А.О Шибаева, О.П. Солдатова

# ВЛИЯНИЕ ПАРАМЕТРОВ МЕТОДА СЛУЧАЙНЫХ ДЕРЕВЬЕВ НА ТОЧНОСТЬ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ В УСЛОВИЯХ МАЛОЙ ВЫБОРКИ

(Самарский университет)

Целью данной работы является изучение влияния варьирования значений параметров на точность решения задачи классификации при построении ансамбля случайных деревьев в условиях малой выборки.

В реальном мире получить достаточно большую выборку входных векторов не всегда представляется возможным в виду высокой стоимости или трудоёмкости получения этих значений. Кроме того, для качественного машинного обучения необходимы сбалансированные данные для каждого класса при решении задач классификации. Поэтому решение задач классификации в условиях малой выборки является сложной проблемой.

Одним из алгоритмов, дающих хорошие результаты при малой неравномерной зашумлённой выборке, является метод ансамбля решающих деревьев или метод случайных лесов [1].

Основная идея метода заключается в использовании большого количества решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но совокупное решение всех деревьев даёт хороший результат. Классификация объектов проводится путём голосования: каждое дерево ансамбля относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

В данной реализации метода были исследованы следующие параметры:

- количество решающих деревьев;
- размер подвыборки критериев из общего числа критериев с повторением;
- количество вариантов каждого из выбранных критериев;
- размер подвыборки векторов с повторением.

В качестве входных данных были использованы открытые данные о типах стекла, в зависимости от химического состава (содержания определённых химических элементов) и показателя преломления [2]. Количество атрибутов, включая класс – 10, всего типов стекла – 6, количество векторов – 214. Значения являются сильно коррелированными и сильно искажёнными. Для каждого класса выборки разного размера: от 9 векторов до 76 на класс.

На рисунке 1 показана зависимость точности классификации ансамбля на тестовой выборке от количества решающих деревьев в ансамбле. Из графика видно, что точность повышается в среднем до 75% правильных ответов при увеличении числа деревьев примерно до 40, дальнейшее увеличения размера ансамбля не даёт существенного прироста точности.

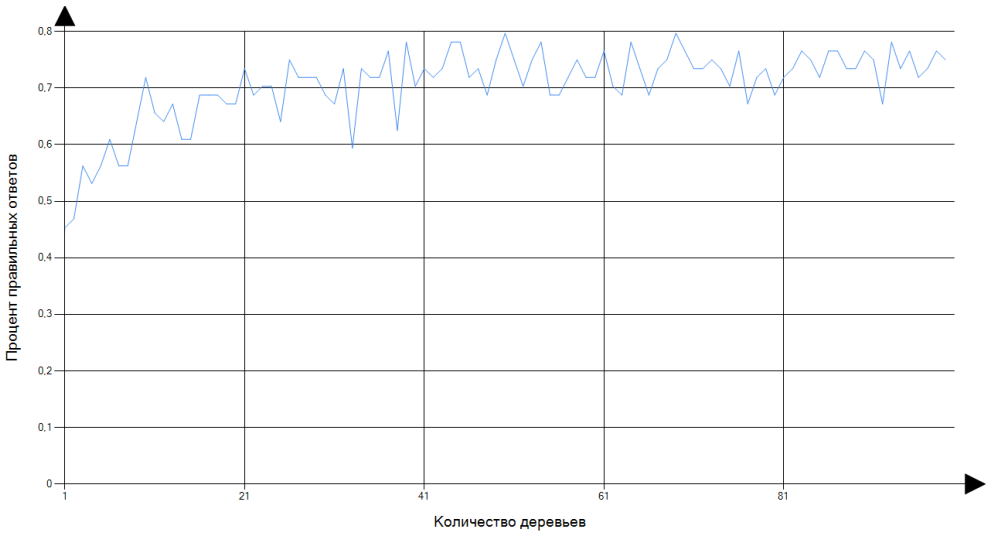


Рисунок 1 – График зависимости точности классификации
от количества решающих деревьев

На рисунке 2 показана зависимость точности классификации от количества критериев, среди которых происходит выбор наилучшего разбиения на конкретном шаге. Так как график представляет собой почти прямую, можно сделать вывод, что количество критериев не оказывает существенного влияния на точность. Однако при проверке в процессе работы программы было выяснено, что при увеличении количества критериев снижается скорость построения деревьев, поэтому рекомендуется использовать $\sqrt{n}$ критериев.

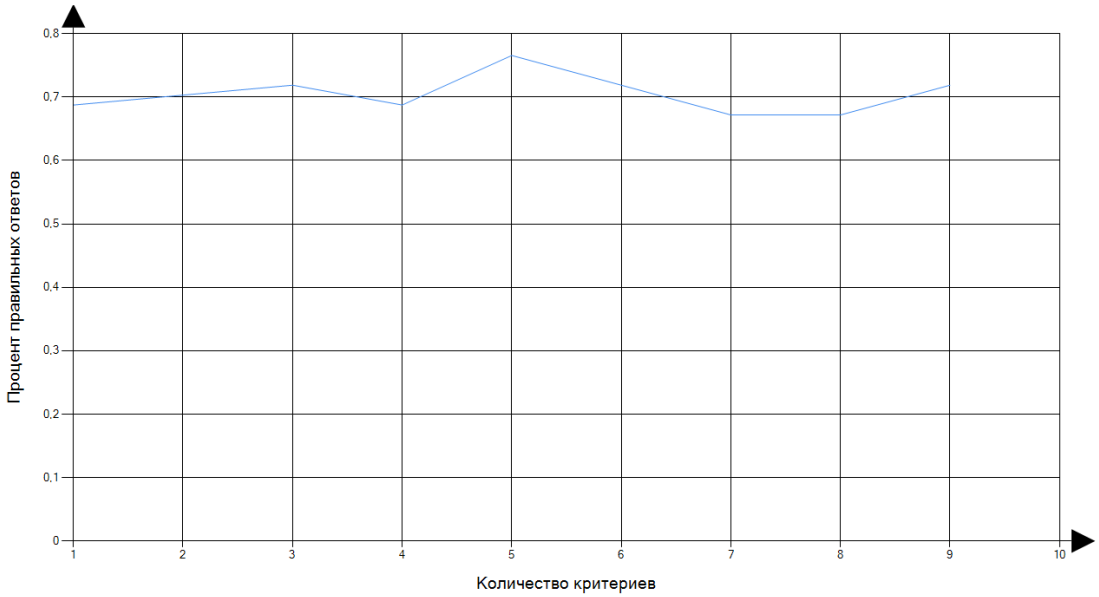


Рисунок 2 – График зависимости точности классификации
от количества критериев для разбиения

На рисунке 3 показан график зависимости точности классификации от количества вариантов каждого критерия, выбранного для разбиения. Так как среди вариантов выбирается наилучшее значение, при увеличении числа крите-

риев падает точность классификации из-за того, что наступает переобучение деревьев. Поэтому при сильно зашумлённых данных предпочтительнее указать как можно меньшее число вариантов.

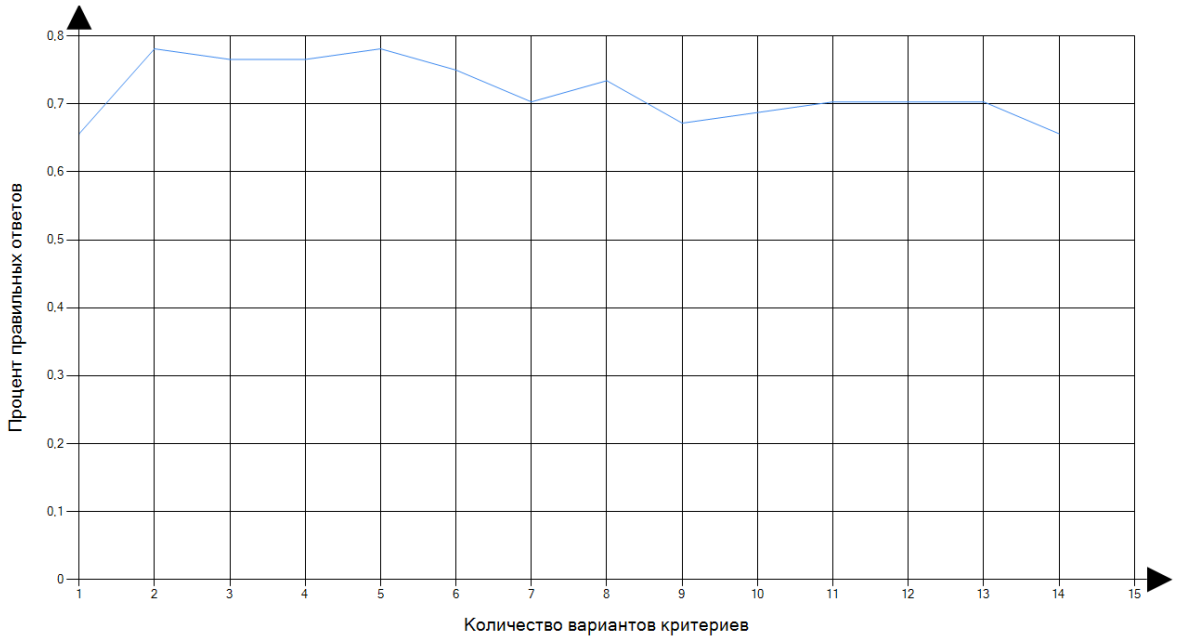


Рисунок 3 – График зависимости точности классификации от количества вариантов критериев

На рисунке 4 показан график зависимости точности классификации от размера подвыборки. Как видно из графика, точность возрастает до значения, равного размеру исходного пула векторов для обучения.

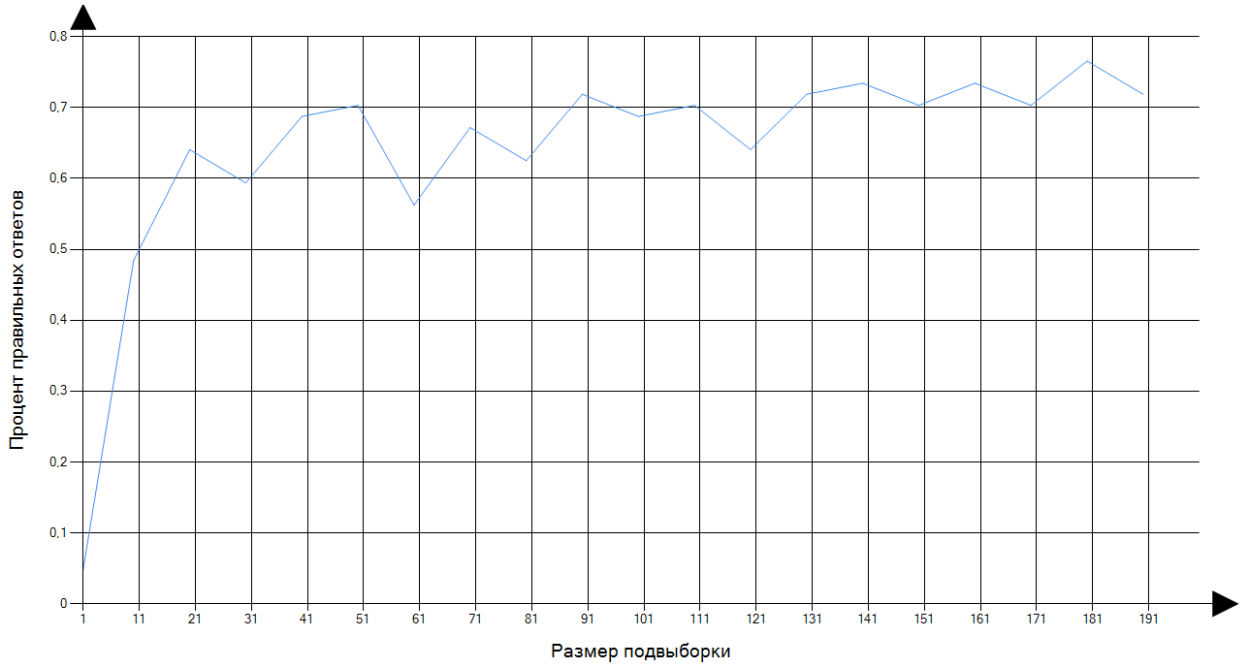


Рисунок 4 – График зависимости точности классификации от количества размера подвыборки

Учитывая выведенные закономерности, установим следующие значения параметров для построения ансамбля решающих деревьев для обучающей выборки типов стекла, содержащей 150 векторов.

На рисунках 5 и 6 приведены параметры построения ансамбля и результаты, полученные на тестовой выборке.

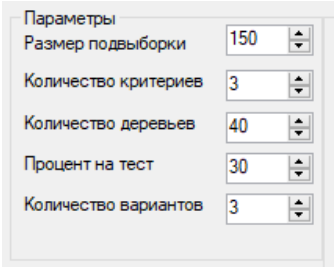


Рисунок 5 – Параметры построения ансамбля

Полученный результат в 75% правильных значений из случайно выбранной заранее тестовой выборки является хорошим для малого количества коррелированных искажённых исходных данных.

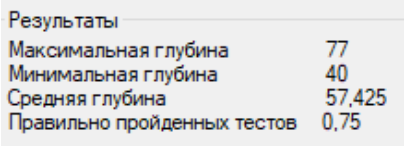


Рисунок 6 – Результаты, полученные на тестовой выборке

**Литература**

1. Breiman, L. «Random Forests» [Текст] / L.Breiman, Machine Learning // 2001. – №45 (1), – 5-32.
2. Kaggle [Электронный ресурс] – // URL: https://www.kaggle.com/.

И.М. Янников, М.В. Телегина

## ОРГАНИЗАЦИЯ БИОМОНИТОРИНГА ЛЕСНЫХ ЭКОСИСТЕМ С ПОМОЩЬЮ ЭКСПЕРТНОЙ СИСТЕМЫ

(ФБГОУ ВО «ИжГТУ имени М.Т. Калашникова», г. Ижевск, Россия)

Важнейшими задачами комплексного исследования загрязнения окружающей природной среды являются: установление источников и выявление пространственной структуры распределения очагов загрязнения, степени их интенсивности и оценка влияния на население [1].

Использование для контроля (мониторинга) за состоянием окружающей природной среды растений - биоиндикаторов является наиболее предпочтительным с точки зрения, как финансовых, так и временных затрат.