DOI 10.58351/2949-2041.2023.4.4.023

Шинкаренко Кирилл Константинович, студент 2 курса,
направления 01.03.02 Прикладная математика и информатика,
Сахалинский государственный университет, Южно-Сахалинск
Shinkarenko Kirill Konstantinovich, Sakhalin State University

Научный руководитель: **Осипов Геннадий Сергеевич,**
д.т.н., профессор кафедры информатики,
Сахалинский государственный университет, Южно-Сахалинск
Gennady S. Osipov, Sakhalin State University

ВВЕДЕНИЕ В ПРИНЦИПЫ ДЕКОМПОЗИЦИИ ОБУЧАЮЩЕЙ ВЫБОРКИ НА ОБУЧАЮЩЕЕ И ТЕСТОВОЕ МНОЖЕСТВА В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ
INTRODUCTION TO THE PRINCIPLES OF DECOMPOSITION OF A TRAINING SAMPLE INTO TRAINING AND TEST SETS IN MACHINE LEARNING TASKS

Аннотация: В статье рассматривается проблема синтеза моделей объектов методами машинного обучения. Исследуется задача разбиения исходных данных (обучающей выборки), которые могут использоваться для построения модели, на обучающее и тестовое множества. Проведенное исследование доведено до практической реализации в среде системы символьной математики.

Abstract: The article deals with the problem of synthesis of object models by machine learning methods. The problem of splitting the initial data (training sample), which can be used to build a model into training and test sets, is investigated. The conducted research has been brought to practical implementation in the environment of the symbolic mathematics system.

Ключевые слова: методы машинного обучения, задача классификации.

Keywords: machine learning methods, classification problem.

Введение

Принцип декомпозиции обучающей выборки в задачах машинного обучения является не самой последней темой, которая позволяет эффективно оценивать качество модели и принимать взвешенные решения на этапе обучения.

В машинном обучении часто возникает необходимость оценки качества модели на новых данных, которых не было во время обучения. Адекватная оценка обобщающей способности модели является ключевым критерием успешности в области машинного обучения. Чтобы достичь надежного и объективного результата оценки, необходимо разделить имеющиеся данные на обучающую и тестовую часть.

Обучающая выборка используется для обучения модели, тогда как тестовая выборка представляет собой независимую часть данных, на которой проверяется качество модели.

В данной статье постараемся доказать важность использования декомпозиции набора данных при машинном обучении.

Постановка задачи

Целью настоящего исследования являлась разработка методологических основ (и их практическая пробация) решения проблемы декомпозиции обучающей выборки (data set) на два непересекающихся подмножества:

1. множества, используемого в процессе обучения для построения модели исследуемого процесса или системы методами машинного обучения (обучающее множество);
2. множества исходных данных, которые не использовались в процессе обучения, а предназначены только для оценки качества построенной модели (тестовое множество).

В работе было проведено аналитическое сравнение следующих методов машинного обучения [1]:

1. метод распределения классов;
2. дерево решений;
3. деревья с градиентным усилением;
4. логистическая регрессия;
5. метод Маркова;
6. наивный Байес;
7. ближайшие соседи;
8. нейронная сеть;
9. случайный лес;
10. машина опорных векторов.

Этап обучения модели

На рисунке 1 представлена функция построения, например, нейросетевого классификатора, которая является отображением множества элементов из обучающего множества (xL) в соответствующий заданный класс объекта (yL).

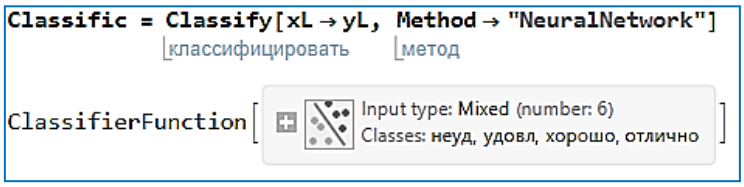

Рис. 1 Построение нейросетевого классификатора

Практическая реализация осуществлялась в среде системы символьной математики Wolfram Mathematica [2, 3].

Полная информация об обучении модели (включая график кривой обучения) приведена на рисунке 2.

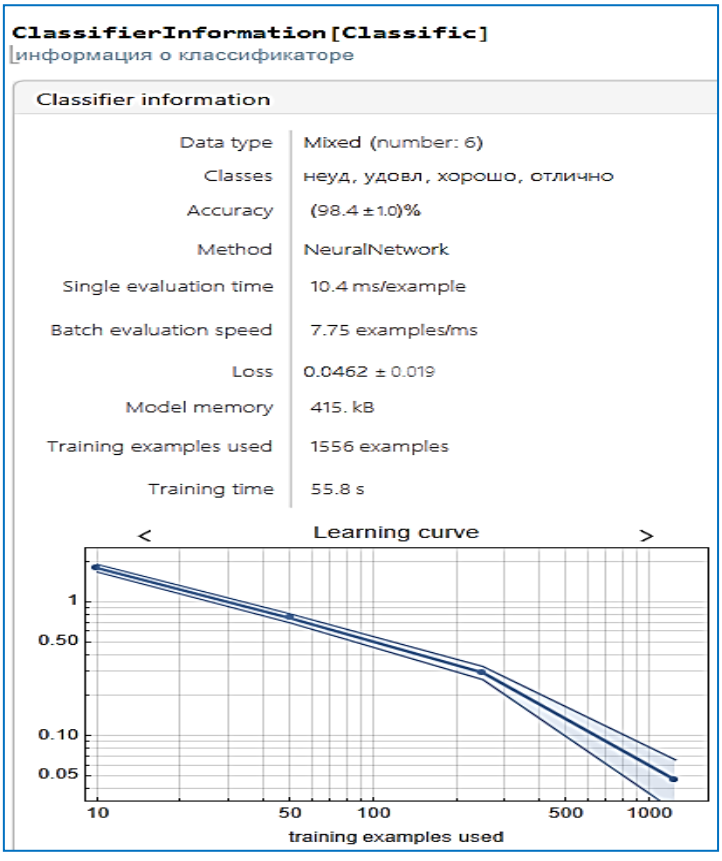

Рис. 2 Информация о классификаторе и кривая его обучения

Тестирование модели на данных, отсутствующих в обучающей выборке

На рисунке 3 приведен этап использования тестового множества для оценки качества, построенной в процессе обучения модели.

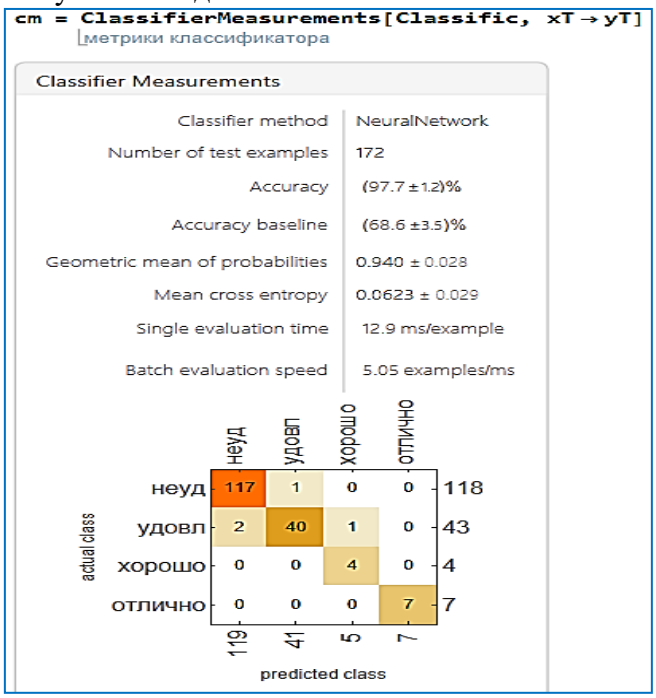

Рис. 3 Классификация на тестовом множестве

В данном случае точность классификации составляет порядка 97%.

Тестирование модели на данных, присутствующих в обучающей выборке

Чтобы доказать наибольшую объективность в оценке классификатора при декомпозиции данных на обучающее и тестирующее множества, проведём ещё одно тестирование, но на данных, которые присутствовали в обучающей выборке.

На рисунке 4 приведен этап использования тестового множества (подмножества обучающего) для оценки качества, построенной в процессе обучения модели.

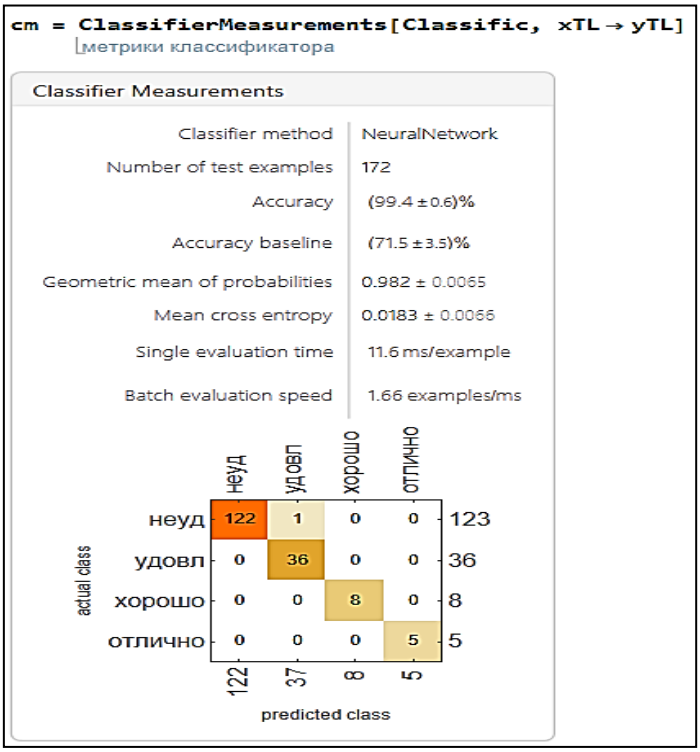

Рис. 4 Классификация на тестовом множестве (подмножестве обучающего)

В данном случае точность классификации составляет порядка 99%.

Аналогичные исследования были проведены для используемых в статье методов машинного обучения.

Выводы и заключение

В результате исследования классификаторов, обученных разными методами, получили таблицу.

Таблица

Основные результаты исследования

№	Метод обучения	Время обучения классификатора, с	Результат тестирования классификатора	
			на данных, не используемых в обучении, точность (%)	на данных, используемых в обучении, точность (%)
1.	Нейронная сеть	55,8	97,7	99,4
2.	Случайный лес	0,481	95,3	94,8
3.	Деревья с градиентным усилением	14,3	94,8	94,2
4.	Машина опорных векторов	27	93,0	93,6
5.	Логистическая регрессия	5,3	92,4	92,4
6.	Дерево решений	0,502	83,7	85,5
7.	Модель Маркова	0,762	83,7	90,7
8.	Наивный Байес	0,414	83,1	83,7
9.	Ближайшие соседи	0,486	83,1	87,2
10.	Распределения классов	5,36	79,7	80,2

Из таблицы видно, что результаты тестирования классификатора на данных, используемых в обучении, и на данных, не используемых в обучении, в 9 из 10 выбранных методов - разные. В 6 из 10 методов результаты отличаются на уровне погрешности. И, наконец, в 4 из 10 методов результаты отличаются более, чем на 1%. Причём результат тестирования на данных, присутствующих в обучающей выборке, лучше, чем результат тестирования на данных, которых нет в обучающей выборке. Следовательно, декомпозиция набора данных на два множества: обучающее и тестирующее, - действительно показывает наиболее объективные результаты.

Небольшая разница в результатах (или её отсутствие) не означает, что декомпозицией можно пренебречь. В данном исследовании использовался далеко не самый большой по объёму набор данных. Большая разница будет видна на многотысячных-миллионных наборах данных, где объективность оценки обучения крайне важна.

В заключение, в данной статье мы рассмотрели важность декомпозиции обучающей выборки на обучающее и тестовое множества в задачах машинного обучения. Этот подход позволяет достичь объективной оценки качества модели на независимых данных, что является ключевым для успешного применения алгоритмов машинного обучения.

Список литературы:

1. Способы обеспечения качества данных для машинного обучения URL: https://habr.com/ru/articles/588266/ (Дата обращения 25.11.2023).

2. Wolfram Language & System Documentation Center URL:

https://reference.wolfram.com/language/ref/Classify.html?v=13.3 (Дата обращения 25.11.2023).

3. Wolfram Mathematica, URL: https://www.wolfram.com/mathematica/ (Дата обращения 25.11.2023).