Applying the K-Nearest Neighbor Algorithm To Predict Breast Cancer

Gunkeet Mutiana
Computer Engineering
Montreal, Canada
gunkeet@gmail.com
40226566

Abstract- This paper documents and showcases the use of a KD-Tree using a K-Nearest-Neighbor algorithm to predict the diagnosis of a patient.

Introduction

Breast cancer is one of the most prevalent cancers to affect women all around the world. Consequently, the accurate diagnosis of their breast cancer can significantly increase the rate of survival. Machine learning has been a great help to doctors to predict breast cancer.

This paper documents the implementation and application of the K-Nearest Neighbor (kNN) algorithm using a KD-Tree data structure to predict breast cancer diagnosis. We use the Wisconsin Breast Cancer dataset to train and test the model. I used the first 10 elements of the dataset where the mean is calculated to predict the diagnoses.

Problem

The problem addressed in this paper is the diagnosis of breast cancer using the Wisconsin Breast Cancer dataset. This dataset has information about 569 patients with many different attributes. We used the elements that contain "mean" in the name to construct, train and test our machine learning model. The objective is to predict the diagnosis of a patient as either malignant or benign based on the 10 mean attributes.

To solve this issue, we implement a kNN algorithm using a KD-Tree data structure. The KD-Tree is used to store data in a multidimensional space which aids the nearest neighbor search. Whilst, the kNN algorithm finds 3 nearest value to our node and making a decision based on the majority of the 3 nearest neighbors found.

Description of model

The model used is the KD-tree that organizes points in a multidimensional space for retrieval. Each node in the KD-Tree represents a point with its attributes and diagnosis and the tree is built recursively. An example as to how the KD-Tree would look like:

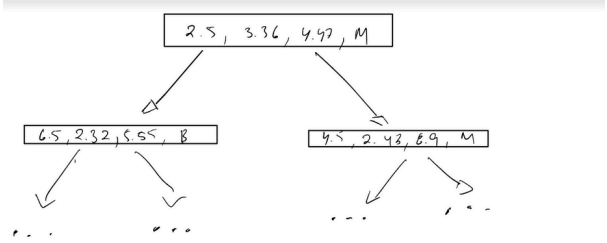

Figure 1: Example of the KD-Tree

The tree shown in figure 1 is an example, however the tree itself in the program will have 10 attributes as well as the diagnosis in the node. They will also contain a reference to the left child node and right child node.

In the kNN algorithm, it finds the k-nearest neighbors using the KD-Tree data structure. The following flowchart represents the way the program executes:

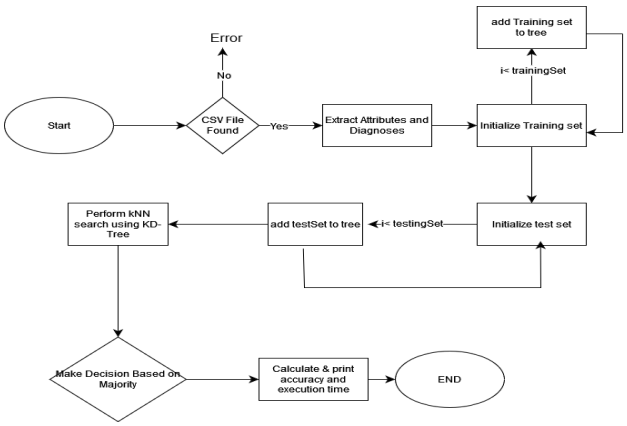

Figure 2: Flowchart of the program

Results

The following table represents the results of the program for different training sizes.

Training Size (N)	Accuracy	Execution time
100	88.0%	20 ms
200	96.0%	40ms
300	92.0 %	102ms
400	88.0%	34ms
500	Was not able to obtain	Was not able to obtain

Data training size (N) of 500 was not possible since the data set might not have enough data for it. Each of these were done using a nearest neighbor k of 3. As you can see, as the training size increases, the execution time increases as well the accuracy. Running the programs multiple times, the accuracy and execution time reduces at 400 training size. A possible source of error for this might be bad logic used in my code

Conclusion

I was successfully able to apply the KD-tree and kNN algorithm to predict breast cancer. However, there is a potential flaw in my code where its inaccuracies increase as the training size increases.