DECISION TREE PRUNING METHOD USING DELAYED SAMPLING

Sergei Mitrofanov

Eugene Semenkin

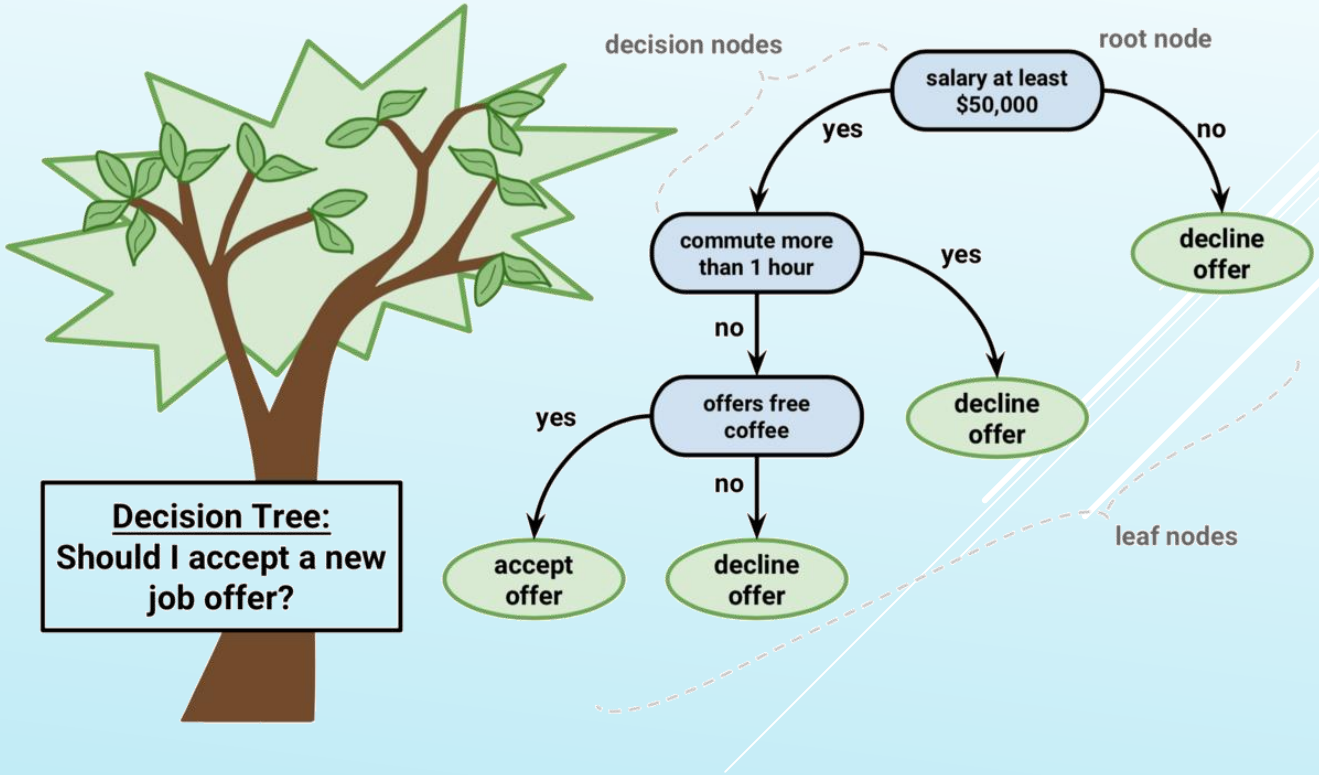
International Conference on Information Technologies (InfoTech), 2024

INTRODUCTION

One of the popular data analysis methods is the decision tree, which effectively classifies and predicts data.

However, decision trees have drawbacks, such as overfitting and difficulty in interpretation.

To resolve these problems, a decision tree pruning method using delayed sampling has been proposed.



DECISION TREE LEARNING ALGORITHMS

Decision tree learning algorithm is a method that is used to build a classification or regression model from data.

The main steps of the decision tree learning algorithm are:

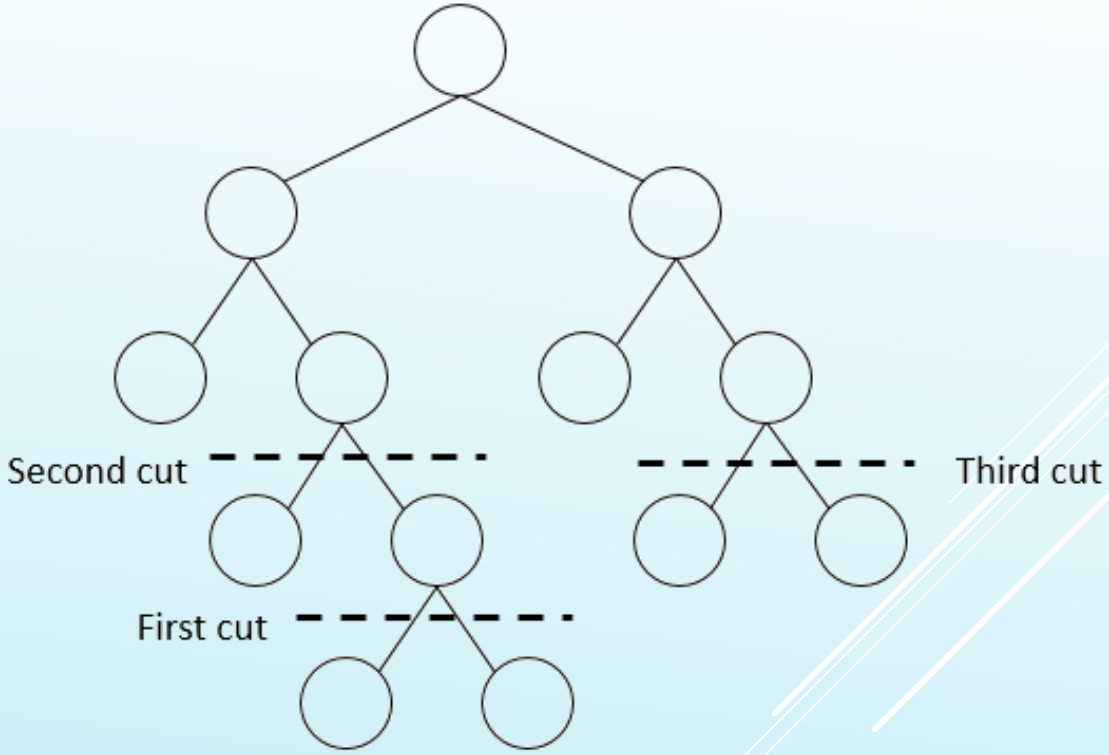
1. Attribute selection.
2. Partitioning.
3. Recursion.
4. Tree construction.

PRUNING DECISION TREES WITH A HOLDOUT SAMPLE

In the proposed approach, the original data set is divided into three parts: training set, delayed pruning set, and test set.

Steps:

1. A decision tree is built on the training sample using the "Measure of Separation" and "Differential Evolution" algorithms. The tree is built until each node contains objects of only one class.
2. Using a delayed sample, the resulting decision tree is evaluated.
3. The algorithm then looks for the longest branch of the decision tree and prunes the two leaf nodes at the end of that branch. When the root node is transformed into a leaf node, the label of the class whose objects were more numerous in the trimmed nodes is fixed in it.
4. The resulting tree is then evaluated again using delayed sampling. If the efficiency of the tree on delayed sampling has increased, then the process of pruning the longest branches continues.
5. As soon as the efficiency of the decision tree decreases due to pruning, the last pruning is canceled, and the resulting tree is considered to be complete.



Tree pruning process.

CLASSIFICATION PROBLEMS

The tasks are taken from the repository. The tasks are selected in such a way as to cover various subject areas.

Task number	Name	Sample size	Number of attributes	Number of target classes
1	Type of car determining	470	18	4
2	Speaker accent recognition	329	12	6
3	Type of cityscape determination	675	147	9
4	Recognizing the stage of hepatitis C	615	12	5
5	Iris variety recognition	150	4	3
6	Parkinson's disease recognition	756	754	2
7	Determination of the need for preventive maintenance of equipment	10000	5	2
8	Defining images by segments	2310	19	7
9	Heart defect recognition	270	13	2
10	Soil type recognition from satellite imagery	6435	36	6
11	Biodegradable chemicals recognition	1055	41	2
12	Workgroup classification by productivity in a garment factory	1197	13	9

RESULTS: EFFICIENCY OF ALGORITHMS

The efficiency of the algorithms was compared as the proportion of correctly classified objects in the test sample.

Task number	Without pruning	With pruning
1	0.692	0.728
2	0.541	0.603
3	0.742	0.773
4	0.879	0.912
5	0.949	0.962
6	0.733	0.779
7	0.902	0.913
8	0.93	0.947
9	0.763	0.806
10	0.831	0.839
11	0.792	0.816
12	0.425	0.451

RESULTS: DEPTH OF DECISION TREES

The interpretability of decision trees is improved, if their depth and, accordingly, the number of rules within the tree, are reduced. This makes trees easier to analyze and interpret, that is especially important in areas such as medicine, finance and others where informed decisions need to be made based on data.

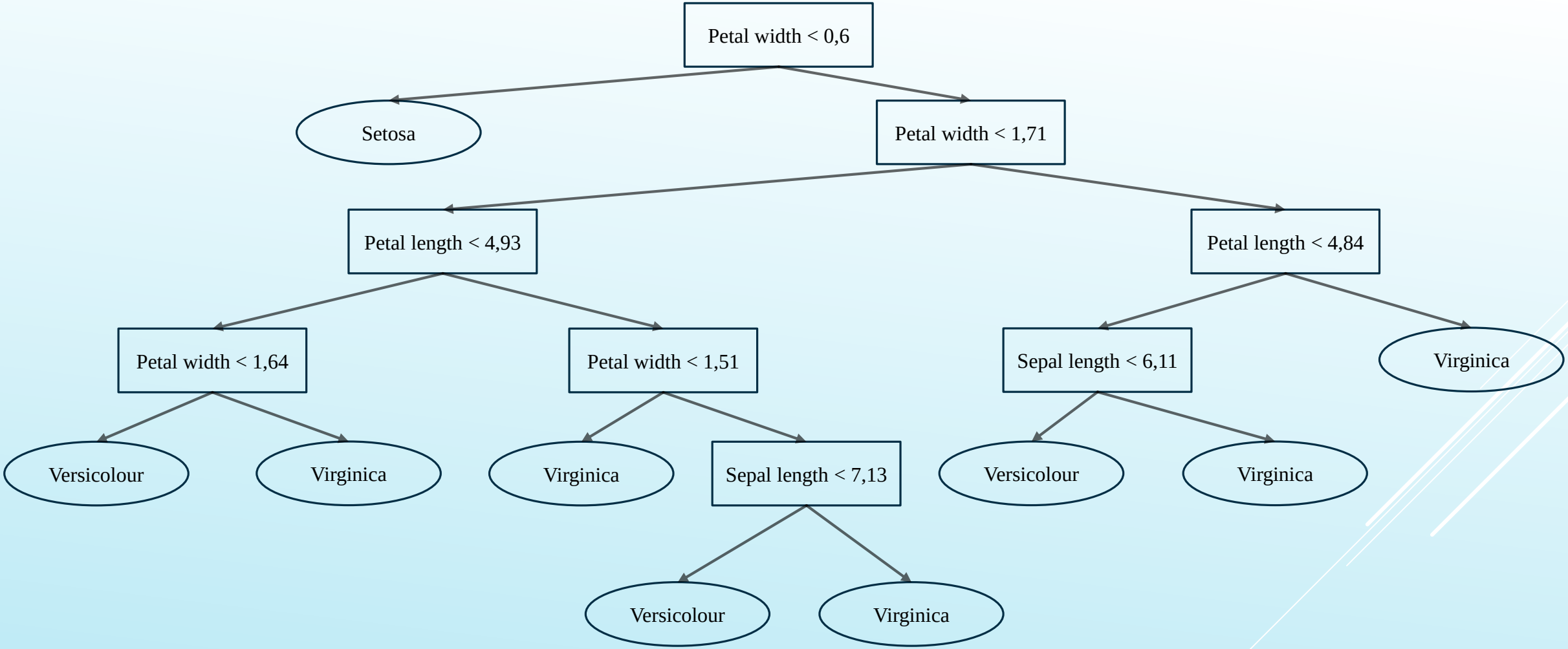
Task number	Without pruning	With pruning
1	13	11
2	11	9
3	11	10
4	9	7
5	6	3
6	9	8
7	17	13
8	13	12
9	9	8
10	17	16
11	13	11
12	15	13

RESULTS: NUMBER OF RULES IN DECISION TREES

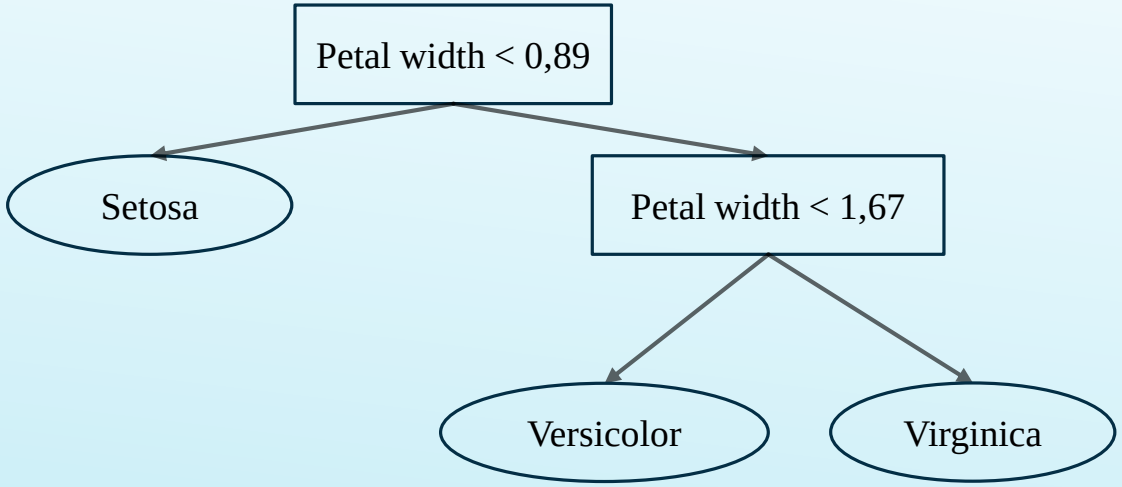
The interpretability of decision trees improves, since their depth and, accordingly, the number of rules within the tree are reduced. This makes trees easier to analyze and interpret, that is especially important in areas such as medicine, finance and others where informed decisions need to be made based on data.

Task number	Without pruning	With pruning
1	68	60
2	48	43
3	60	56
4	29	23
5	7	2
6	36	34
7	158	128
8	61	58
9	34	29
10	401	397
11	96	92
12	345	327

RESULTS: EXAMPLE OF AN *UNPRUNED* DECISION TREE FOR PROBLEM 5



RESULTS: EXAMPLE OF A *PRUNED* DECISION TREE FOR PROBLEM 5



RESULTS: STATISTICAL ANALYSIS, STUDENT T-TEST

Task number	T-TEST VALUES
1	**3.561**
2	1.589
3	**3.349**
4	2.07
5	1.087
6	**3.035**
7	0.148
8	1.085
9	**2.177**
10	1.009
11	1.699
12	1.708

BEST RESULTS

Task number	Unpruned / Pruned			
	Efficiency	Depth	Number of rules	T-Test
1	0.692 / 0.728	13 / 11	68 / 60	**3.561**
2	0.541 / 0.603	11 / 9	48 / 43	1.589
3	0.742 / 0.773	11 / 10	60 / 56	**3.349**
5	0.949 / 0.962	6 / 3	7 / 2	1.087
6	0.733 / 0.779	9 / 8	36 / 34	**3.035**
7	0.902 / 0.913	17 / 13	158 / 128	0.148
9	0.763 / 0.806	9 / 8	34 / 29	**2.177**

CONCLUSION

- The delayed sampling method allows to increase the accuracy and generalization ability of models, as well as prevent their overfitting.
- Despite the fact that the delayed sampling method has a number of advantages, its use cannot always provide an optimal solution to the problem.
- Overall, the delayed sampling method is a valuable tool for optimizing decision trees in machine learning.
- The interpretability of the obtained results is an important aspect of the application of the delayed sampling method in machine learning. In addition, the interpretability of the results allows stakeholders to better understand what factors influence the model's decision making.

THANKS FOR ATTENTION!