УДК 004.85

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ В СРЕДЕ WOLFRAM MATHEMATICA

Хлебородова Лидия Дмитриевна
студент
Научный руководитель: **Осипов Геннадий Сергеевич**
д.т.н., зав. кафедрой Информатики
ФГБОУ ВО «Сахалинский государственный университет»

Аннотация: Апробирована унифицированная методология решения задачи классификации сложных объектов на больших данных методами машинного обучения в среде Wolfram Mathematica. Проведен сравнительный анализ точности классификации 9 методами на примере базы данных по автомобилям, содержащей как числовую, так и символьную информацию.

Ключевые слова: задача классификации, методы машинного обучения, компьютерный анализ.

COMPARATIVE ANALYSIS OF MACHINE LEARNING METHODS FOR SOLVING THE CLASSIFICATION PROBLEM IN THE WOLFRAM MATHEMATICA ENVIRONMENT

Khleborodova Lidiia Dmitrievna
Osipov Gennady Sergeevich

Abstract: A unified methodology for solving the problem of classifying complex objects on big data using machine learning methods in the Wolfram Mathematica environment has been tested. A comparative analysis of the accuracy of classification by 9 methods is carried out on the example of a database on cars containing both numerical and symbolic information.

Key words: classification problem, machine learning methods, computer analysis.

Структура обучающей выборки

На рисунке 1 представлен фрагмент обучающей выборки [1], в которой m = 1728 результата наблюдений, n = 6 входных переменных: $x = \{x_1, x_2, .., x_6\}$ и одна выходная переменная y.

x_1	x_2	x_3	x_4	x_5	x_6	y
vhigh	vhigh	2	2	small	low	unacc
med	med	2	more	big	med	acc
vhigh	vhigh	2	2	small	high	unacc
vhigh	vhigh	2	2	med	low	unacc
low	low	5more	more	big	low	unacc
low	low	5more	more	big	med	good
low	low	5more	more	big	high	vgood
vhigh	vhigh	2	2	med	med	unacc

Рис. 1. Фрагмент обучающей выборки

Постановка задачи

Объектом и предметом исследования является проблема построения классификатора $f : x \to y$,

где $x = (x_1, x_2, .., x_6)$ – вектор числовых и строковых данных;

y – идентификатор класса, к которому относится объект.

Цель исследования – проведение сравнительного анализа применимости различных методов машинного обучения [2] для решения задачи классификации.

Описание исходных данных

Наименование входных данных (x):

- Общая цена *(total price)*:

1. цена покупки (*purchase price*);
2. стоимость обслуживания (*cost of maintenance*).

- Техническте характерирстики (technical parameters):

3. количество дверей (*number of doors*);
4. вместимость (людей для перевозки) (*capacity* (*people for transportation*));
5. размер багажника (*trunk size*);
6. безопасность автомобиля (*car safety*).

На рисунке 2 приведена иерархическая структура исходных данных

Рис. 2. Сетевая модель исходных данных

В таблице 1 приведены возможные значения для $n = 6$ входных переменных и наименования четырех классов.

Таблица 1

Множество значений переменных

x_1	x_2	x_3	x_4	x_5	x_6	y
vhigh, high, med, low	*vhigh, high, med, low*	2, 3, 4, 5*more*	2, 4, *more*	*small, med, big*	*low, med, high*	*unacc, acc, good, vgood*

Методы исследования

Для достижения цели исследования реализовывались следующие методы машинного обучения [3, 4, 5]:

1. дерево решений (DecisionTree);
2. градиентный бустинг (GradientBoostedTrees;
3. логистическая регрессия (LogisticRegression);
4. модель Маркова (Markov);
5. наивный Байес (NaiveBayes);
6. ближайший сосед (NearestNeighbors);

7. нейронная сеть (NeuralNetwork);
8. случайный лес (RandomForest);
9. опорные вектора (SupportVectorMachine).

Компьютерное моделирование и практическая апробация системы классификации на большом массиве разнородных данных выполнялись на базе аналитической платформы Wolfram Mathematica [6], имеющей в своем арсенале алгоритмы решения задач построения систем искусственного интеллекта.

Основные результаты

На рисунке 3 приведен оператор построения модели классификации, например, методом Логистической регрессии.

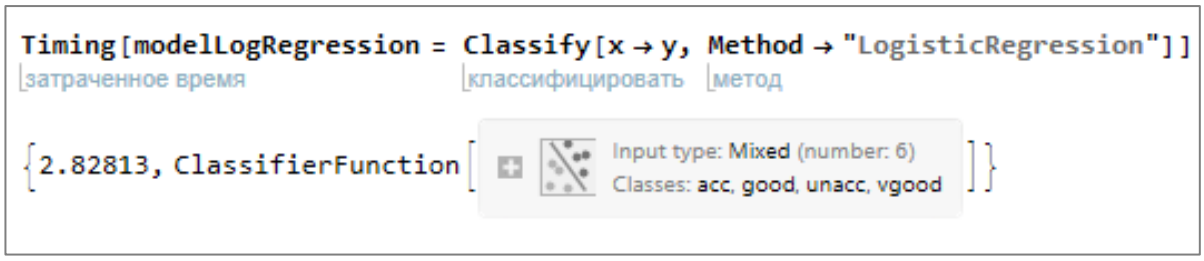

Рис. 3. Обучение модели

Подробная информация о параметрах процесса обучения приведена на рисунке 4.

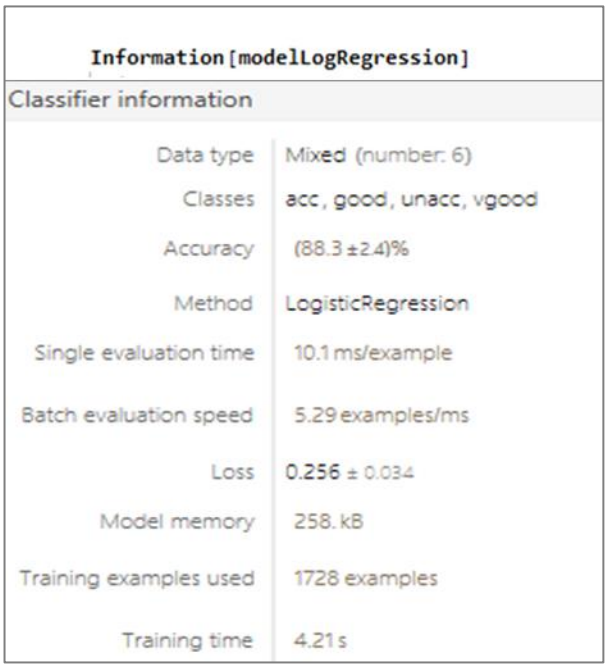

Рис. 4. Информация о процессе обучения

Дополнительная графическая информация об изменении в процессе обучения точности классификации и кривая обучения на тестовом наборе данных приведены на рисунке 5.

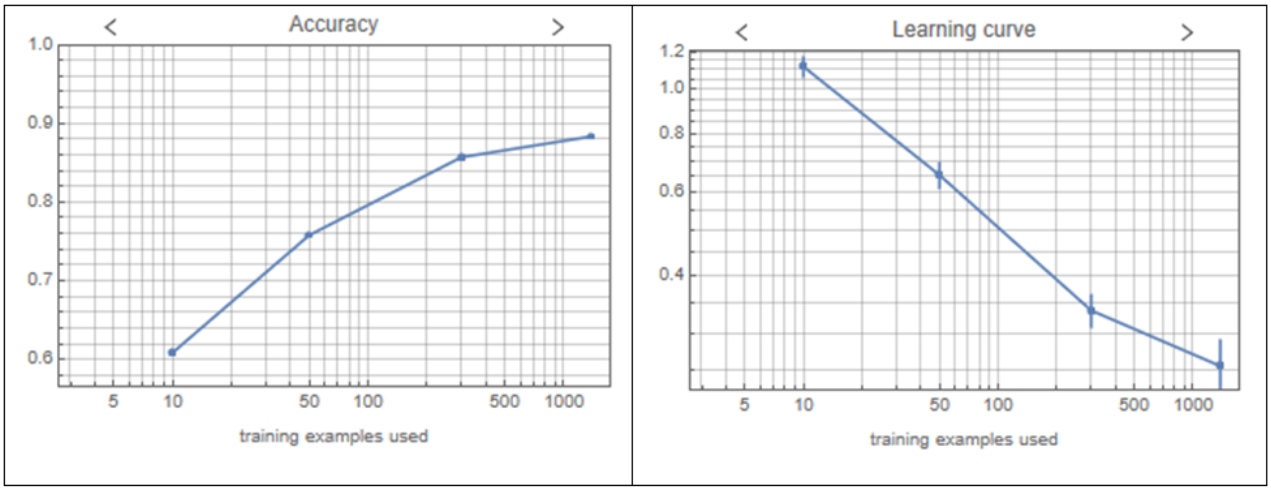

Рис. 5. Изменение точности классификации и кривая обучения

Операторы проверки системы классификации, обученной методом случайного леса, на тестовой выборке приведены на рисунке 6.

```
setfortest = x → resultRandomForest;

ClassifierMeasurements[modelRandomForest, setfortest]
|метрики классификатора
```

Рис. 6. Тестирование модели на обучающей выборке

На рисунке 7 даны сведения о процессе обучения системы классификации и матрица ошибок распознавания классов.

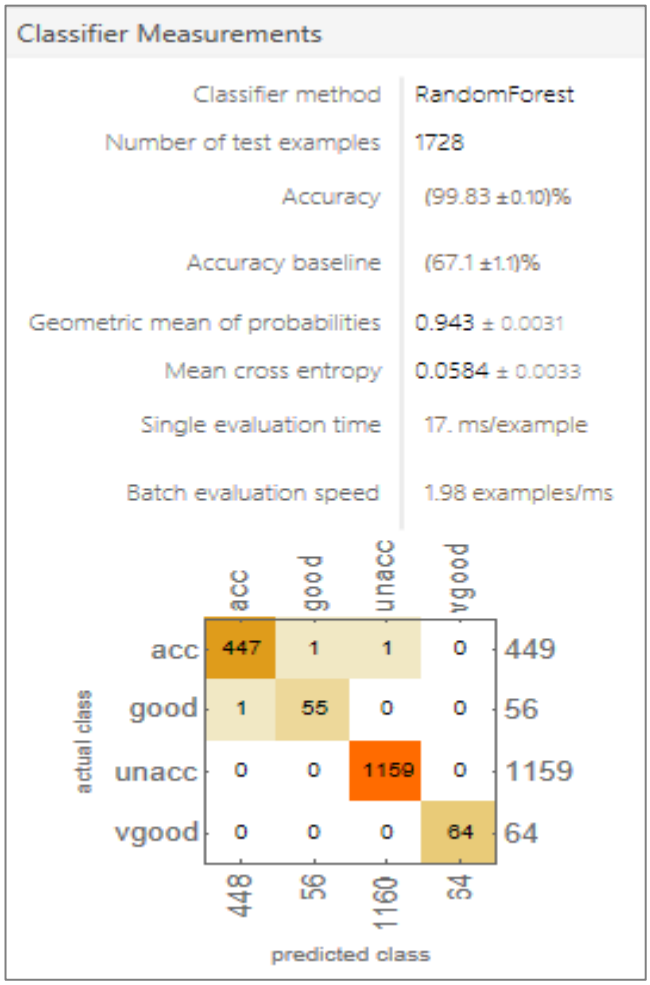

Рис. 7. Информация об ошибках классификации

Основные результаты исследования – показатели качества классификации различными методами представлены в таблице 2.

Таблица 2

Основные оценки качества классификации

Название метода	Время на моделирование, с	Ошибка классификации, %
Дерево решений	5.77	14.4
Градиентный бустинг	3.78	5.3
Логистическая регрессия	1.19	10.0
Модель Маркова	0.52	10.8
Наивный Байес	0.31	13.4
Ближайший сосед	0.27	21.4
Нейронная сеть	11.39	0.4
Случайный лес	2.33	7.1
Опорные вектора	8.36	10.1

После нормирования экспериментальных данных и, учитывая преимущество критерия ошибки классификации над временем обучения модели (в соответствии со шкалой сравнений Саати), получим оценки степени предпочтительности методов машинного обучения применительно к задаче классификации, представленные на рисунке 8.

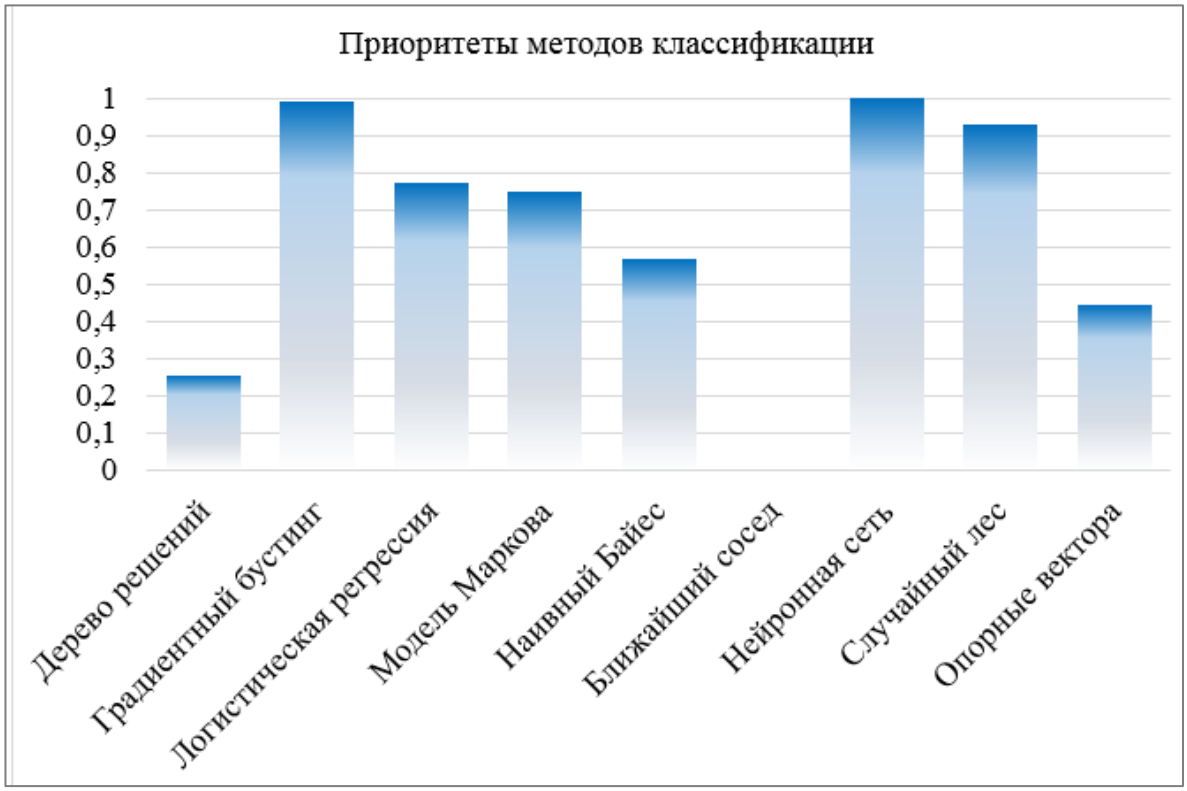

Рис. 8. Ранжирование методов машинного обучения

Из полученных данных следует, что наивысшим приоритеттом обладает метод классификации, основанный на использовании искусственной нейронной сети.

Выводы

Произведен сравнительный анализ 9 методов машинного обучения для решения задачи классификации автомобилей на примере обучающей выборки, состоящей из неоднородных данных (числовых и строковых) большой размерности.

Список литературы

1. Car Evaluation Data Set.
2. URL: https://archive.ics.uci.edu/ml/datasets/car+evaluation.
3. А. Дьяконов «Введение в анализ данных и машинное обучение». URL: https://alexanderdyakonov.files.wordpress.com/2017/06/book_boosting_pdf.pdf (Дата обращения 03.12.2021).
4. Ким Н.Г. Прогнозирование котировок ценных бумаг методами линейной регрессии, дерева решений и с помощью многослойной нейронной сети / Н.Г. Ким, Л.Д. Хлебородова // Студент года 2021: Сборник статей Международного учебно-исследовательского конкурса в 6-ти частях, Петрозаводск, 19 мая 2021 года. – Петрозаводск: Международный центр научного партнерства «Новая Наука», 2021. – С. 288-292. – DOI 10.46916/02062021-4-978-5-00174-249-4.
5. Хлебородова Л.Д. Сравнение методов машинного обучения для задачи прогнозирования в среде Wolfram Mathematica / Л.Д. Хлебородова, Г.С. Осипов // Постулат. 2021. № 10 (72). С. 10.
6. Хлебородова Л.Д. Исследование применимости метода градиентного бустинга для решения задачи прогнозирования / Л.Д. Хлебородова // Лучшая исследовательская статья 2021: сборник статей II Международного научно-исследовательского конкурса, Петрозаводск, 01 ноября 2021 года. – Петрозаводск: Международный центр научного партнерства «Новая Наука» (ИП Ивановская И.И.), 2021. – С. 278-283. – DOI 10.46916/08112021-4-978-5-00174-363-7.
7. Stephen Wolfram. An Elementary Introduction to the Wolfram Language. URL: https://www.wolfram.com/language/elementary-introduction/2nd-ed/ (Дата обращения 03.12.2021).