Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(национальный исследовательский университет)
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

ОТЧЁТ ПО ЗАДАНИЮ №4

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Классификация с помощью дерева решений

Выполнил
студент группы КЭ-120
Глизница Максим Николаевич
E-mail: letadllo@mail.ru

Челябинск 2021

1. Задание

Выполните классификацию набора данных из задания 3 с помощью построения дерева решений, фиксируя критерий выбора атрибута разбиения (information gain, gain ratio, index gini) и варьируя соотношение мощностей обучающей и тестовой выборок (от 60%:40% до 90%:10% с шагом 10%). Выполните визуализацию построенных деревьев решений.

Вычислите показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера. Выполните визуализацию полученных результатов в виде диаграмм.

2. Краткие сведения о наборах данных

Использованный набор данных:

Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). Содержит информацию о произведённых партиях красного вина "Vinho Verde" (такую как pH и содержание хлоридов), а также оценку качества от 1 до 10 (реально в наборе данных присутствуют значения от 3 до 8). Как предлагается в пояснении к набору данных, высококачественным можно считать вино с оценкой 7 и выше. Таким образом, в наборе данных присутствуют 217 партий высокого качества и 1382 партий низкого качества.

3. Краткие сведения о средствах реализации

Для реализации методов была использована библиотека scikit-learn, включающая в себя множество алгоритмов для анализа данных.

Репозиторий по дисциплине: https://github.com/Airpllane/DAAlgorithms.

Каталог для задания: 4. DecTree.

4. Визуализация показателей качества

Для визуализации были использованы соотношения мощностей обучающей и тестовой выборки от 60%:40% до 90%:10% с шагом 10%. Был использован классификатор DecisionTreeClassifier, реализующий алгоритм классификации с помощью дерева решений. В качестве критерия выбора атрибута разбиения была выбрана энтропия. Для того, чтобы дерево было легче визуализировать, его глубина была ограничена 3 уровнями.

Была выполнена визуализация всех построенных деревьев. Результаты визуализации приведены на рис. 1.

Рис. 1. Результаты визуализации деревьев

Из рисунка можно увидеть, что все построенные деревья используют разные признаки. Однако среди них можно увидеть сходства: например, первым использованным признаком во всех деревьях является содержание алкоголя.

Далее была выполнена визуализация четырёх метрик качества на выбранных соотношениях. Результат визуализации приведён на рис. 2.

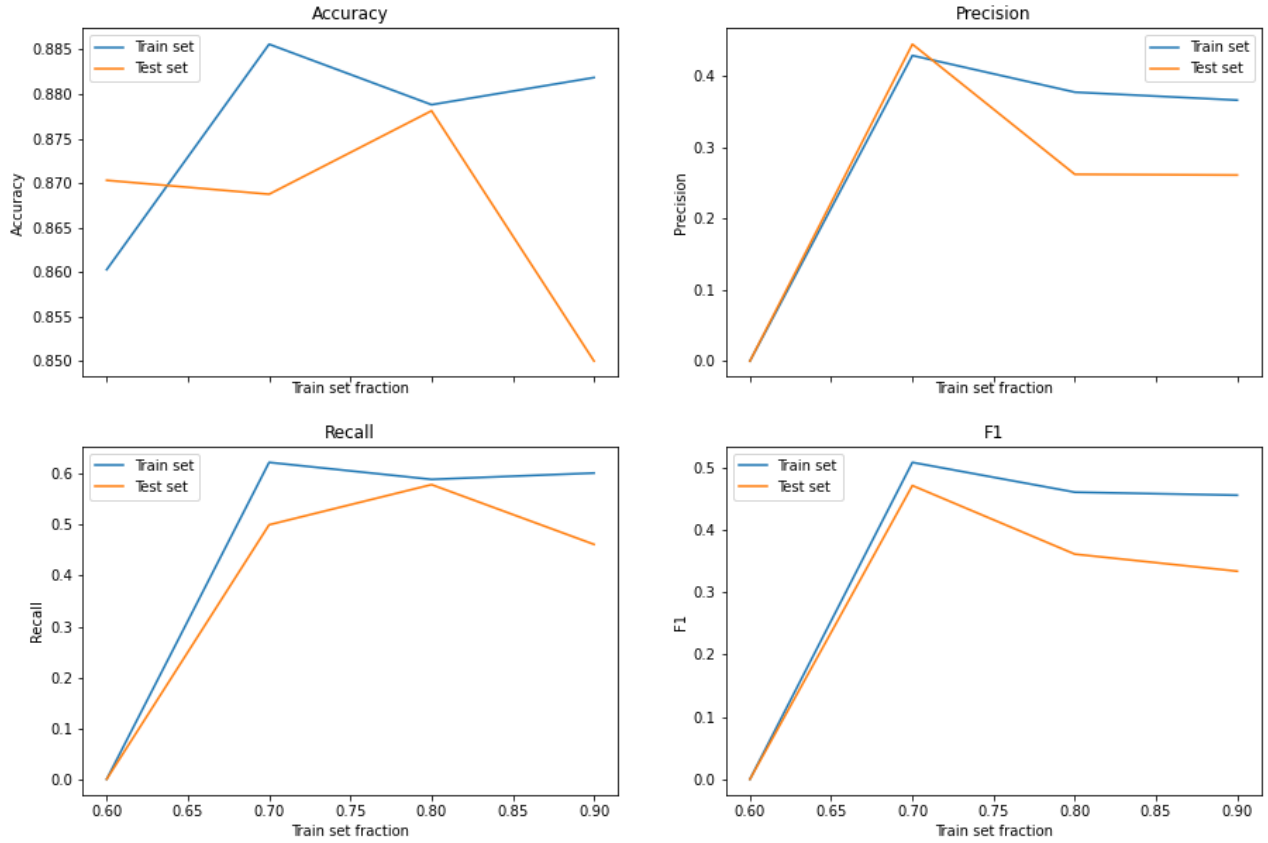

Рис. 2. Результаты визуализации метрик качества

Можно увидеть, что в соответствии с большинством метрик, наилучшим классификатором оказался построенный на разбиении 70%:30%.

Как и в прошлом задании, классификация должна стремиться уменьшить количество ложноотрицательных результатов (то есть низкокачественное вино не должно попасть в категорию высококачественного). Таким образом, наиболее важной метрикой является recall (полнота), которая зависит от количества ложноотрицательных результатов. На этот раз метрика имеет наибольшее значение среди всех, кроме точности, что говорит, что качество этого классификатора несколько выше, чем построенного в предыдущем задании.