopularflowers/morepopularflowers/iris [Accessed 3 Oct. 2019].

[3] R. A. FISHER, Sc.D., F.R.S, "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS," Annals of Eugenics, Pages 179-188. [online] Available at: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x [Accessed 03 Oct.. 2019].

[4] Martinez, G. (2019). CPSC-51100: Statistical Programming. [online] GitHub. Available at https://github.com/IzzyStyle/Statistical-Programming/tree/master/Assigments [Accessed 03 Oct.. 2019].