Ансамблевая классификация

Кузьмичева Ольга, КЭ-403

Задание

1. Разработайте программу, которая выполняет классификацию заданного набора данных с помощью одной из техник ансамблевой классификации. Параметрами программы являются набор данных, ансамблевая техника (бэггинг, случайный лес или бустинг), количество участников ансамбля, а также параметры в соответствии с выбранной техникой ансамблевой классификации.
2. Проведите эксперименты на наборах данных из задания Классификация с помощью дерева решений, варьируя количество участников ансамбля (от 50 до 100 с шагом 10).
3. Выполните визуализацию полученных результатов в виде следующих диаграмм:
   - показатели качества классификации в зависимости от количества участников ансамбля для заданного набора данных; нанесите на диаграмму соответствующие значения, полученные в задании «Классификация с помощью дерева решений».
4. Подготовьте отчет о выполнении задания и загрузите отчет в формате PDF в систему. Отчет должен представлять собой связный и структурированный документ со следующими разделами:
   - формулировка задания;
   - гиперссылка на каталог репозитория с исходными текстами, наборами данных и др. сопутствующими материалами;
   - рисунки с результатами визуализации;
   - пояснения, раскрывающие смысл полученных результатов.

В соответствии с заданием была разработана программа, которая выполняет классификацию заданного набора данных с помощью одной из техник ансамблевой классификации. Для задачи классификации была использована библиотека sklearn. В репозитории размещено два файла: 4lab_classifications.ipynb – реализация классификации с помощью одной из техник ансамблевой классификации, grades.csv – используемый датасет.

Эксперимент

Для классификации была выбрана техника случайного леса. В результате классификации датасета (grades - сведения об оценках школьников за письменную контрольную работу), были получены показатели качества классификации для следующего количества участников ансамбля: 50, 60, 70, 80, 90, 100.

На рисунке 1 изображена таблица показателей качества наиболее оптимальной классификации (при количестве участников ансамбля = 60).

	precision	recall	f1-score	support
2	0.67	0.67	0.67	3
3	0.33	0.50	0.40	2
4	1.00	1.00	1.00	1
4-	1.00	0.50	0.67	2
5	1.00	0.50	0.67	2
5-	0.50	1.00	0.67	1
accuracy			0.64	11
macro avg	0.75	0.69	0.68	11
weighted avg	0.74	0.64	0.65	11

Рисунок 1 – Показатели качества классификации

Аккуратность (accuracy) обозначает долю данных, по которым классификатор принял правильное значение (в данном случае, удалось достичь значения 0.64). Макроусредненная оценка (macro avg) – это среднее арифметическое всех оценок F1 для каждого класса. Средневзвешенная оценка – это среднее значение всех оценок F1 для каждого класса с учетом поддержки каждого класса.

На рисунке 2 представлена визуализация показателей качества классификации в зависимости от количества участников ансамбля для заданного набора данных. Можно заметить, что наилучшие результаты достигаются при количестве участников ансамбля равном 60 и 90.

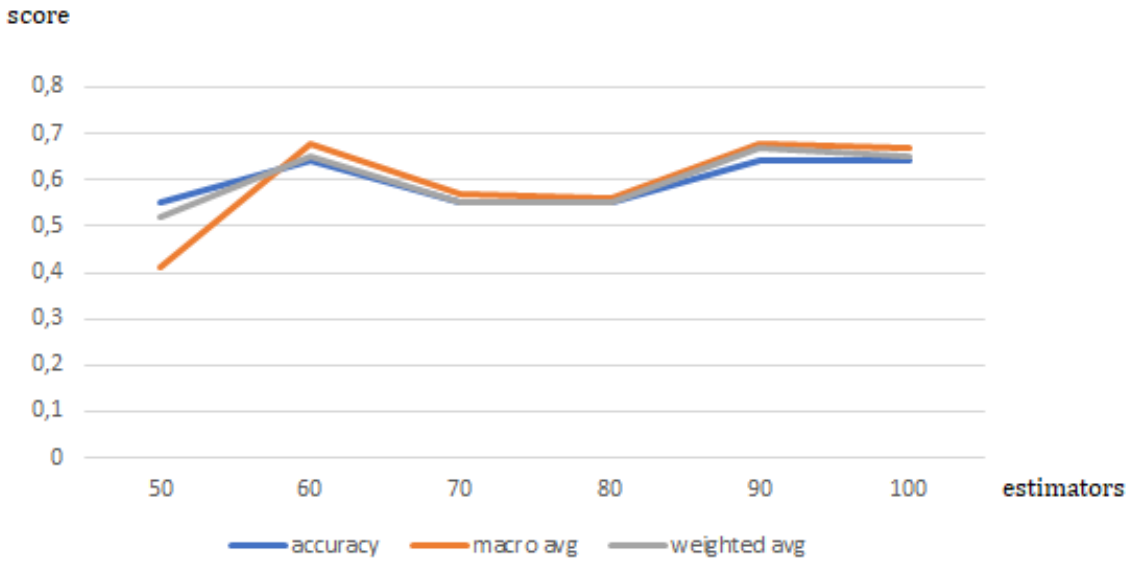

Рисунок 2 – Показатели качества классификации в зависимости от количества участников ансамбля

На рисунке 3 представлено сравнение лучшего результата ансамблевой классификации и лучшего результата классификации с помощью дерева решений. Можно заметить, что параметры точности ансамблевой классификации превосходят параметры точности классификации с помощью дерева решений.

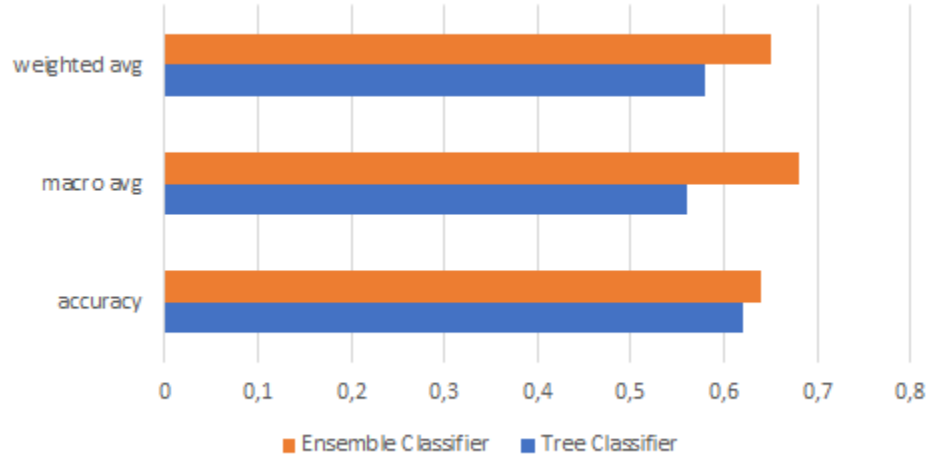

Рисунок 3 – Сравнение ансамблевой классификации и классификации с помощью дерева решений