Exploring Association between Optimal Depth and Number of Features used in Decision Tree

[1]Sunanda Mulik, [2]Prasad Gokhale, [3]Poonam Sawant
[1]Assistant Professor, [2]Associate Professor, [3]Assistant Professor,
[1]Department of Computer Science,
[1]Vishwakarma University Pune, India.

Abstract : **Decision Tree (DT) is a widely used predictive modelling tool that has applications spanning a wide range of areas including business, energy modelling, medicine and remote sensing. Decision Tree is a non-parametric supervised learning method used for both classification and regression tasks. Building optimal decision tree has always been an NP problem. A number of factors affect the accuracy of decision trees and a lot of research has been done in this area. In the present work we attempted to explore a relationship between number of features used for modelling and optimal depth of decision tree. Different datasets were tested and it was observed that there exists some kind of relationship between these two. The work may help reduce the testing time significantly by setting an upper bound on depth value during testing for optimal depth.**

IndexTerms **- Decision tree, predictive modelling, optimal depth, accuracy.**

I. INTRODUCTION

Decision Tree is a widely used predictive modelling tool that has applications spanning a wide range of areas including business, energy modelling, medicine and remote sensing (Sharma, H., Kumar S., 2016). It is one of the most popular and practical methods for supervised learning. Decision Tree is a non-parametric method used for both classification and regression tasks. DTs use divide and conquer strategy where at each step it splits a data set based on different attribute conditions and limits the search space. Each internal node contains a split criteria and each leaf node contains a class label (Pujari, A., 2001). The length of the maximum sequence of nodes from root node to the leaf is called the depth of tree.

Optimal decision tree is a well-researched problem in data mining. DTs are heuristic in nature. The accuracy of DT depends on factors like depth of the tree, number of samples considered to split a node, number of samples considered in leaf node, features used in prediction etc. Optimal values of these factors have significant effect on the accuracy of the algorithm. A lot of research has been done in attempt to find optimal DT. A novel approach is proposed for inferring an optimal DT with a minimum depth based on the incremental generation of Boolean formulas (Avellaneda F., 2020). PyDL8.5, a Python library is introduced to infer depth-constrained Optimal Decision Trees (ODTs) (Aglin, G., Nijssen, S., and Schaus, P., 2020).

Our work focuses on optimal depth of the tree. It is known that the lower values of depth lead to under fitting whereas larger values lead to over fitting. So it is crucial task to find optimal value of tree depth that leads to acceptable level of accuracy. It takes significant amount of time to test the tree for different depth values. In this work we tried to explore if there exists any relationship between number of features used in the model and optimal depth of decision tree.

II. MATERIAL AND METHOD

We have considered three datasets here. DT classification algorithm is tested for all datasets for different depth values keeping other parameters constant. Each experiment is repeated hundred times and classification accuracies are measured. Average accuracies for different depth values are computed and compared for each experiment.

Platform Used:
Operating System: Windows 10
Python Package: Anaconda3

Experiment 1: Binary Classification Task- Liver Patient Dataset

Table 1: Description of Liver Patient Dataset

Dataset Size	No. of features	Target	Type of features
(583,11)	10	1	Mix

The dataset contains 10 features. After checking for multi-collinearity, seven features, namely, age, gender, DB, alkphos, sgpt, ALB, A_G are selected. Class is the target variable that can either have a value 1 or 2.
Here we built decision trees for depth ranging from one to three times the number of features i.e. one to twenty one.

Experiment 2: Binary Classification Task- Diabetes Dataset

Table 2: Description of Diabetes Dataset

Dataset Size	No. of Features	Target	Type of Features
(768,9)	8	1	Continuous

Features include Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age. Target variable is 'Outcome' that indicates disease status of the person. It can take one of the two values 1 and 0 (whether the person is suffering from disease or not).
DTs are built for depths ranging from one to two times the number of features i.e. 1 to 16.

Experiment 3: Binary classification task- SECOM dataset

Table 3: Description of SECOM Dataset

Dataset Size	No. of features	Target	Type of features
(1567,592)	591	1	Continuous

The dataset contains 591 features out of which 20 highly correlated features are selected. DTs for depth ranging from one to two times the number of features (1 to 40) are built.

III. Results and Discussion

Results of the three experiments are given below. The findings and the limitations of the work are discussed and further research directions are given.

3.1 Results of Experiment 1

Figure 1 shows Accuracy Graph for different values of depth in experiment1. Figure 2 shows the average accuracy obtained from Figure 1.

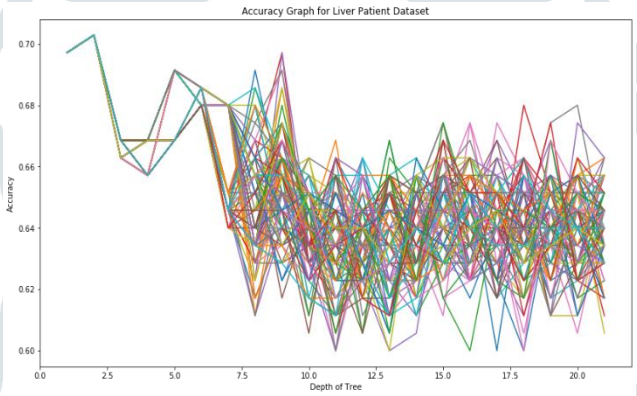

Figure 1: Accuracy Graph for liver patient dataset for different depth values

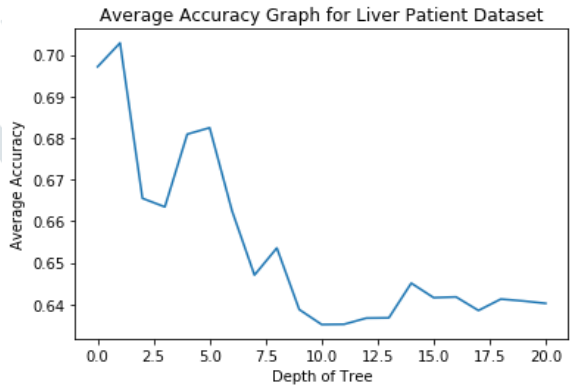

Figure 2: Average Accuracy Graph for liver patient dataset for different depth values

From Figure 2 it is seen that the maximum average accuracy is obtained at depth two which is far less than half the number of features used.

3.2 Results of Experiment 2

Figure 3 shows Accuracy Graph for different values of depth parameter in experiment 2.

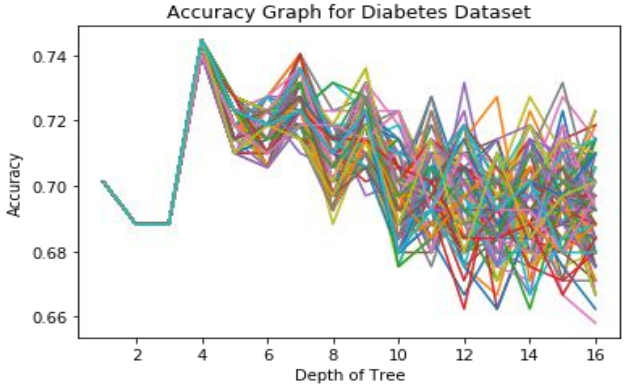

Figure 3: Accuracy Graph for liver patient dataset for different depth values

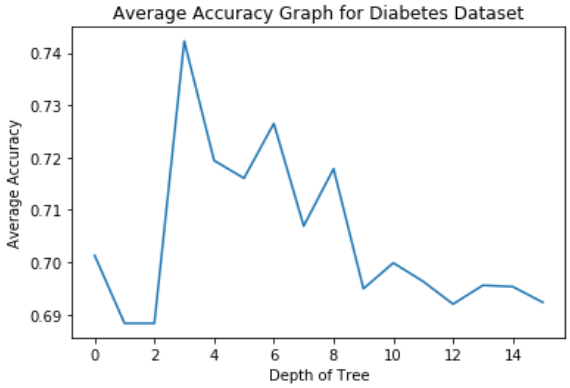

Figure 4: Average Accuracy Graph for liver patient dataset for different depth values

From Figure 4 it is seen that the maximum average accuracy is obtained at depth three which is less than half the number of features used.

3.3. Results of Experiment 3

Figure 5 shows accuracy graph for different depth levels for SECOM data classification task. Average accuracy is shown in Figure 6. The maximum average accuracy is obtained at depth one only.

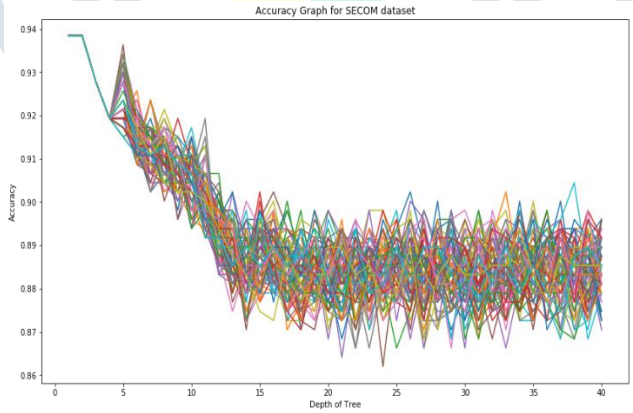

Figure 5: Accuracy Graph for SECOM dataset for different depth values

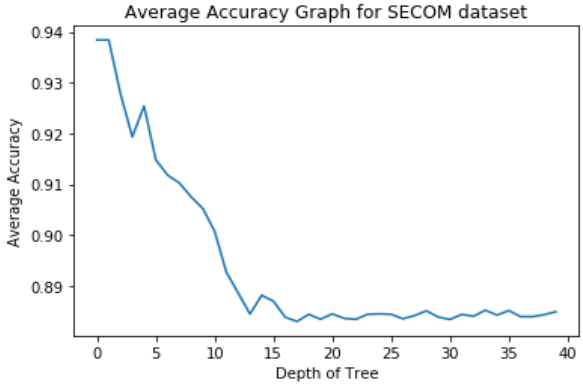

Figure 6: Average Accuracy Graph for SECOM dataset for different depth values

3.4 Findings

From the results, it can be seen that we can get optimal depth for decision tree at values less or equal to the half of the number of features used in the model.

Optimal depth <= (½) number of features used

IV. Limitations and future Scope

In this work we tried to discover relationship between number of features used in the model and the optimal value for depth of the tree. Main limitation of the present work is that we have only considered classification tress and not regression trees. The work can be extended for regression trees and results validated.

The research can further be extended in various directions. We can explore the effect of number of attribute values on optimal depth. Also it can be tested for multiclass problem to check the effect of number of classes on depth value. We can also check the effect of sample class distribution on the depth. Impact of data types of features on the depth value can be examined as well.

V. Conclusions

Often, a considerable amount of time is spent on testing for optimal depth for decision tree. The results here show that there is a relationship between number of features and optimal depth value. The optimal depth can be found at values less than half the number of features used in the model. This work may help set upper bound on the depth value for testing and thus reduce the testing time significantly.

VI. Acknowledgment

We sincerely thank to Vishwakarma University Pune for its constant support and encouragement. We thank Dr.Vaishali Bhosale for the intellectual discussions we had from time to time that aid in the research process. Finally, our deepest gratitude to our family that treasured our hard work and help us keep the spirit during tough times.

References

[1] Sharma, H., Kumar S. 2016. A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR), 5(4), 2094-2097

[2] Avellaneda, F. 2020 Efficient Inference of Optimal Decision Trees. Association for the Advancement of Artificial Intelligence. Available electronically at http://florent.avellaneda.free.fr/dl/AAAI20.pdf,

[3] Aglin, G., Nijssen, S., and Schaus, P. 2020 PyDL8.5: a Library for Learning Optimal Decision Trees. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), Demonstrations Track, 5222-5224

[4] Verwer, S., Zhang, Y. 2019 Learning optimal classification trees using a binary linear program formulation. Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 1625–1632

[5] Bertsimas, D., and Dunn, J. 2017 Optimal classification trees. Machine Learning 106(7), 1039–1082

[6] Sieling, D. 2008 Minimization of decision trees is hard to approximate. Journal of Computer and System Sciences 74(3), 394–403,

[7] Witten, I., Frank, E., Hall, M., and Pal, C. 2016 Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann,

[8] Mittal, K., Khanduja, D., Tewari, P. 2017. An Insight into "Decision Tree Analysis. World Wide Journal of Multidisciplinary Research and Development, 3(12), 111-115

[9] Pujari, A. 2001 Data Mining Techniques, Universities Press.