COMP 551: Applied Machine Learning Report for Project 1

Zijian Pei, Xiaoxiao Shang, Yueng Zhang

February 10, 2022

Contents

1 Abstraction

This report reviews what we accomplished for the first project of course COMP551. The content of this project is applying K-Nearest Neighbors and Decision Trees, which are two practical and powerful algorithms for machine learning, on two data-sets from the Machine Learning Repository of University of California Irvine.

The experimental results show that Decision Trees Algorithm achieved slightly better accuracy than K-Nearest Neighbors Algorithm if computation complexity is not considered as a key factor. Various hyper-parameters and diverse cost/distance functions are also used in the comparison of the accuracy of different models. The experiment reflects that a decent choice of K for K-Nearest Neighbors Algorithm is around 5;the best choice of max depth for Decision Trees Algorithm ranges from 5 to 8 for [4]Hepatitis Data Set and 16 to 23 for [3]Diabetic Retinopathy Debrecen Data Set. For the distance functions, all distance functions fit the Hepatitis Data Set well, while Manhattan Distance and Minkowski Distance are better for Diabetics Retinopathy Debrecen Data Set considered on the same K value. For the cost function, Gini Index is slightly better but some randomness remains. Eventually, we observed vital features of both data sets; decision boundary plots associated with plausible hyper-parameter values for both models are generated as well.

2 Introduction

[5]Hepatitis means inflammation of the liver. It is often caused by a virus, but other factors can be trigger. It can be short-term infection or long-term infection. Signs of hepatitis can include fatigue, nausea, dark urine, join pain and so on. Among five types of infectious hepatitis, hepatitis A, hepatitis B, hepatitis C are most common. For noninfectious ones, one of the major causes are heavy alcohol use. In our data set, there are 80 complete data entries with each having 19 attributes and 1 class variable. The conclusion is that [4]ALBUMIN and ASCITES are important features. The highest accuracy is reached by Decision Trees Model with depth around 5 for all three cost functions.

[6]Diabetic Retinopathy is a diabetes complication that affects eyes. Symptoms may not show until the condition progresses. The major signs include spots or dark string floating in one's vision, blurred vision, vision loss and so on. In our experiment, there are 1151 complete data entries with each having 19 features and 1 class variable. The conclusion is that [2]number of Mas found at confidence levels of 0.5 and 0.6 are important features. The highest accuracy is reached by Decision Trees Model with depth 16 and Gini index as the cost function.

3 Data sets

Hepatitis Data Set and Diabetic Retinopathy Debrecen Data Set both contain numerical and binary features, which are covariates that need to be examined to build the models. The binary features of Hepatitis Data Set are either 1 or 2 whereas those of Diabetic Retinopathy Debrecen Data Set are either 0 or 1. The range of each numerical feature varies. Meanwhile, both data sets contain a class variable: *Die* and *Live* in Hepatitis data set, and *Signs of DR* and *No Signs of DR* in Diabetic Retinopathy Debrecen Data Set. The class variable is the one needs to be predicted by the models.

We cleaned the data by removing rows with missing values. We generated a correlation diagram and chose two variables with the highest correlation with the class variable for training. The number of complete data entries we can use in Hepatitis Data Set is 80 and in Diabetic Retinopathy Debrecen Data Set is 1151. The discrepancy in numbers of data sets affect the performances of K-Nearest Neighbors and Decision trees to some extent. We later split the two data sets into training data sets and test data sets. For Hepatitis Data Set, 60 data entries are considered as the training data set, and the rest data entries are test data set. For Diabetic Retinopathy Debrecen Data Set, 800 data entries are considered as the training data set, and the rest data entries are test data set. In the extra investigation part, we also analyzed the effects on performance of the models when the ratio of the test set with respect to the whole data set varies.

4 Result

4.1 Experiment 1: Compare the Accuracy of K-Nearest Neighbors Algorithm and Decision Trees Algorithm

Model	Minimum (Testing)	Maximum (Testing)
KNN	K = 15 (0.80)	K = 5 (0.95)
Decision Tree	Depth = 1 (0.80)	Depth = 5 (0.95)

Table 1: Hepatitis Data Set: Minimum and Maximum Testing Accuracy for K-Nearest Neighbors and Decision Trees

Model	Minimum (Testing)	Maximum (Testing)
KNN	K = 51 (0.647)	K = 8 (0.755)
Decision Tree	Depth = 3 (0.638)	Depth = 23 (0.795)

Table 2: Diabetic Retinopathy Debrecen Data Set: Minimum and Maximum Testing accuracy for K-Nearest Neighbors and Decision Trees

For Hepatitis Data set, the minimum and maximum accuracy is consistent regardless of the model types. Similarly, for the Retinopathy Debrecen Data Set, the minimum and maximum accuracy is close regardless of the model types.

4.2 Experiment 2: Different K values for K-Nearest Neighbors

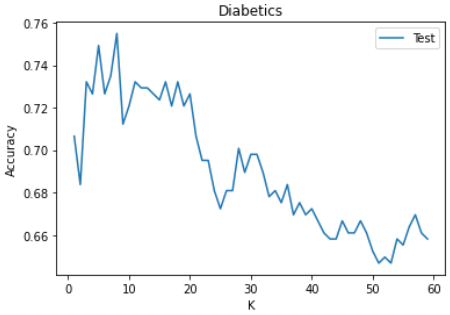

Fig. 1. Diabetic Retinopathy Debrecen: Test Data Accuracy for Different K Values

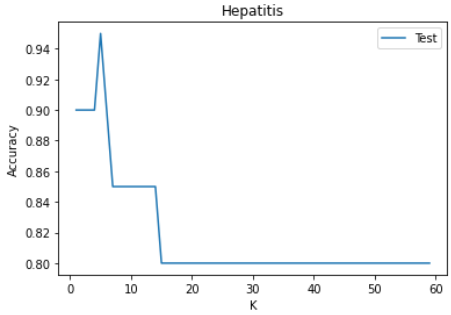

Fig. 2. Hepatitis: Test Data Accuracy for Different K Values

In the second experiment, the curve of the test accuracy initially has a dramatic wiggle because the model is overfitting when t is small; the curve becomes more steady as k increases. Meanwhile, the accuracy decreases gradually and becomes relatively flat as k becomes large. Meanwhile, a extreme large k value should yield an underfitting model.

4.3 Experiment 3: Different Maximum Tree Depth for Decision Trees

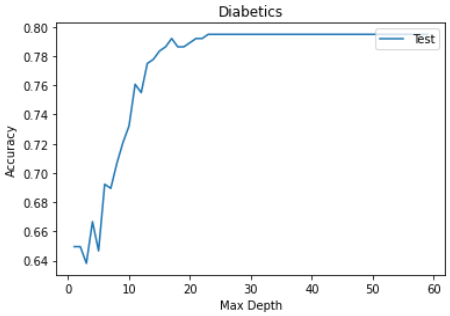

Fig. 3. Diabetic Retinopathy Debrecen: Test Data Accuracy for Max Depth

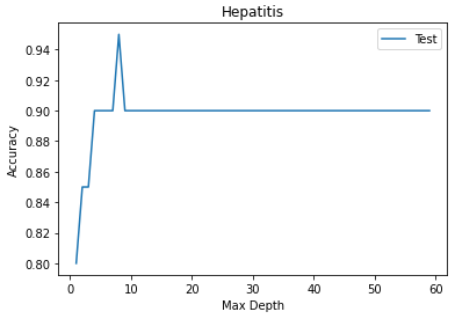

Fig. 4. Hepatitis: Test Data Accuracy for Max Depth

In the third experiment, we can tell that the test accuracy increases sharply when depth is small and remains constant when it is large.

4.4 Experiment 4: Different Distance/Cost Functions

For the K-Nearest Neighbors model, the overall shape for the Euclidean, Manhatton and Minkowski distance functions given different k values are similar as there is no discernible differences among the three lines on both data sets. These three distance functions are all preferable to be implemented on Hepatitis Data Set, whereas Minkowski distance is the most optimal on Diabetic Retinopathy Debrecen Data Set. On the other hand, for the Decision Trees model, we found that the Gini Index cost function has the highest accuracy among different cost functions on both data sets.

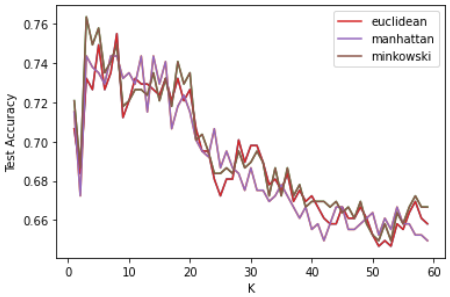

Fig. 5. Diabetic Retinopathy Debrecen: Test Data Accuracy vs K values with Different Distance functions

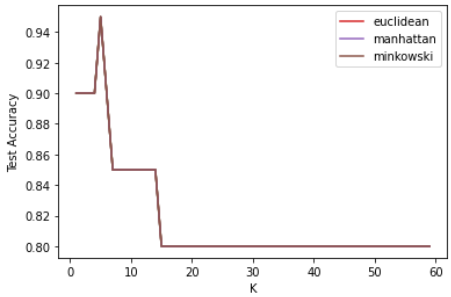

Fig. 6. Hepatitis: Test Data Accuracy vs K values with Different Distance functions

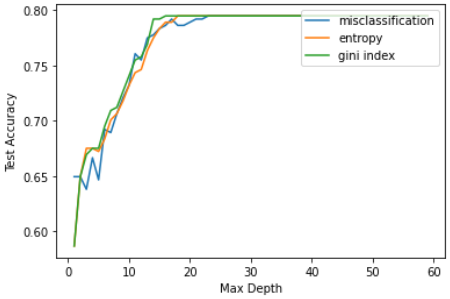

Fig. 7. Diabetic Retinopathy Debrecen: Test Data Accuracy vs Depth with Different Cost functions

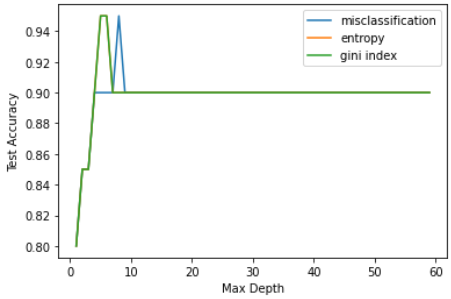

Fig. 8. Hepatitis: Test Data Accuracy vs Depth with Different Cost functions

4.5 Experiment 5: Decision Boundary

We selected two features for each data set to evaluate the behavior of models for simplicity, [3]numbers of Mas found at the confidence level 0.5 and 0.6 for Diabetic Retinopathy Debrecen Data Set and [4]ALBUMIN and ASCITES for Hepatitis Data Set. We chose the plausible hyper-parameter values to display the plots.

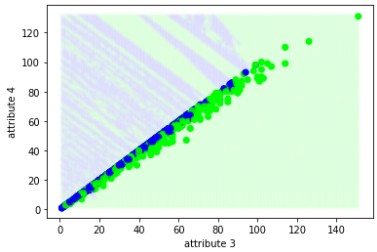

Fig. 9. Decision Boundary for KNN,K= 1 (Diabetics)

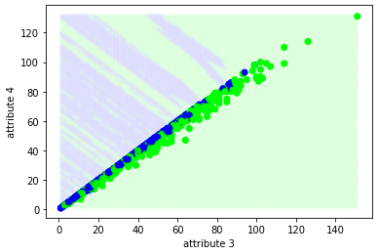

Fig. 10. Decision Boundary for KNN,K= 6 (Diabetics)

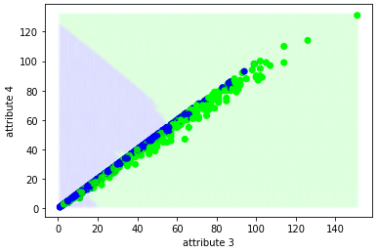

Fig. 11. Decision Boundary for KNN,K= 40 (Diabetics)

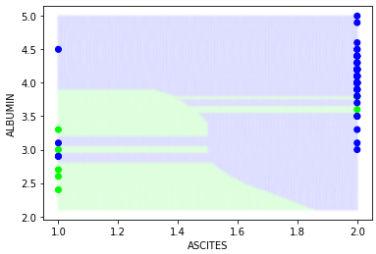

Fig. 12. Decision Boundary for KNN,K= 1 (Hepatitis)

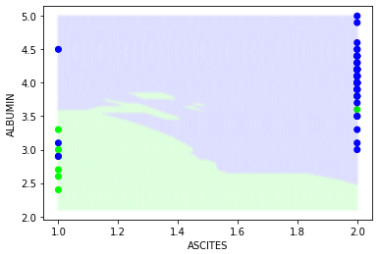

Fig. 13. Decision Boundary for KNN,K= 6 (Hepatitis)

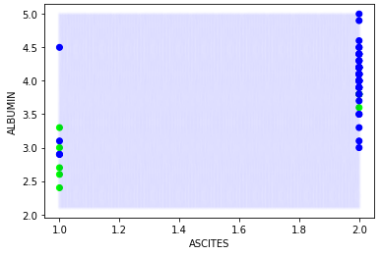

Fig. 14. Decision Boundary for KNN,K= 40 (Hepatitis)

In general, for the Hepatitis Data Set, the decision boundary plot of K-Nearest Neighbor Algorithm initially is some green lines that span the entire graph from left to right, which clearly indicates the model is too specific as there is a dividing line at almost every blue-green junction points. However, when the K value is too large, for example 40, the plot whole region in the plot becomes glue, the model then becomes too generalized.

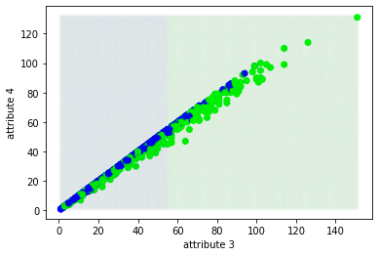

Fig. 15. Decision Boundary for Decision Trees,depth = 1 (Diabetics)

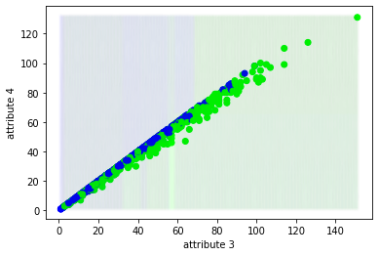

Fig. 16. Decision Boundary for Decision Trees,depth = 5 (Diabetics))

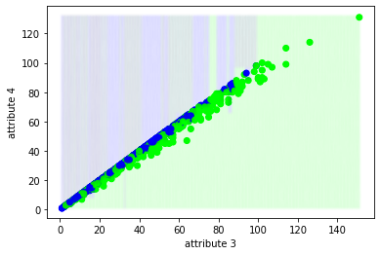

Fig. 17. Decision Boundary for Decision Trees,depth = 59 (Diabetics)

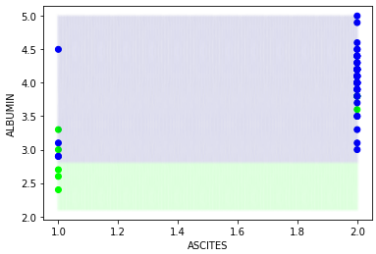

Fig. 18. Decision boundary for Decision Tree,depth = 1 (Hepatitis)

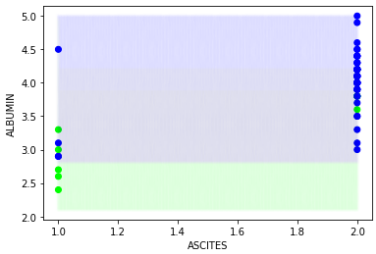

Fig. 19. Decision boundary for Decision Tree,depth = 5 (Hepatitis)

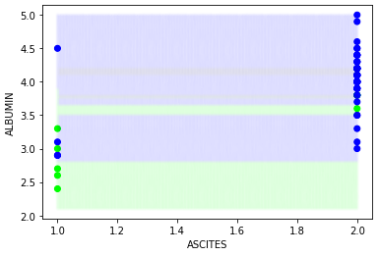

Fig. 20. Decision boundary for Decision Tree,depth = 59 (Hepatitis)

The decision boundary plot of Decision Trees Algorithm initially is a well-divided rectangles, which means the model is too generalized. However, as the depth value increases to a large value, there should exist green lines, which means the model is over-fitting. For the Diabetics data set, the decision boundary plot of K-Nearest Neighbor Algorithm initially contains some lines as the other data set, and becomes vague as the value of K increases. The decision boundary plot of Decision Trees Algorithm initially has clear boundaries, and as the depth value increases, the whole region becomes blue, which means the model is over-fitting.

4.6 Experiment 6: Extra Experiment

The purpose of extra experiment is to find the relationship between test accuracy and test set's ratio with respect to the whole data set. We found that there is negative correlation between test accuracy and test set's ratio. When k is small, the accuracy plummets. As k increases, the curve of accuracy fluctuates within a small range for a wide range of k values and then decreases sharply. It is worth noting that the accuracy drops drastically if the ratio is extremely low or high because of insufficient data.

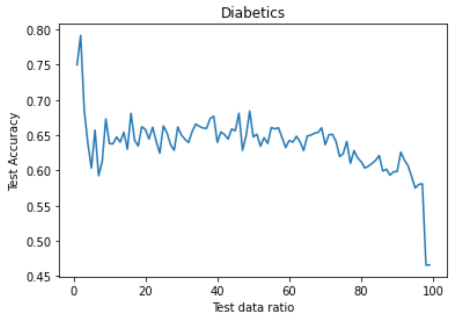

Fig. 21. relationship between test data ratio and accuracy for Diabetics

5 Discussion and Conclusion

Firstly, we run hyper-parameters from 1 to 60 on both Decision Trees and K-Nearest Neighbors models in order to find the best model with the best accuracy. In general, Decision Trees and K-Nearest Neighbors are both preferable because the maximum and minimum accuracy on both data sets with different models are either close or the same. We studied the behavior of K-Nearest Neighbors and Decision Trees models separately in the following experiments.

We test different k values and investigated how the hyper-parameters affect the test data accuracy for each data set. We cannot pick very small or very big k, as extreme k values should decrease the accuracy. In conclusion, the appropriate k should be around 8 for the Diabetic Retinopathy Debrecen Data Set and 6 for the Hepatitis Data Set.

Similarly, in the experiment 3, we should choose depth around 23-60 for the Hepatitis Data Set and 8 for Diabetic Retinopathy Debrecen Data Set respectively. We found that the results are sometimes totally different for different distributions of data if the training set and the test set are randomly shuffled. For experiment 4, we tried many times on different distributions of data and the shapes of lines vary dramatically. Thus we cannot easily derive a conclusion solely based on this.

6 Contributions

Zijian Pei: Decision Trees Algorithm Implementation, Data Cleaning, Correlation Computation, data analysis, Report Writing
Xiaoxiao Shang: Experiments Replication, Data Analysis, Decision Boundary Plots Generation, Report Writing
Yuteng Zhang: Experiments Replication, K-Nearest Neighbors Algorithm Implementation, Data Analysis, Report Writing
Code for building [2] K-Nearest Neighbors and [1] Decision Tree Algorithms references from tutorial

References

[1] Teaching Assistant. *Decision Tree*. Website. https://colab.research.google.com/github/mravanba/comp551-notebooks/blob/master/DecisionTree.ipynb#scrollTo=9czgntjHFFTg. 2021.

[2] Teaching Assistant. *K-Nearest Neighbors*. Website. https://colab.research.google.com/drive/1y8B-yAkP57GaTf3wuQySwkDlcDtxyUbI?usp=sharing. 2021.

[3] University of California Irvine. *Diabetic Retinopathy Debrecen Data Set Data Set*. Website. https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set. 2014.

[4] University of California Irvine. *Hepatitis Data Set*. Website. https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set. 1988.

[5] CDC. *What is Viral Hepatitis*. Website. https://www.cdc.gov/hepatitis/abc/index.htm#:~:text=Hepatitis%20means%20inflammation%20of%20the,medical%20conditions%20can%20cause%20hepatitis.. 2021.

[6] Mayo Clinic. *Diabetic retinopathy*. Website. https://www.mayoclinic.org/diseases-conditions/diabetic-retinopathy/symptoms-causes/syc-20371611#:~:text=Diabetic%20retinopathy%20(die%2Duh%2D,or%20only%20mild%20vision%20problems.. 2021.