

Министерство образования и науки Российской Федерации
Санкт-Петербургский государственный технический университет
Институт прикладной математики и механики
Кафедра «Телематика»

ЛАБОРАТОРНАЯ РАБОТА

ПО ТЕМЕ

«Алгоритмы композиции»

по направлению 02.04.01.02 «Организация и управление суперкомпьютерными системами»

Выполнил:

Студент гр. 13643.1 Титов А.И.

Проверил: Уткин Л.В.

Санкт-Петербург

2019

Оглавление

Постановка задачи	3
1 Набор данных «Vehicle»	4
2 Набор данных «Glass»	4
3 Использование KNN в алгоритме бустинга	5

Постановка задачи

1. Исследовать зависимость тестовой ошибки от количества деревьев в ансамбле для алгоритма `adaboost.M1` на наборе данных `Vehicle` из пакета `mlbench` (обучающая выборка должна состоять из 7/10 всех прецедентов, содержащихся в данном наборе данных). Построить график зависимости тестовой ошибки при числе деревьев, равном 1, 11, 21, . . . , 301, объясните полученные результаты.
2. Исследовать зависимость тестовой ошибки от количества деревьев в ансамбле для алгоритма `bagging` на наборе данных `Glass` из пакета `mlbench` (обучающая выборка должна состоять из 7/10 всех прецедентов, содержащихся в данном наборе данных). Построить график зависимости тестовой ошибки при числе деревьев, равном 1, 11, 21, . . . , 201, объясните полученные результаты.
3. Реализовать бустинг алгоритм с классификатором K ближайших соседей. Сравнить тестовую ошибку, полученную с использованием данного классификатора на наборах данных `Vehicle` и `Glass`, с тестовой ошибкой, полученной с использованием единичного дерева классификации.

1 Набор данных «Vehicle»

Был построен классификатор на основе алгоритма бустинга adaboost. Алгоритм был предложен Йоавом Фройндом и Робертом Шапире в 1996ом году. Алгоритм использует дерево решений в качестве классификатора. Был построен график ошибки в зависимости от количества использованных слоев классификаторов (рис. 1). Было установлено, что минимальная ошибка достигается при количестве деревьев: 31, 61, 111.

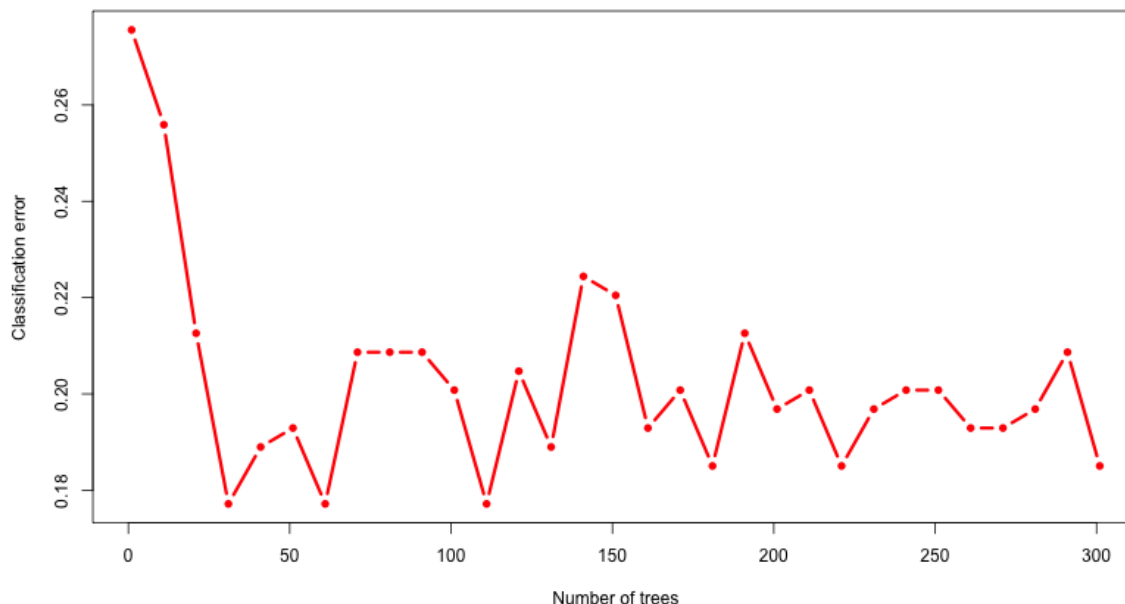


Рис. 1. Зависимость ошибки от количества использованных деревьев для набора «Vehicle»

2 Набор данных «Glass»

Был построен классификатор на основе алгоритма bagging (bootstrap bagging). Алгоритм был предложен Лео Брейманом в 1994ом году. Алгоритм также использует дерево решений в качестве классификатора, но подход использования иной в сравнении с adaboost. Был построен график ошибки в зависимости от количества использованных слоев классификаторов (рис. 2). Было установлено, что минимальная ошибка достигается при количестве деревьев: 41, 81, 101, 131, 141, 151, 161, 191.

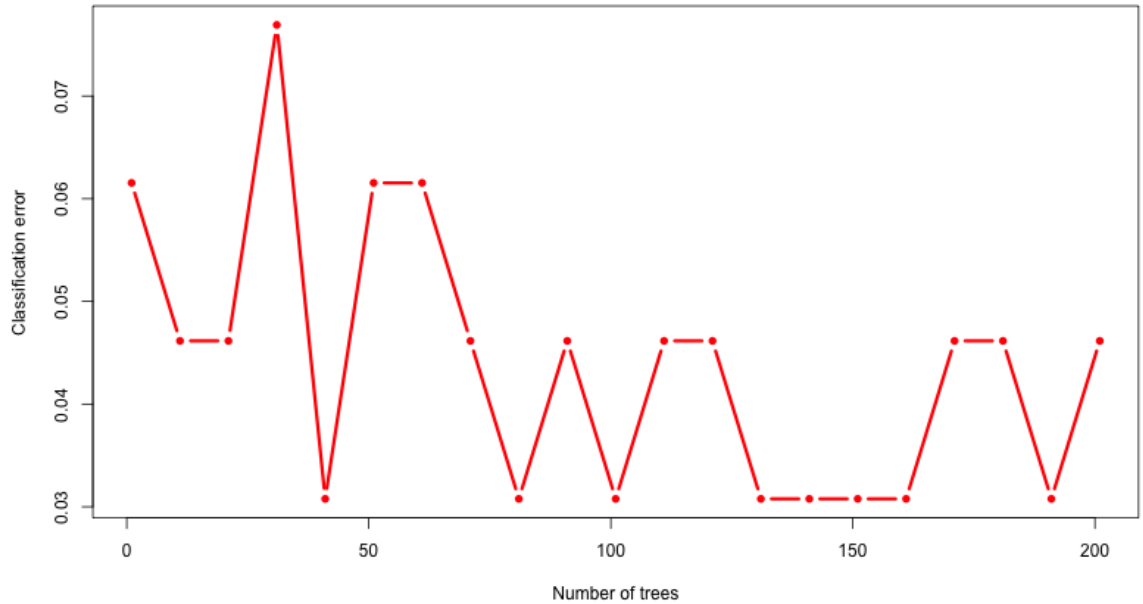


Рис. 2. Зависимость ошибки от количества использованных деревьев для набора «Glass»

3 Использование KNN в алгоритме бустинга

Задание ясно, но труднореализуемо по причине отсутствия готовых решений в библиотеках, таких как были в предыдущих заданиях («adaboost» и «bagging»). Существует реализация с помощью инструментария библиотеки «caret», но такое решение слишком ручное, поэтому очень трудозатратно и потребует много времени для реализации. Поэтому в рамках выполнения лабораторной работы были изучены варианты реализации задания, но само задание выполнено не было. Ко всему прочему - требуемый вывод был все же сделан. Применение расширенного с помощью бустинга алгоритма knn может привести к повышению точности результатов, но не всегда это имеет смысл, так как сам по себе подход последовательного ансамбля может оказаться весьма ресурсозатратным и дать не достаточный прирост точности, а в некоторых случаях даже проиграть по точности перед применением более простых методов.