

**АНАЛИЗ ЭФФЕКТИВНОСТИ РАСПОЗНАВАНИЯ РУКОПИСНЫХ
СИМВОЛОВ НАИБОЛЕЕ ПОПУЛЯРНЫМИ МЕТОДАМИ
КОНТРОЛИРУЕМОГО МАШИННОГО ОБУЧЕНИЯ**
Кузенков Р.С. Email: Kuzenkov1136@scientifictext.ru

*Кузенков Роман Сергеевич – магистрант,
кафедра прикладной математики,
Смоленский государственный университет, г. Смоленск*

Аннотация: в статье рассмотрены наиболее распространенные методы классификации, используемые в машинном обучении: дерево принятия решений, логистическая регрессия, наивный байесовский классификатор, метод опорных векторов с линейным и квадратичным ядром, метод ансамблей. Проведен сравнительный анализ их эффективности при решении задачи распознавания цифр из набора данных MNIST. Построена кривая обучения указанным методам и сделаны выводы о скорости их обучения и величине обучающей выборки, необходимой для их обучения.

Ключевые слова: машинное обучение, классификация, mnist, эффективность, большие данные.

**ANALYSIS OF EFFICIENCY OF RECOGNITION OF MANUSCRIPT SYMBOLS BY
MOST POPULAR METHODS OF CONTROLLED MACHINE LEARNING**
Kuzenkov R.S.

*Kuzenkov Roman Sergeevich – Undergraduate,
DEPARTMENT OF APPLIED MATHEMATICS,
SMOLENSK STATE UNIVERSITY, SMOLENSK*

Abstract: the most common classification methods used in machine learning are considered: decision tree, logistic regression, naive Bayesian classifier, support vector machine with linear and quadratic kernel, ensemble method. A comparative analysis of their effectiveness in solving the problem of recognition of numbers from the MNIST data set is carried out. A learning curve for these methods is constructed and conclusions are drawn about the speed of their training and the amount of training sampling necessary for their training.

Keywords: machine learning, controlled learning, classification, MNIST, efficiency, big data.

УДК 519.254

Популярность искусственного интеллекта и машинного обучения неуклонно растет. Алгоритмы машинного обучения на сегодняшний день можно по праву считать наимоощнейшим инструментом прогнозирования, применимым в области Big Data. В данном исследовании приведен сравнительный анализ наиболее популярных методов контролируемого обучения [1]. Методы контролируемого обучения тренируются и выявляют закономерности на помеченных данных, а затем в результате обучения строится модель для осуществления прогнозирования на новом наборе данных. С помощью указанных методов решалась задача распознавания рукописных цифр из набора данных MNIST. Для решения поставленной задачи использовалась библиотека Scikit-learn языка программирования Python. Изображения с цифрами были взяты из набора данных Scikit-learn в нормализованном виде, а затем преобразованы в набор значений, представляющих оттенки серого цвета [2].

Из группы алгоритмов метода ансамблей был выбран алгоритм Random Forest («случайный лес»), наиболее подходящий для решения поставленной задачи.

Для каждого алгоритма была построена матрица несоответствий, строки которой представляют собой цифры, изображенные на рисунках из набора, а столбцы – цифры, предсказанные алгоритмом. Например, наивный классификатор 20 раз верно определил единицу и в четырех случаях ошибся в распознавании единицы, перепутав ее с двойкой (рис. 1).

	0	1	2	3	4	5	6	7	8	9
0	32	0	0	0	1	0	0	0	0	0
1	0	20	4	0	0	0	0	0	2	2
2	0	1	31	0	0	0	0	0	1	0
3	0	0	1	30	0	0	0	0	2	1
4	0	0	0	0	46	0	0	0	0	0
5	0	0	0	0	0	39	1	0	0	7
6	0	0	0	0	1	0	34	0	0	0
7	0	0	0	0	0	0	0	34	0	0
8	0	2	0	0	0	0	0	0	28	0
9	0	0	0	0	1	1	0	3	1	34

Рис. 1. Матрица несоответствий наивного классификатора Байеса

В дальнейшем был проведен анализ методов классификации при разном разбиении данных на тренировочные и тестовые с целью изучения скорости обучения и эффективности методов. Доля точных прогнозов при разном разбиении данных в соотношениях 4:1 и 1:1 представлена в таблицах 1 и 2.

Таблица 1. Результат классификации при доле тестовой выборки 20%

Метод классификации	Процент правильно предсказанных цифр
Метод опорных векторов с квадратичным ядром	98,3
Логистическая регрессия	96,4
Метод ансамблей	96,1
Метод опорных векторов с линейным ядром	96,1
Наивная байесовская классификация	91,1
Дерево принятия решений	85,3

Таблица 2. Результат классификации при доле тестовой выборки 50%

Метод классификации	Процент правильно предсказанных цифр
Метод опорных векторов с квадратичным ядром	98,2
Логистическая регрессия	94,9
Метод опорных векторов с линейным ядром	93,3
Метод ансамблей	92,4
Наивная байесовская классификация	89,4
Дерево принятия решений	83,9

На рисунке 2 приведена кривая обучения для всех методов, на основании которой можно сделать ряд выводов:

- в наибольшей степени от размера тренировочной выборки зависят метод ансамблей и дерево принятия решений;
- кривая обучения наивного байесовского классификатора является очень пологой, из чего можно сделать вывод о том, что данному алгоритму достаточно небольшой выборки данных для обучения;
- наибольший процент точных предсказаний имеет метод опорных векторов с квадратичным ядром;
- метод ансамблей, метод опорных векторов и логистическая регрессия показали результат выше среднего, при этом метод ансамблей требует большего объема тренировочных данных.

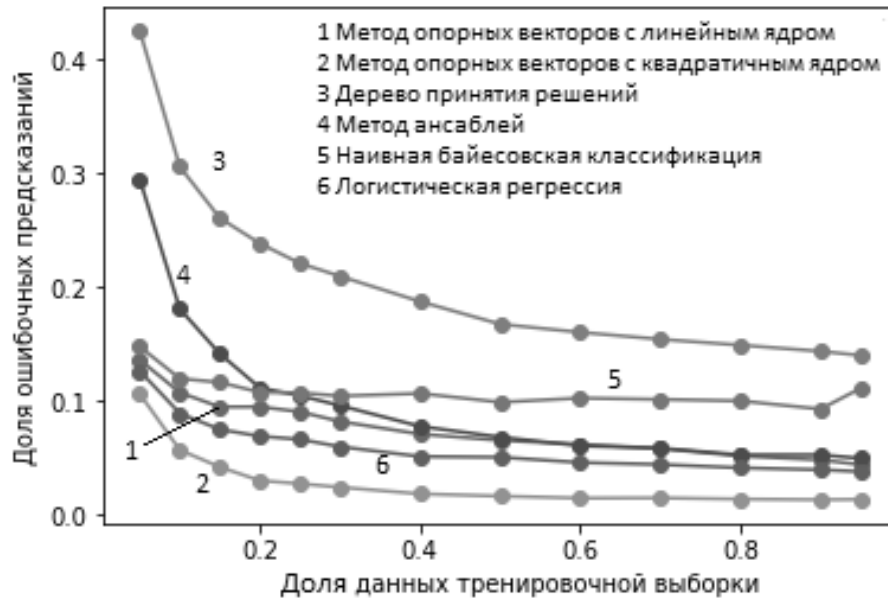


Рис. 2. Кривая обучения исследуемых методов

Список литературы / References

1. 10 главных алгоритмов машинного обучения. [Электронный ресурс]. Режим доступа: <http://ru.datasides.com/code/algorithms-machine-learning/> (дата обращения: 23.06.2017).
2. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. СПб.: Питер, 2017. 336 с.:ил. (Серия «Библиотека программиста»).