DOI 10.37539/2949-1991.2023.9.9.026
УДК 004.855.5

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, Yuzhno-Sakhalinsk

Научный руководитель:
Осипов Геннадий Сергеевич, д.т.н., профессор кафедры Информатики,
Сахалинский государственный университет, Южно-Сахалинск
Osipov Gennady Sergeevich, Sakhalin State University, Yuzhno-Sakhalinsk

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ФОРМАТА ПРЕДСТАВЛЕНИЯ ОБУЧАЮЩЕЙ ВЫБОРКИ НА ТОЧНОСТЬ КЛАССИФИКАТОРОВ, ПОСТРОЕННЫХ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ
INVESTIGATION OF THE INFLUENCE OF THE TRAINING SAMPLE PRESENTATION FORMAT ON THE ACCURACY OF CLASSIFIERS CONSTRUCTED BY MACHINE LEARNING METHODS

Аннотация: Проведено исследование влияние вида исходных данных (обучающей выборки) на качество обучения интеллектуальных классификаторов, синтезируемых методами машинного обучения.

Выполнено аналитическое сравнение и практическая апробация в системе символьной математики трех базовых методов машинного обучения.

Abstract: The influence of the type of initial data (training sample) on the quality of training of intelligent classifiers synthesized by machine learning methods has been studied.

Analytical comparison and practical approbation of three basic machine learning methods in the system of symbolic mathematics are carried out.

Ключевые слова: обучающая выборка, классификация объектов, машинное обучение.
Keywords: training sampling, object classification, machine learning.

Введение

Двадцать первый век – век развития информационных технологий. С середины двадцатого века и по сей день человечество стремится овладеть загадочным цифровым миром. Одним из направлений развития стала область машинного обучения.

Современное развитие этой области подразумевает использование большого объема данных для обучения и оценки моделей классификации. При этом, формат представления этих обучающих выборок оказывает значительное влияние на точность получаемых классификаторов. В данном исследовании мы сосредоточимся на анализе этого влияния и поиском оптимального формата представления обучающих выборок для достижения максимальной точности классификации.

Формальная постановка задачи

Объектом и предметом исследования является проблема построения обучаемых классификаторов вида $f: \boldsymbol{x} \to y$,

где $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ – вектор количественных и символьных характеристик объекта;

y – идентификатор класса к которому относится объект.

Цель исследования – проведение сравнительного анализа различных форм представления исходных данных (обучающей выборки) и исследования их влияния на точность синтеза интеллектуальных классификаторов на базовых методах машинного обучения.

Практическая апробация методологии исследования проблемы осуществлялась на основе экспертной системы для оценки автомобилей [1].

Таким образом, в данном случае классификатор представим в виде:

$$f\colon (x_1, x_2, \ldots, x_6) \to y.$$

В качестве исследовательской аналитической платформы использовалась система символьной математики Wolfram Mathematica [2], которая является одной из наиболее современных сред моделирования систем искусственного интеллекта и реализации методов машинного обучения.

Для аналитического исследования применялись следующие классические методы машинного обучения:

1. Логистическая регрессия;
2. Метод Маркова;
3. Многослойная нейронная сеть.

Анализировались различные виды представления исходных данных:

1. Исходные англоязычные данные [1].
2. Соответствующие характеристики объектов, представленные на русском языке одним словом;
3. Многословные русскоязычные характеристики и идентификаторы классов;
4. Априорное цифровое представление исходных данных (оцифровка пользователем);
5. Оцифровка в среде используемой аналитической платформы.

1 Исходные англоязычные данные

1.1 Исходные данные

Множество возможных значений характеристик объектов и идентификаторов классов представлены в таблице 1

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*

1.2 Методы решения

1.2.1 Логистическая регрессия

На рисунке 1 представлена информация о классификаторе, обученном с помощью метода Логистической регрессии, о процессе его обучения и соответствующая ему кривая обучения.

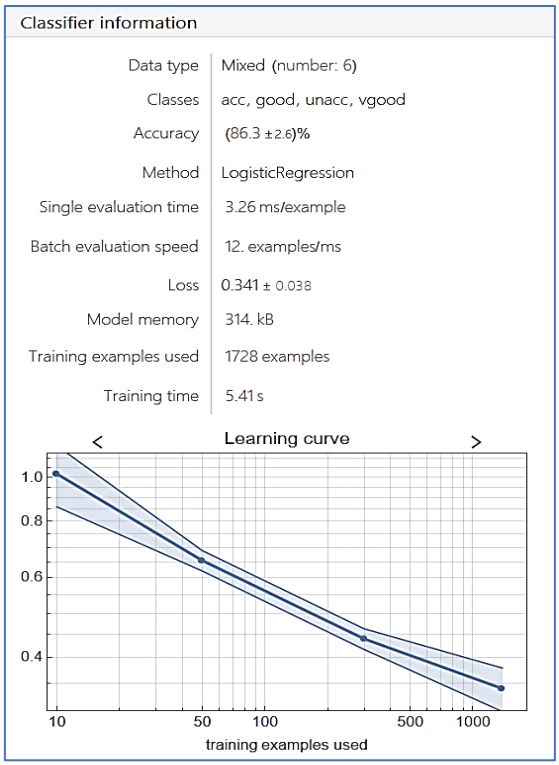

Рис. 1 Информация о классификаторе (метод Логистической регрессии).

Оценка точности построенного классификатора методом логистической регрессии представлена на рисунке 2.

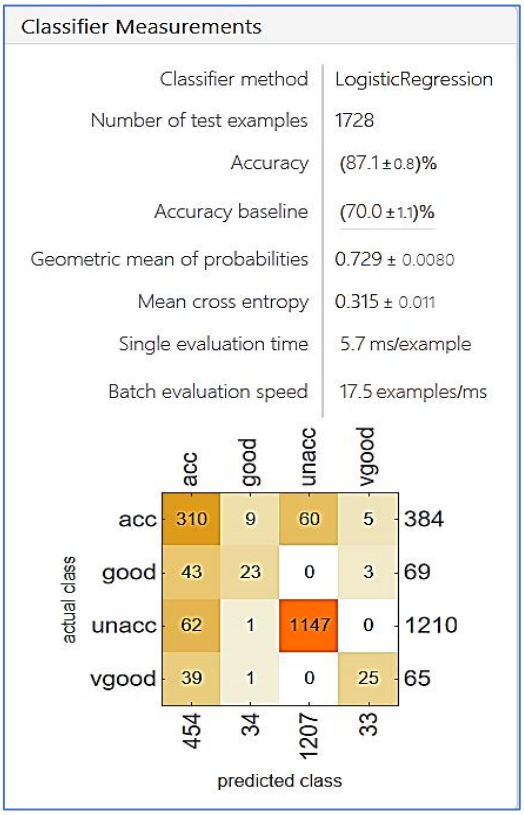

Рис. 2 Сводная матрица ошибок классификации

1.2.2 Модель Маркова

На рисунке 3 представлена информация о классификаторе, обученном с помощью метода Маркова, о процессе его обучения и соответствующая ему кривая обучения.

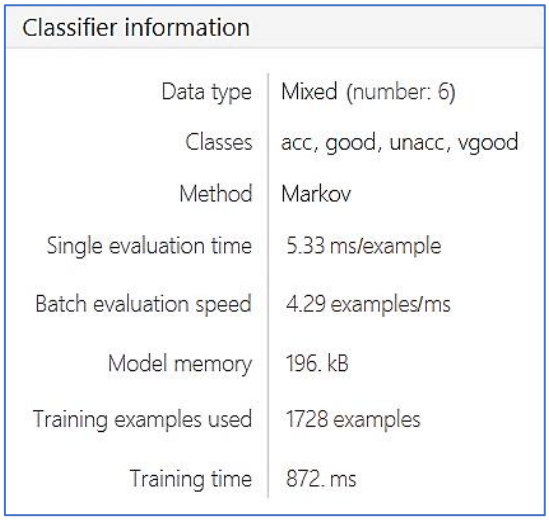

Рис. 3 Информация о классификаторе (метод Маркова)

Оценка точности построенного классификатора методом Маркова представлена на рисунке 4.

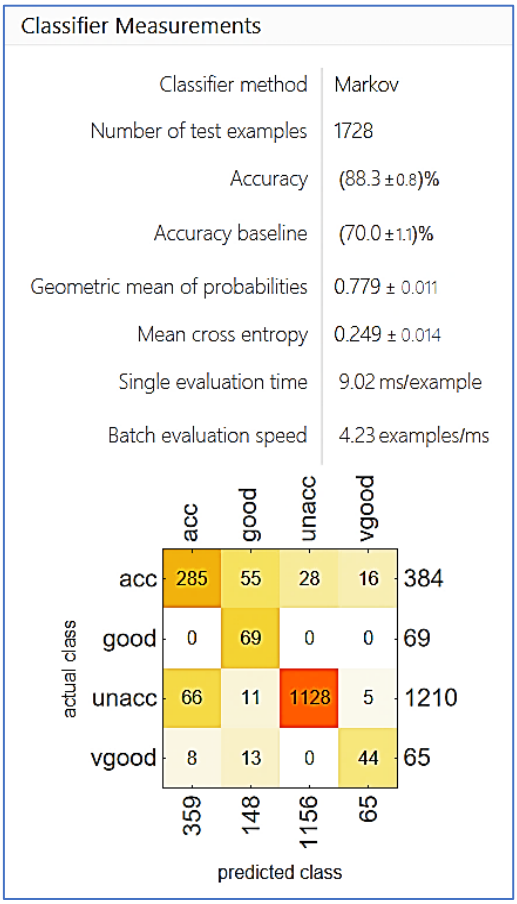

Рис. 4 Сводная матрица ошибок классификации

1.2.3 Многослойная нейронная сеть

На рисунке 5 представлена информация о классификаторе, обученном с помощью метода Многослойной нейронной сети, о процессе его обучения и соответствующая ему кривая обучения.

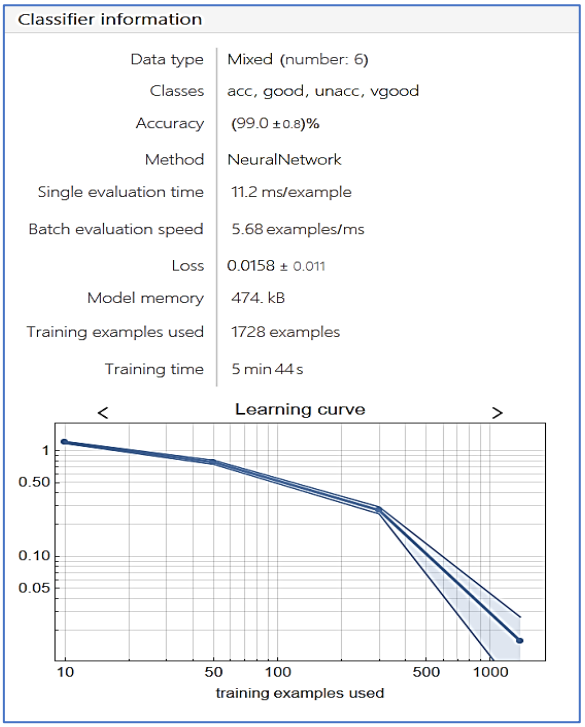

Рис. 5 Информация о классификаторе (метод Многослойной нейронной сети)

Оценка точности построенного классификатора представлена на рисунке 6.

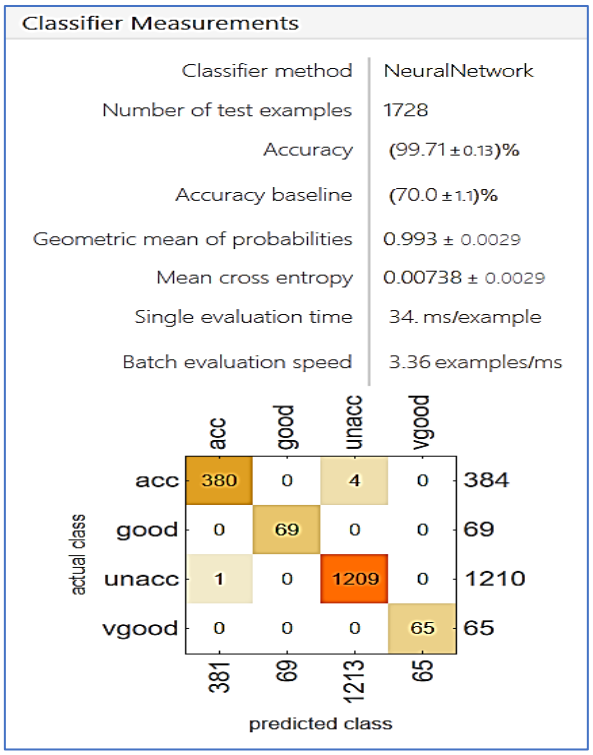

Рис. 6 Матрица ошибок обучения нейросетевого классификатора

1.2.4 Сравнение методов обучения

В таблице 2 представлена точность обучения классификаторов различными методами машинного обучения.

Таблица 2

Сравнение точности обучения

Метод машинного обучения	Логистическая регрессия	Модель Маркова	Нейронная сеть
Точность классификации, %	~87,1	~88,3	~99,7

2 Однословный русский

2.1 Исходные данные

Множество возможных значений характеристик объектов и идентификаторов классов представлены в таблице 3

Таблица 3

Множество значений переменных

x_1	x_2	x_3	x_4	x_5	x_6	y
высоченная, высокая, средняя, низкая	высоченная, высокая, средняя, низкая	2, 3, 4, >=5	2, 4, >4	маленькая, средняя, большая	низкая, средняя, высокая	неуд, удовл, хорошо, отлично

2.2 Методы решения

2.2.1 Логистическая регрессия

На рисунке 7 представлена информация о классификаторе, обученном с помощью метода Логистической регрессии, о процессе его обучения и соответствующая ему кривая обучения.

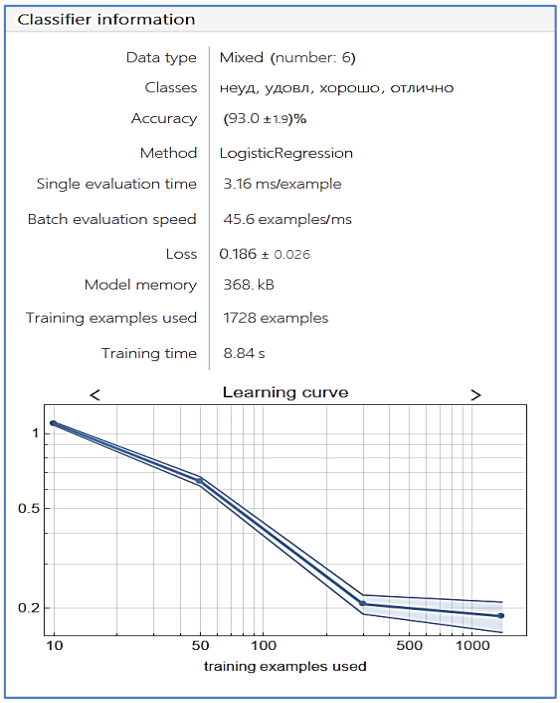

Рис. 7 Информация о классификаторе (метод Логистической регрессии).

Оценка точности построенного классификатора представлена на рисунке 8.

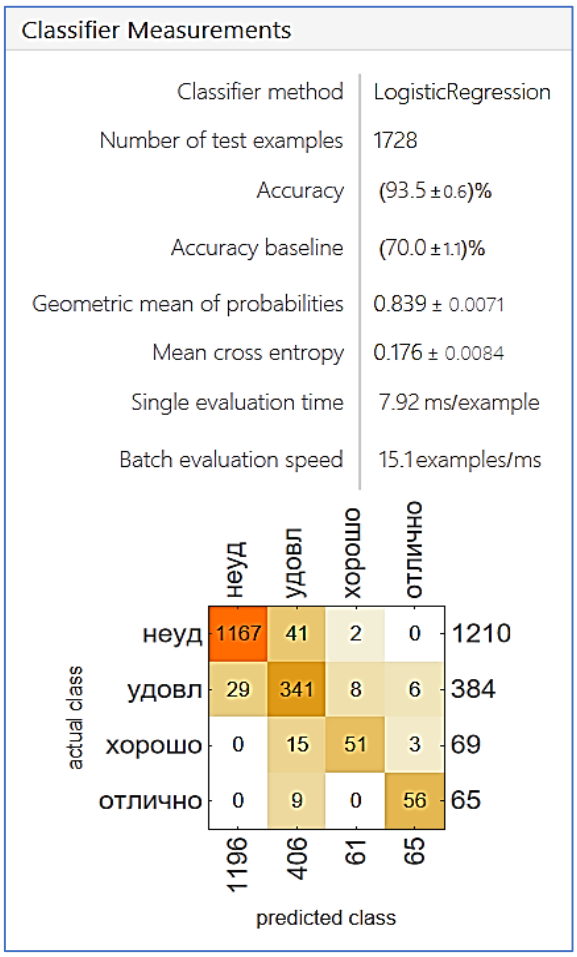

Рис. 8 Матрица ошибок обучения

2.2.2 Модель Маркова

На рисунке 9 представлена информация о классификаторе, обученном с помощью метода Маркова, о процессе его обучения и соответствующая ему кривая обучения.

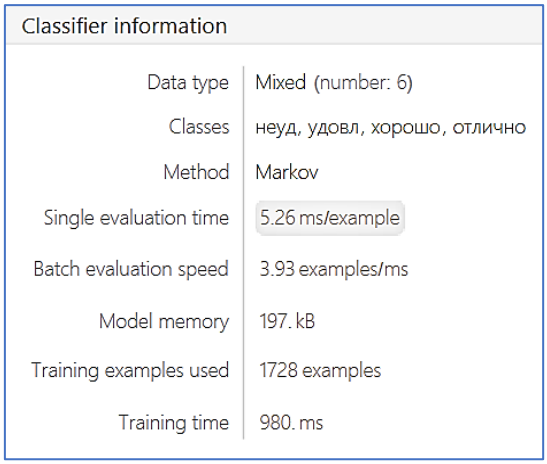

Рис. 9 Информация о классификаторе (метод Маркова)

Оценка точности построенного классификатора представлена на рисунке 10.

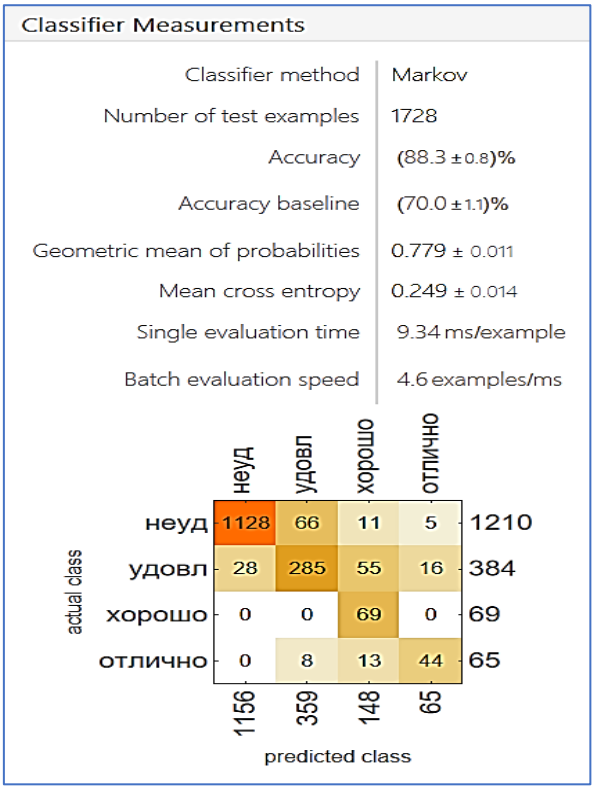

Рис. 10 Матрица ошибок классификации

2.2.3 Многослойная нейронная сеть

На рисунке 11 представлена информация о классификаторе, обученном с помощью метода Многослойной нейронной сети, о процессе его обучения и соответствующая ему кривая обучения.

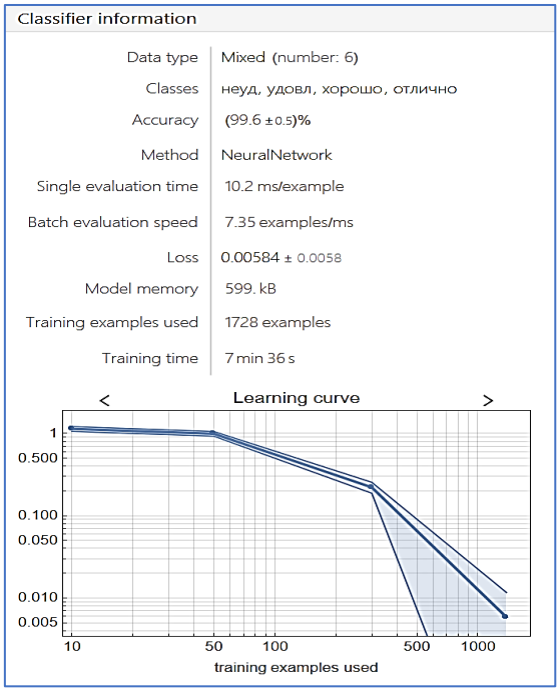

Рис. 11 Информация о классификаторе (метод Многослойной нейронной сети)

Оценка точности построенного нейросетевого классификатора представлена на рисунке 12.

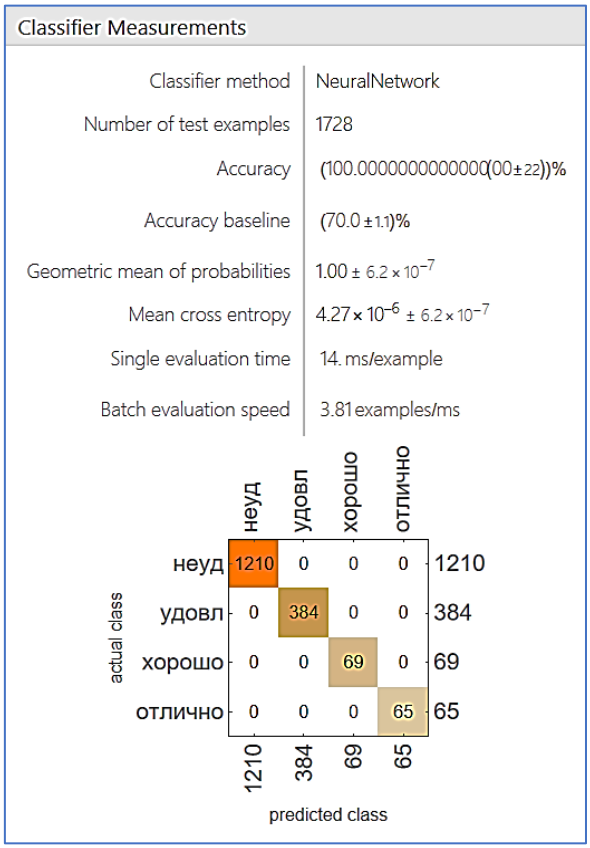

Рис. 12 Матрица ошибок обучения нейросетевого классификатора

2.2.4 Сравнение методов обучения

В таблице 4 представлена точность обучения классификаторов исследуемыми методами машинного обучения.

Таблица 4

Сравнение точности обучения

Метод машинного обучения	Логистическая регрессия	Модель Маркова	Нейронная сеть
Точность классификации, %	~93,5	~88,3	~100

3 Многословный русский

3.1 Исходные данные

Множество возможных значений характеристик объектов и идентификаторов классов представлены в таблице 5.

Таблица 5

Множество значений переменных

x_1	x_2	x_3	x_4	x_5	x_6	y
очень высокая,	*очень высокая,*	*2*	*2*	*маленькая,*	*низкая,*	*неуд,*
высокая,	*высокая,*	*3*	*4*	*средняя,*	*средняя,*	*удовл,*
средняя,	*средняя,*	*4*	*>4*	*большая*	*высокая*	*хорошо,*
низкая	*низкая*	*>=5*				*отлично*

3.2 Методы решения

3.2.1 Логистическая регрессия

На рисунке 13 представлена информация о классификаторе, обученном с помощью метода Логистической регрессии, о процессе его обучения и соответствующая ему кривая обучения.

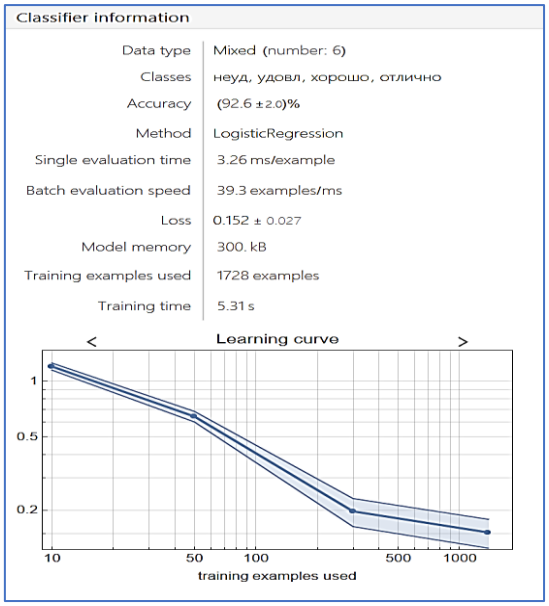

Рис. 13 Информация о классификаторе (метод Логистической регрессии).

Оценка точности построенного классификатора представлена на рисунке 14.

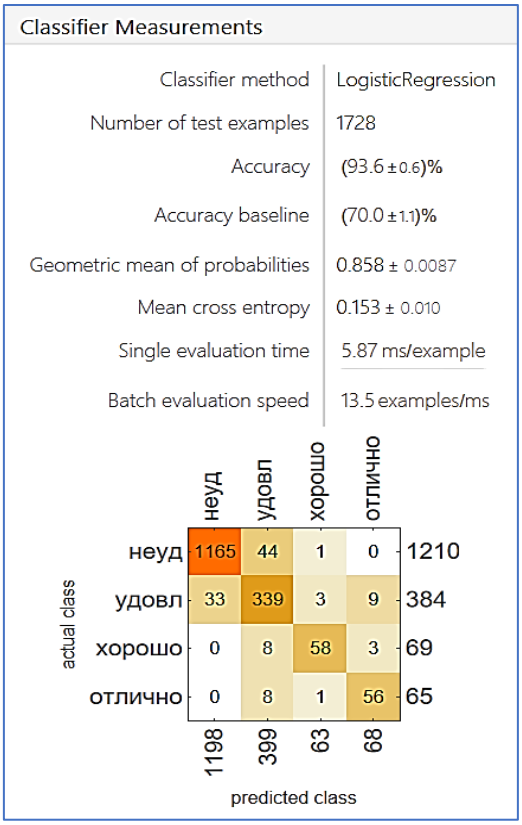

Рис. 14 Матрица ошибок обучения

3.2.2 Модель Маркова

На рисунке 15 представлена информация о классификаторе, обученном с помощью метода Маркова, о процессе его обучения и соответствующая ему кривая обучения.

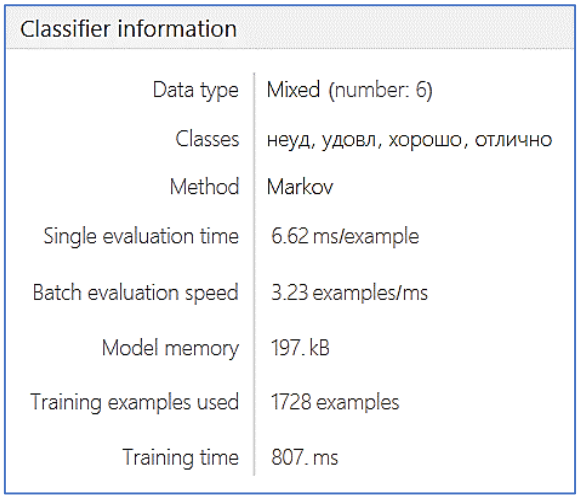

Рис. 15 Информация о классификаторе (метод Маркова)

Оценка точности построенного классификатора представлена на рисунке 16.

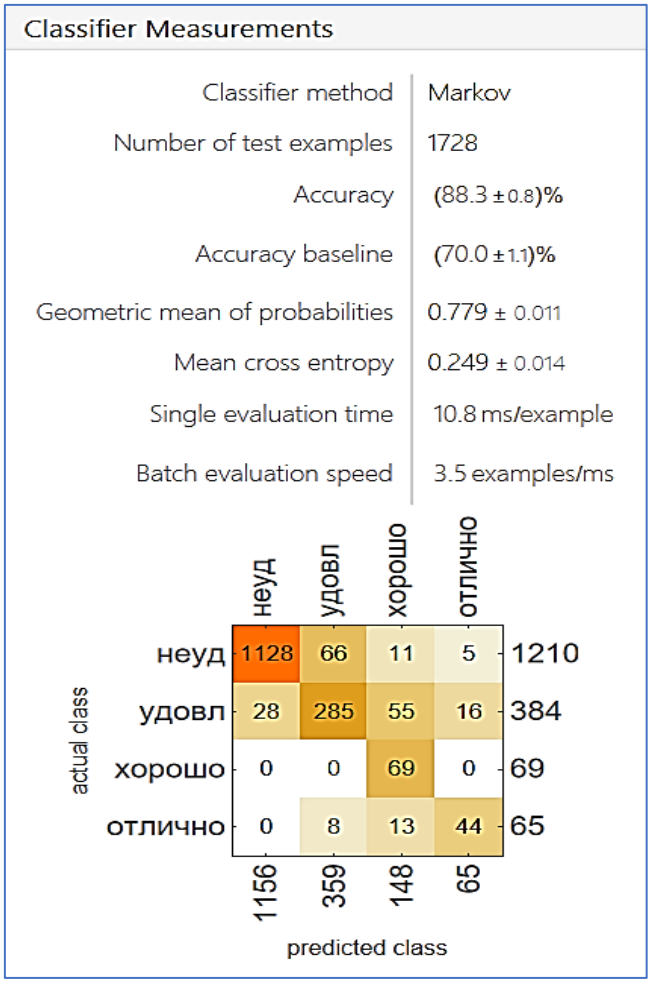

Рис. 16 Матрица ошибок классификации

3.2.3 Многослойная нейронная сеть

На рисунке 17 представлена информация о классификаторе, обученном с помощью метода Многослойной нейронной сети, о процессе его обучения и соответствующая ему кривая обучения.

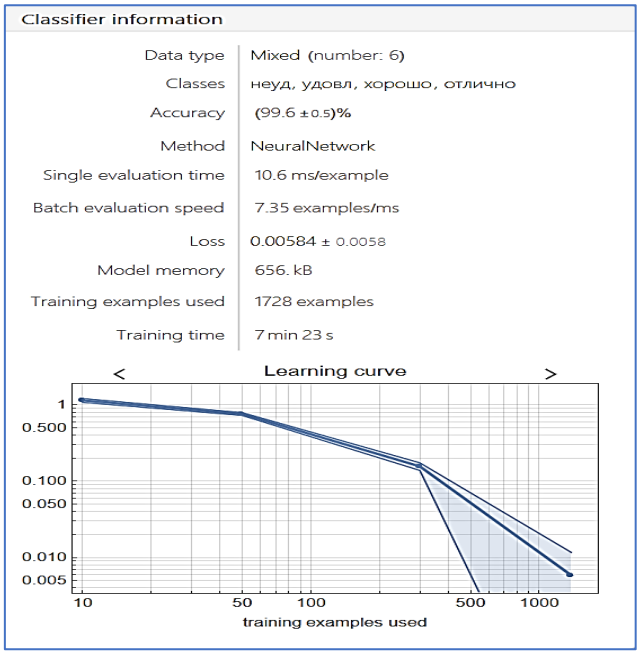

Рис. 17 Информация о классификаторе (метод Многослойной нейронной сети)

Оценка точности построенного нейросетевого классификатора представлена на рисунке 18.

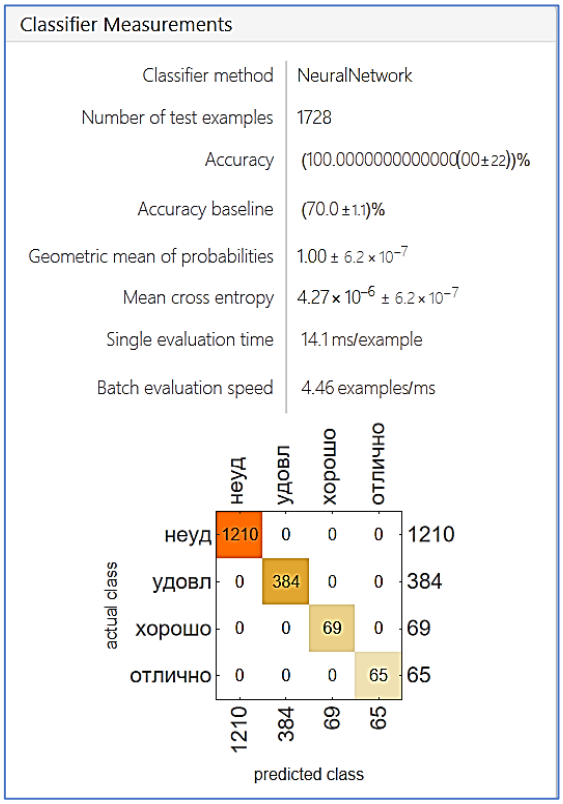

Рис. 18 Матрица ошибок обучения нейросетевого классификатора

3.2.4 Сравнение методов обучения

В таблице 6 представлена точность обучения классификаторов исследуемыми методами машинного обучения

Таблица 6

Сравнение точности обучения

Метод машинного обучения	Логистическая регрессия	Модель Маркова	Нейронная сеть
Точность классификации, %	~93,6	~88,3	~100

4 Цифровое представление исходных данных пользователем

4.1 Исходные данные

Множество возможных значений характеристик объектов и идентификаторов классов представлены в таблице 7.

Таблица 7

Множество значений переменных

x_1	x_2	x_3	x_4	x_5	x_6	y
4	*4*	*2*	*2*	*1*	*1*	*1*
3	*3*	*3*	*4*	*2*	*2*	*2*
2	*2*	*4*	*8*	*3*	*3*	*3*
1	*1*	*5*				*4*

4.2 Методы решения

4.2.1 Логистическая регрессия

На рисунке 19 представлена информация о классификаторе, обученном с помощью метода Логистической регрессии, о процессе его обучения и соответствующая ему кривая обучения.

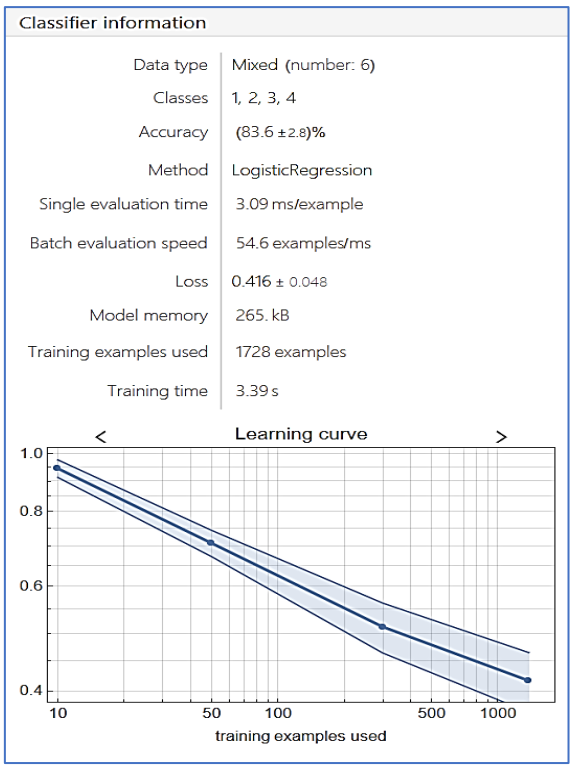

Рис. 19 Информация о классификаторе (метод Логистической регрессии)

Оценка точности построенного классификатора представлена на рисунке 20.

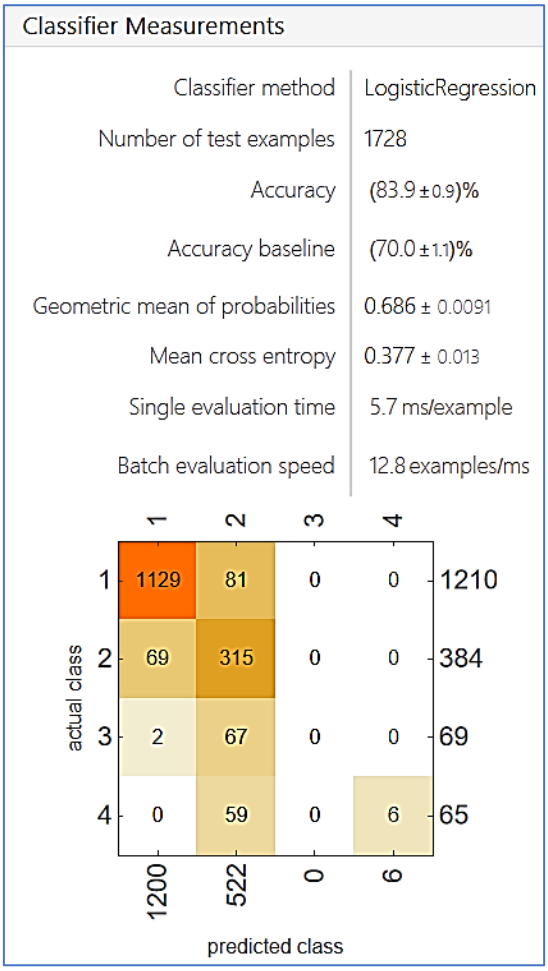

Рис. 20 Матрица ошибок обучения

4.2.2 Модель Маркова

На рисунке 21 представлена информация о классификаторе, обученном с помощью метода Маркова, о процессе его обучения и соответствующая ему кривая обучения.

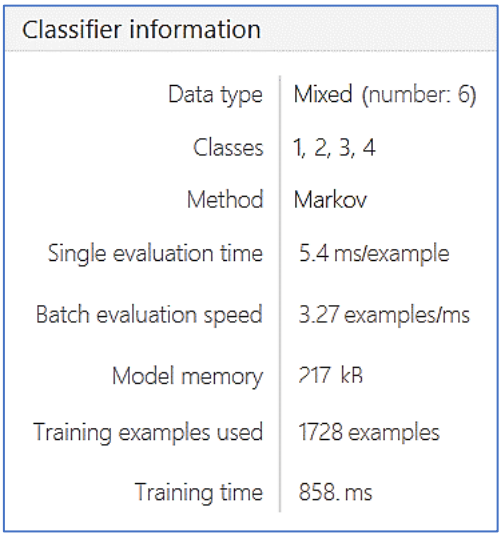

Рис. 21 Информация о классификаторе (метод Маркова)

Оценка точности построенного классификатора представлена на рисунке 22.

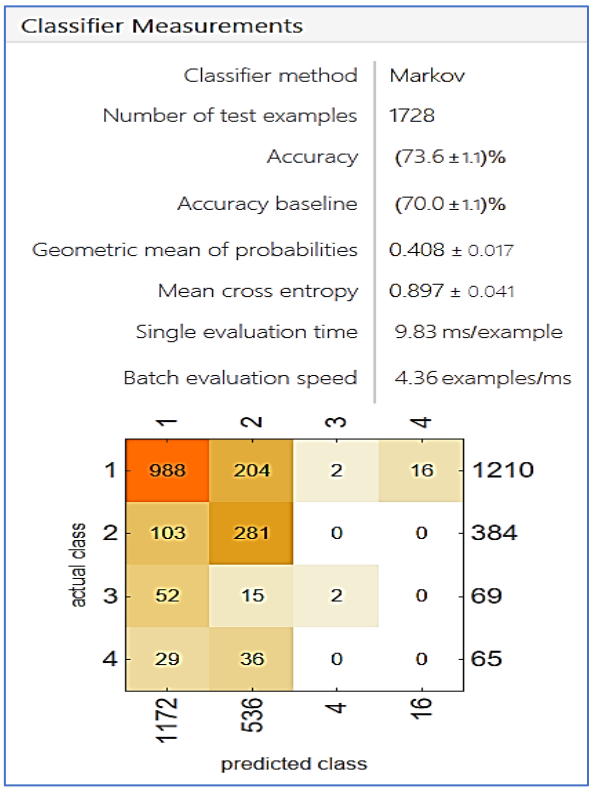

Рис. 22 Матрица ошибок классификации

4.2.3 Многослойная нейронная сеть

На рисунке 23 представлена информация о классификаторе, обученном с помощью метода Многослойной нейронной сети, о процессе его обучения и соответствующая ему кривая обучения.

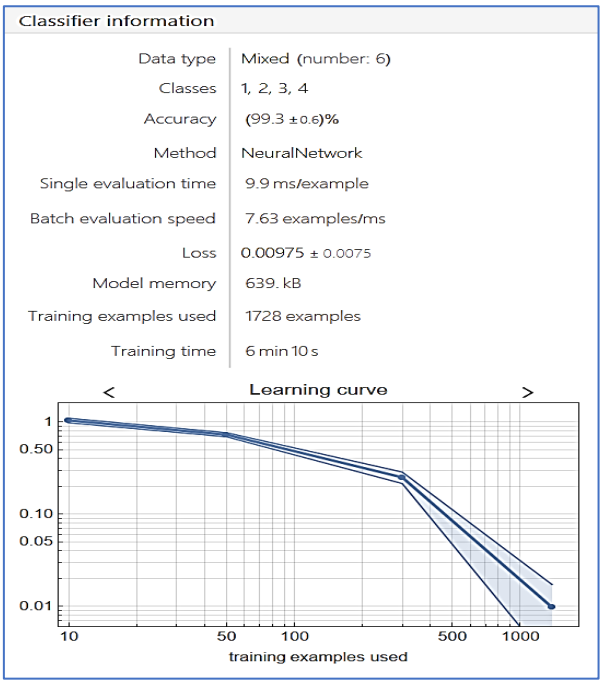

Рис. 23 Информация о классификаторе (метод Многослойной нейронной сети)

Оценка точности построенного нейросетевого классификатора представлена на рисунке 24.

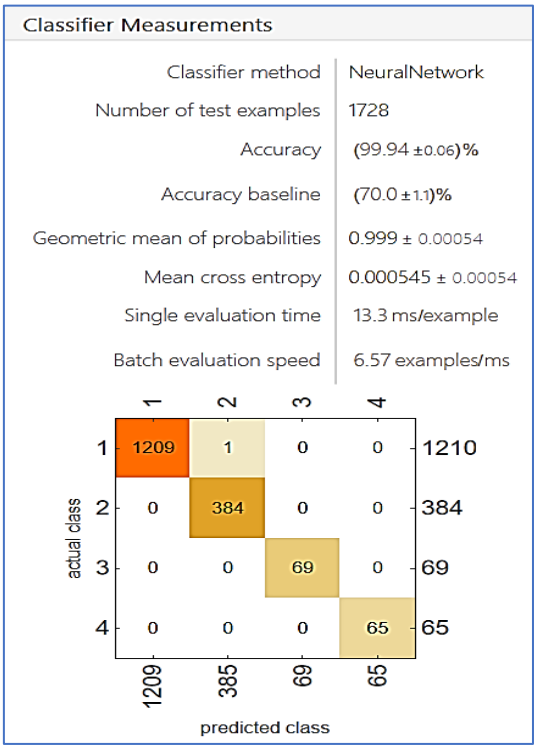

Рис. 24 Матрица ошибок обучения нейросетевого классификатора

4.2.4 Сравнение методов обучения

В таблице 8 представлена точность обучения классификаторов исследуемыми методами машинного обучения

Таблица 8

Сравнение точности обучения

Метод машинного обучения	Логистическая регрессия	Модель Маркова	Нейронная сеть
Точность классификации, %	~83,9	~73,6	~99,9

5 Автоматическая оцифровка значений исходных данных в системе

5.1 Исходные данные

Примеры возможных значений оцифрованных характеристик объектов и идентификаторов классов представлены в таблице 9.

Таблица 9

Примеры значений оцифрованных переменных

x_1	x_2	x_3	x_4	x_5	x_6	y
0.6836502286886217	*1.5881572349320594*	*0.5934967922671609*	*0.8425971741729311*	*0.8660254037844386*	*2.048943757547123*	*acc*
0.6282772388863171	*0.0909773396921357 3*	*0.6031607248629 96*	*0.8615185591569103*	*1.7320508075688772*	*2.1191917673893665*	*acc*
0.5729042490840126	*1.4062025555475788*	*0.6128246574588311*	*0.8804399441408894*	*0.8660254037844388*	*2.1894397772316 1*	*Good*
1.9307734172461777	*1.6334262533750854*	*0.3917062089092 5684*	*0.0450264342359 7426*	*0.8660254037844386*	*1.3032925811813 96*	*vgood*

5.2 Методы решения

5.2.1 Логистическая регрессия

На рисунке 25 представлена информация о классификаторе, обученном с помощью метода Логистической регрессии, о процессе его обучения и соответствующая ему кривая обучения.

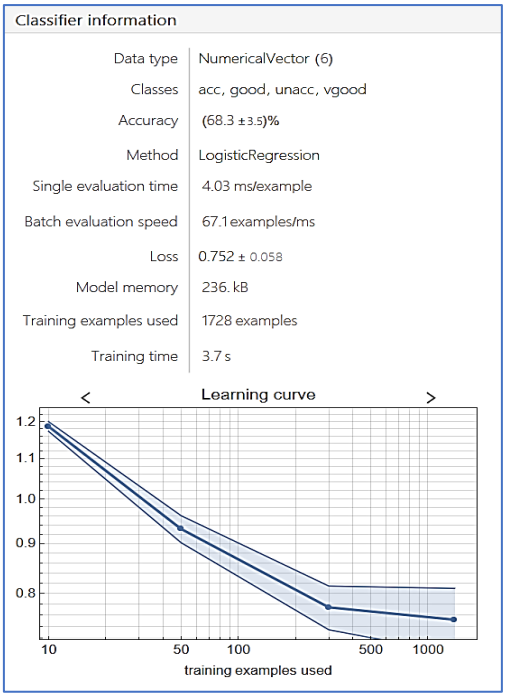

Рис. 25 Информация о классификаторе (метод Логистической регрессии)

Оценка точности построенного классификатора методом логистической регрессии представлена на рисунке 26.

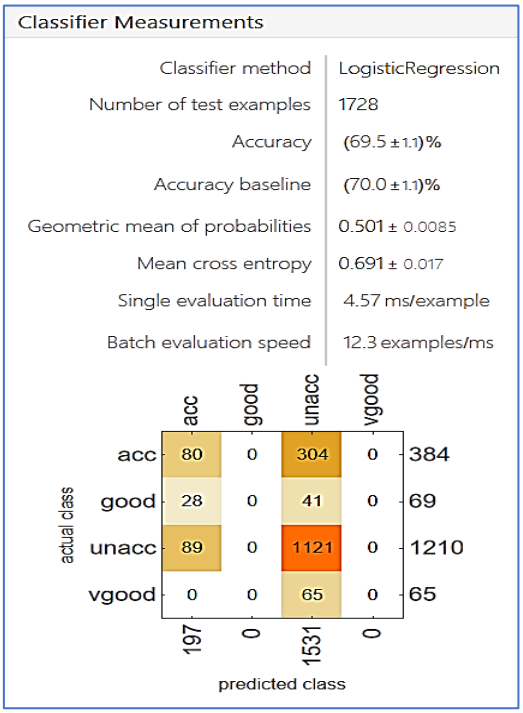

Рис. 26 Сводная матрица ошибок классификации

5.2.2 Модель Маркова

На рисунке 27 представлена информация о классификаторе, обученном с помощью метода Маркова, о процессе его обучения и соответствующая ему кривая обучения.

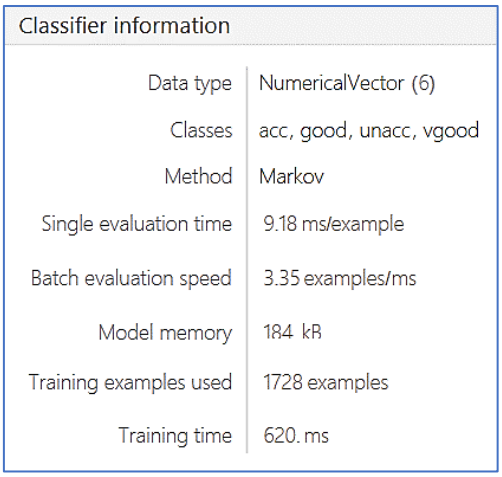

Рис. 27 Информация о классификаторе (метод Маркова)

Оценка точности построенного классификатора методом Маркова представлена на рисунке 28.

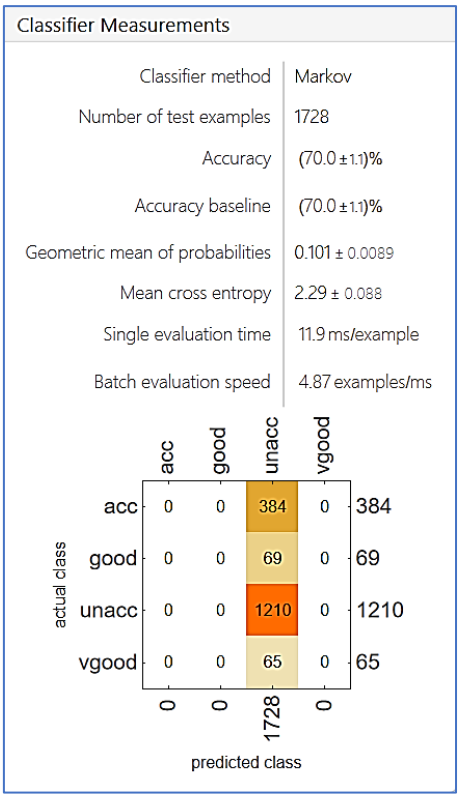

Рис. 28 Сводная матрица ошибок классификации

5.2.3 Многослойная нейронная сеть

На рисунке 29 представлена информация о классификаторе, обученном с помощью метода Многослойной нейронной сети, о процессе его обучения и соответствующая ему кривая обучения.

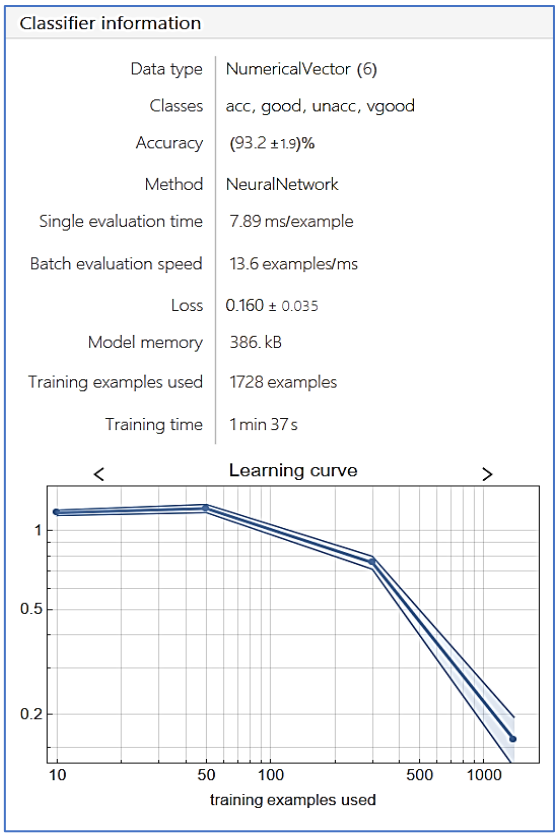

Рис. 29 Информация о классификаторе (метод Многослойной нейронной сети)

Оценка точности построенного классификатора представлена на рисунке 30.

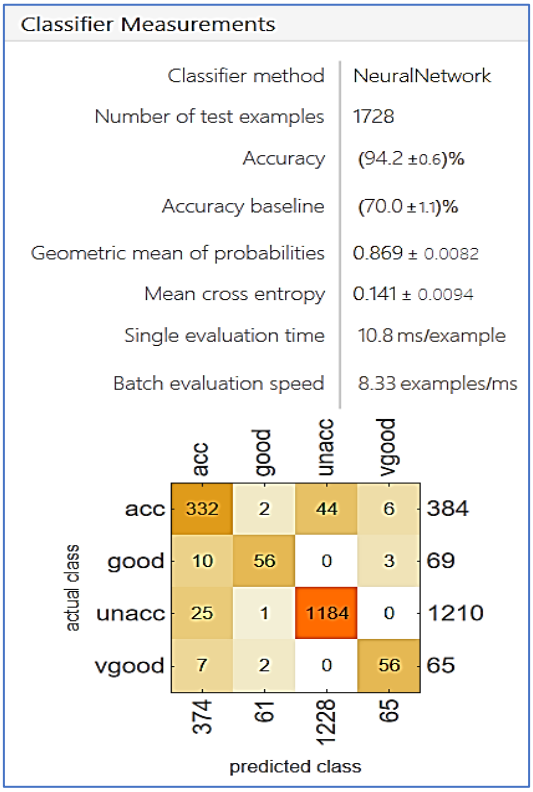

Рис. 30 Матрица ошибок обучения нейросетевого классификатора

5.2.4 Сравнение методов обучения

В таблице 10 представлена точность обучения классификаторов исследуемыми методами машинного обучения

Таблица 10

Сравнение точности обучения

Метод машинного обучения	Логистическая регрессия	Модель Маркова	Нейронная сеть
Точность классификации, %	~69,5	~70	~94,2

Выводы и заключение

В результате исследования мы получили таблицу 11:

Таблица 11

Итоговая таблица сравнения различных форматов представления исходных данных

Метод обучения	Логистическая регрессия	Модель Маркова	Нейронная сеть	
Представление	Точность обучения, %			Средняя точность, %
Англоязычные исходные данные	~87,1	~88,3	~99,7	~91,7
ЗЗОднословные русские данные	~93,5	~88,3	~100	~93,93
Многословные русские данные	~93,6	~88,3	~100	~93,97
Оцифрованные вручную исходные данные	~83,9	~73,6	~99,9	~85,8
Оцифрованные программой исходные данные	~69,5	~70	~94,2	~77,9

По данным таблицы видно, что с небольшим отрывом в ~2% (в пределах погрешности) с англоязычными данными (~91,7%) однословные русские данные (~93,93%) и многословные русские данные (~93,97) оказались наиболее пригодны для обучения классификаторов выбранными методами.

Худшими для обучения оказались исходные данные, оцифрованные встроенными в ПО Wolfram Mathematica средствами: понижением размерности (~77,9%). Оцифрованные вручную данные показали хороший результат (~85,8%), однако он является на порядок ниже, чем результат языковых данных.

Делая вывод, отметим, что языковые исходные данные в текущих условиях показали результат лучший, нежели цифровые исходные данные.

Список литературы:

1. Car Evaluation Data Set. URL: https://archive.ics.uci.edu/ml/datasets/car+evaluation
2. Wolfram Mathematica Official WebsiteURL: https://www.wolfram.com/mathematica/
3. Метод Логистической Регрессии. URL: https://habr.com/ru/companies/io/articles/265007/
4. Модель Маркова.URL: https://habr.com/ru/articles/455762/
5. Метод Многослойной Нейронной Сети.URL: https://habr.com/ru/articles/198268/