Национальный исследовательский университет “МЭИ”

Институт автоматики и вычислительной техники

Кафедра прикладной математики

## АНАЛИЗ ДАННЫХ

## Лабораторная работа №1

## Тема: «Алгоритм CART»

**Выполнила:**

Карпеева Виктория

Группа: А-13м-19

Москва, 2019

**Выполнение работы**

1. Реализовать алгоритм CART для построения дерева решений. Процесс решения должен отображаться на форме. На каждом этапе работы алгоритма вычисляется подходящий признак для разбиения на основе анализа impurity узлов. Необходимо показать выбранный признак и его значение impurity.

В результате выполнения данной лабораторной работы была реализована программа, позволяющая строить дерево решений для данных, содержащих числовые признаки.

Тестировать работу программы будем с использованием Iris Dataset.

В качестве тестовой выборки будем использовать первые 10 значений из каждого класса, а в качестве обучающей - оставшиеся 120 значений.

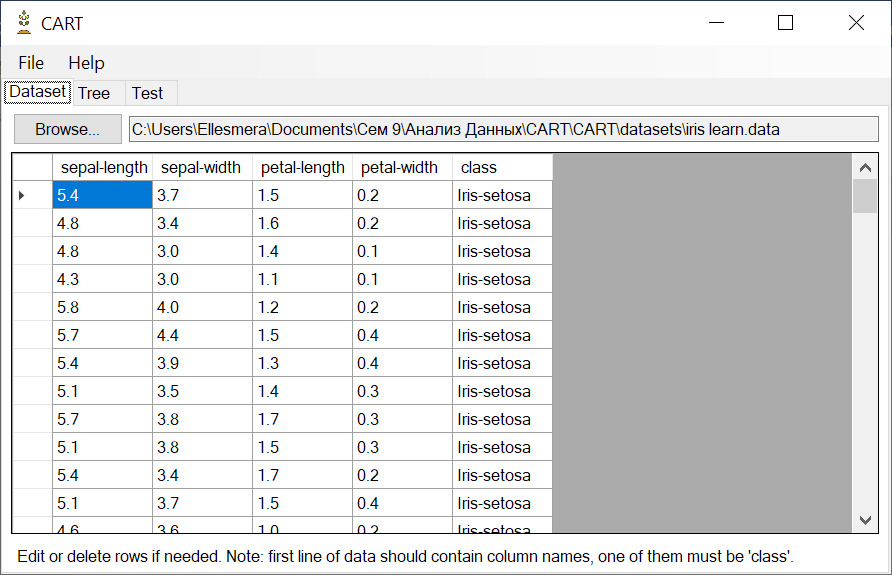


Рис. 1. Вкладка с исходными данными (обучающей выборкой)

1. Построить дерево решений для различного количества уровней и оценить точность классификации. Полученные результаты отобразить в таблице и на графике. Сделать вывод об оптимальном значении количества уровней дерева решений.

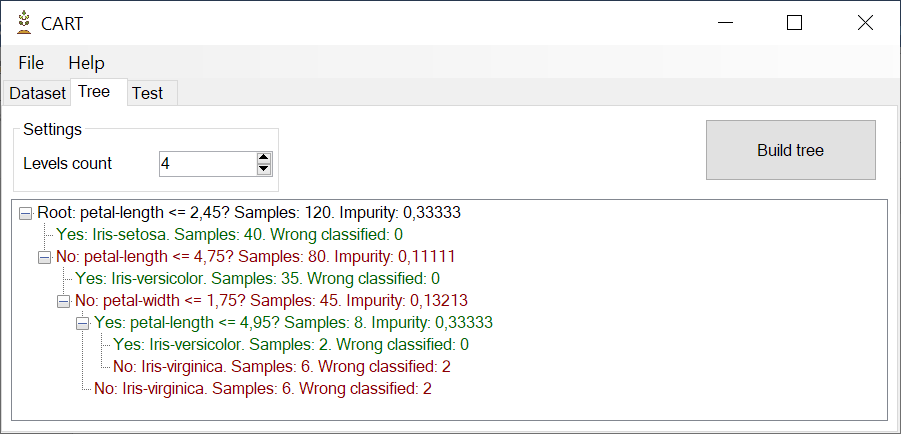


Рис. 2. Вкладка с построенным деревом

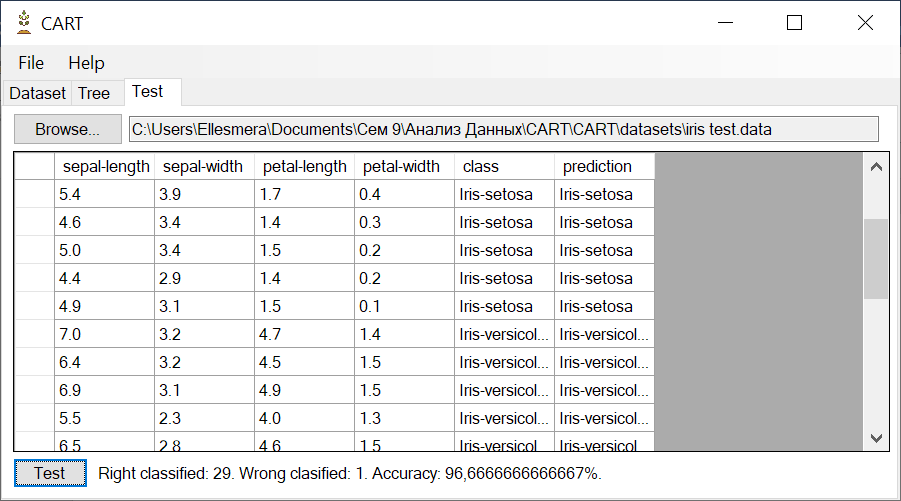


Рис. 3. Вкладка с результатом классификации тестовой выборки

Построим деревья для различных значений уровней и занесём данные в таблицу.   
Стоит заметить, что максимальная глубина, достигаемая на данной обучающей выборке, равна 6. При дальнейшем увеличении числа уровней, глубина дерева не увеличивается.

|  |  |
| --- | --- |
| Количество уровней | Точность, % |
| 1 | 66,6 |
| 2 | 93,3 |
| 3 | 93,3 |
| 4 | 96,6 |
| 5 | 96,6 |
| 6 | 96,6 |
| 7 | 96,6 |

Из таблицы можно сделать вывод, что оптимальным числом уровней будет являться число 4, так как дальше дерево усложняется, но точность классификации уже не повышается.

1. После построения основного дерева решений необходимо проверить его устойчивость. Для этого следует удалить некоторые записи из обучающего множества и проверить работу (качество классификации) дерева решений на той же тестовой выборке. Выполнить данную процедуру три раза. Сформировать таблицу, на каждой строке которой будут содержаться данные об общем количестве записей в обучающей выборке, количестве объектов в тестовой выборке, количестве правильно классифицированных и количестве неправильно классифицированных объектов. Сравнить результаты.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | Размер обучающей выборки | Размер тестовой выборки | Число правильно классифицированных объектов | Число правильно классифицированных объектов | Точность |
| 1 | 96 | 30 | 29 | 1 | 96,6% |
| 2 | 88 | 30 | 29 | 1 | 96,6% |
| 3 | 41 | 30 | 27 | 3 | 90% |

Анализируя полученные данные, можно сделать вывод о том, что полученное дерево достаточно устойчиво, так как для значительного снижения точности классификации пришлось удалить из обучающей выборки почти 2/3 значений.