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 №3

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Байесовская классификация

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

Челябинск 2021

1. Задание

Выполните классификацию набора данных с помощью Байесовской классификации, варьируя соотношение мощностей обучающей и тестовой выборок от 60%:40% до 90%:10% с шагом 5%.

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: 3. Bayes.

4. Визуализация показателей качества

Для визуализации были использованы соотношения мощностей обучающей и тестовой выборки от 60%:40% до 90%:10% с шагом 5%. Был использован классификатор GaussianNB, реализующий алгоритм наивной байесовской классификации с использованием нормального (Гауссовского) распределения вероятностей.

Была выполнена визуализация четырёх метрик качества на выбранных соотношениях. Результаты визуализации приведены на рис. 1.

Из метрик можно увидеть, что, в целом, качество классификации получилось более высоким в соответствии с метриками accuracy и precision, и менее высоким в соответствии с recall и f1.

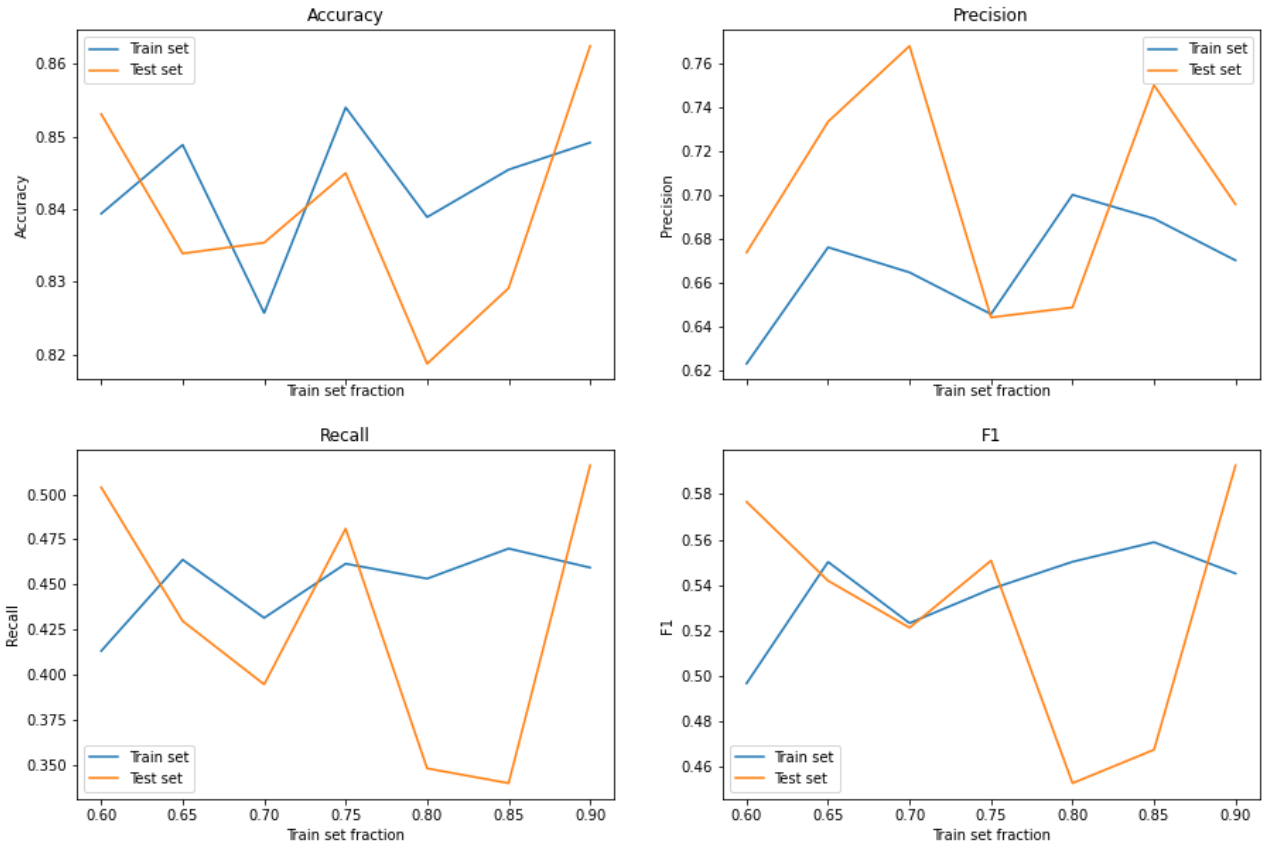

Рис. 1. Результаты визуализации

Поскольку использованный набор данных содержит данные о качестве вина, классификация должна стремиться уменьшить количество ложноотрицательных результатов (то есть низкокачественное вино не должно попасть в категорию высококачественного). Таким образом, наиболее важной метрикой можно считать recall (полноту), которая зависит от количества ложноотрицательных результатов. Эта метрика имеет наиболее низкие значения среди четырёх исследованных, что говорит о том, что качество классификации ниже, чем может показаться изначально.