DOI 10.37539/2949-1991.2023.3.3.019
УДК 004.855.5

Рогаль Сергей Александрович, студент,
Сахалинский государственный университет, Южно-Сахалинск
Rogal Sergey Alexandrovich, Sakhalin State University, Yuzhno-Sakhalinsk

Шибанов Вячеслав Сергеевич, студент,
Сахалинский государственный университет, Южно-Сахалинск
Shibanov Vyacheslav Sergeevich, Sakhalin State University, Yuzhno-Sakhalinsk

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

ИССЛЕДОВАНИЕ МЕТОДОЛОГИИ СНИЖЕНИЯ РАЗМЕРНОСТИ ОБУЧАЮЩЕЙ ВЫБОРКИ В ЗАДАЧЕ СИНТЕЗА КЛАССИФИКАТОРОВ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

INVESTIGATION OF THE METHODOLOGY FOR REDUCING THE DIMENSION OF THE TRAINING SAMPLE IN THE TASK OF CLASSIFIER SYNTHESIS BY MACHINE LEARNING METHODS

Аннотация: Исследуется проблема снижения размерности обучающей выборки, представляющей собой большой массив разнородных данных по критерию минимизации потери полезной информации. Предложена методология решения задачи синтеза и снижения размерности классификаторов на базе методов машинного обучения. Произведена практическая апробации методологии решения задачи в среде пакета символьной математики.

Abstract: The problem of reducing the dimension of the training sample, which is a large array of heterogeneous data by the criterion of minimizing the loss of

useful information, is investigated. A methodology for solving the problem of synthesis and dimensionality reduction of classifiers based on machine learning methods is proposed. The practical approbation of the methodology for solving the problem in the environment of the symbolic mathematics package is carried out.

Ключевые слова: методы машинного обучения, проблема классификации объектов.

Keywords: machine learning methods, the problem of object classification.

Введение

Одно из ключевых направлений развития систем искусственного интеллекта в настоящее время базируется на формализации исследуемых систем, представляемых большими массивами данных, методами машинного обучения с использованием технологий самообучения, бустинга и снижения размерности обучающей выборки.

Практически значимые обучающие выборки имеют большое количество переменных (характеристик, факторов), определяющих функционирование исследуемых систем или объектов. Поэтому при построении классификаторов таких систем на реальных исходных данных возникает, так называемое, «проклятие размерности», которое практически сводит на нет решение проблемы классификации и требует снижения количества используемых переменных, синтезируя из них наборы меньшей размерности.

Поэтому целью настоящего исследования является отработка методологии снижения размерности входного массива данных по критерию минимизации потери полезной информации при синтезе интеллектуальных классификаторов методами машинного обучения.

Формальная постановка задачи

На основании экспертных заключений известна базовая классификация вида:

$$\boldsymbol{x} = \left(x_1, x_2, \ldots, x_n \right) \Rightarrow y .$$

Количество информации, содержащейся в классификаторе, обозначим через $I(x \Rightarrow y)$.

Требуется построить семейство классификаторов $\mathbf{z}_i$ размерности $m<n$ и выбрать из них тот, который обеспечивает минимальную потерю полезной информации, т.е.:

$$\left\| I(x \Rightarrow y) - I\left(z_i \underset{i \le m}{\Rightarrow} y\right) \right\| \to \min .$$

Термин «количество информации» в классификаторе определяется индивидуально для каждой задачи, это может быть, например, процент правильно классифицированных им объектов.

Используемые методы и методологии

Построение классификаторов осуществлялось в среде системы символьной математики Wolfram Mathematica, являющейся одной из наиболее универсальных аналитических платформ, ориентированных на решение задач искусственного интеллекта и машинного обучения [1].

На рисунке 1 приведен фрагмент текста в среде Wolfram Mathematica, реализующий обучение системы классификации, например, нейросетевым методом.

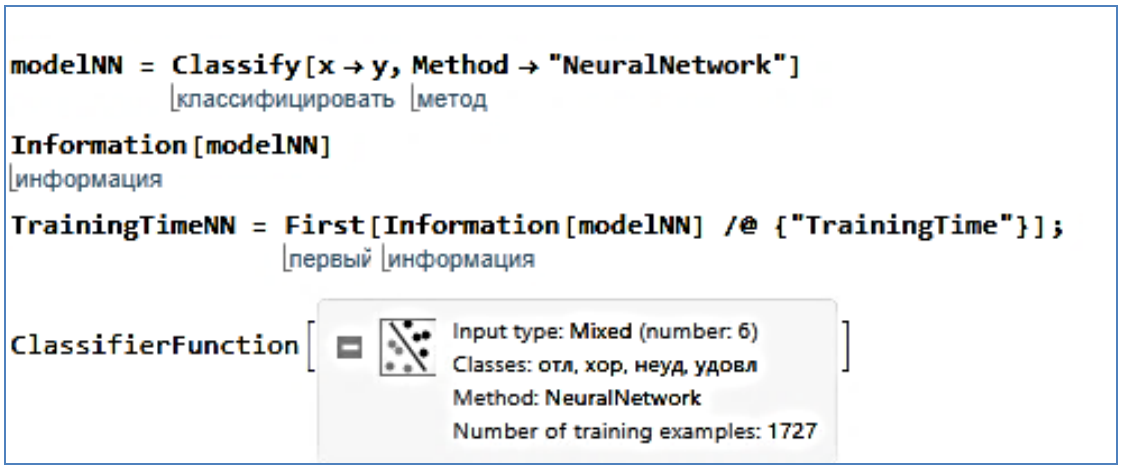

Рис. 1 Организация обучения модели.

В данном случае осуществлялось априорное разбиение исходных объектов [2], характеризуемых смешанными (символьными, буквенными и цифровыми показателями) на 4 класса («отл», «хор», «удовл» и «неуд»).

Информация о параметрах обучения классификатора представлена на рисунке 2.

Out[34]=

Classifier information

Data type	Mixed (number: 6)
Classes	отл, хор, неуд, удовл
Accuracy	(99.1 ±0.7)%
Method	NeuralNetwork
Single evaluation time	5.98 ms/example
Batch evaluation speed	13.3 examples/ms
Loss	0.0305 ± 0.015
Model memory	514. kB
Training examples used	1727 examples
Training time	52.8 s

Рис. 2 Информация о классификаторе

Кривая обучения и изменение точности классификации в процессе обучения модели представлены на рисунке 3.

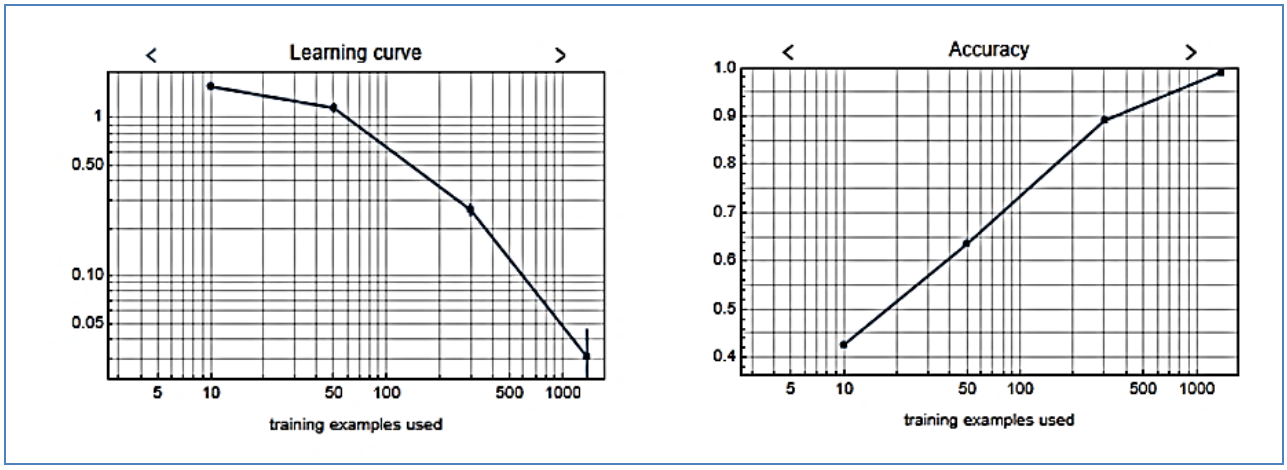

Рис. 3 Графики кривых обучения

В таблице 1 для исследуемых методов машинного обучения [3] представлены их идентификаторы, которые для краткости записи используются в дальнейшем и точность решения задачи классификации, рассчитываемая как процент отношения правильно классифицируемых объектов к их общему числу в обучающей выборке.

Таблица 1

Оценки точности классификации

№	Метод	ID	Точность %
1	Logistic Regression	LGR	94
2	Decision Tree	DTR	81
3	Gradient Boosted Trees	GBT	98
4	Markov	MM	92
5	Naïve Bayes	NB	87
6	Neural Network	NN	100
7	Nearest Neighbors	NNB	80
8	Random Forest	RF	95
9	Support Vector Machine	SVM	93

На рисунке 4 представлена диаграмма абсолютных величин точности решения задачи классификации для 9 методов машинного обучения (в соответствии с данными из таблицы 1).

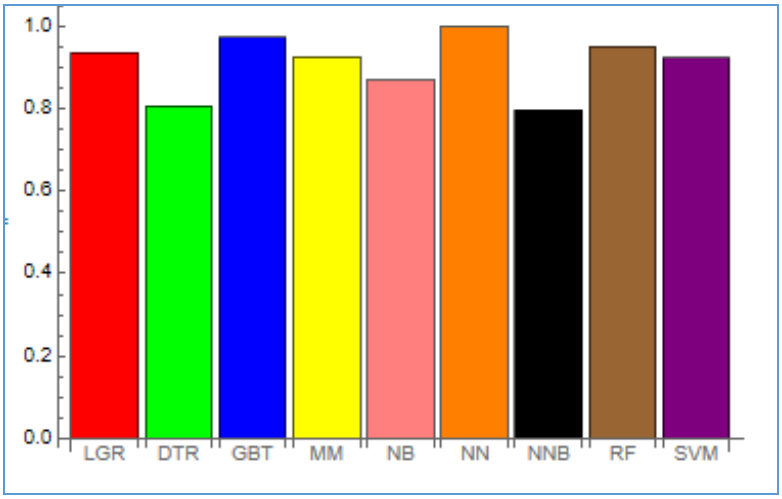

Рис. 4 Диаграмма величин точности классификации.

Основные результаты

На рисунке 5 в качестве примера приведена полная информация о тестировании построенного нейросетевого классификатора по всей обучающей выборке. В данном случае точность классификации составляет 100%, ошибок нет, все 1727 объекта классифицированы верно.

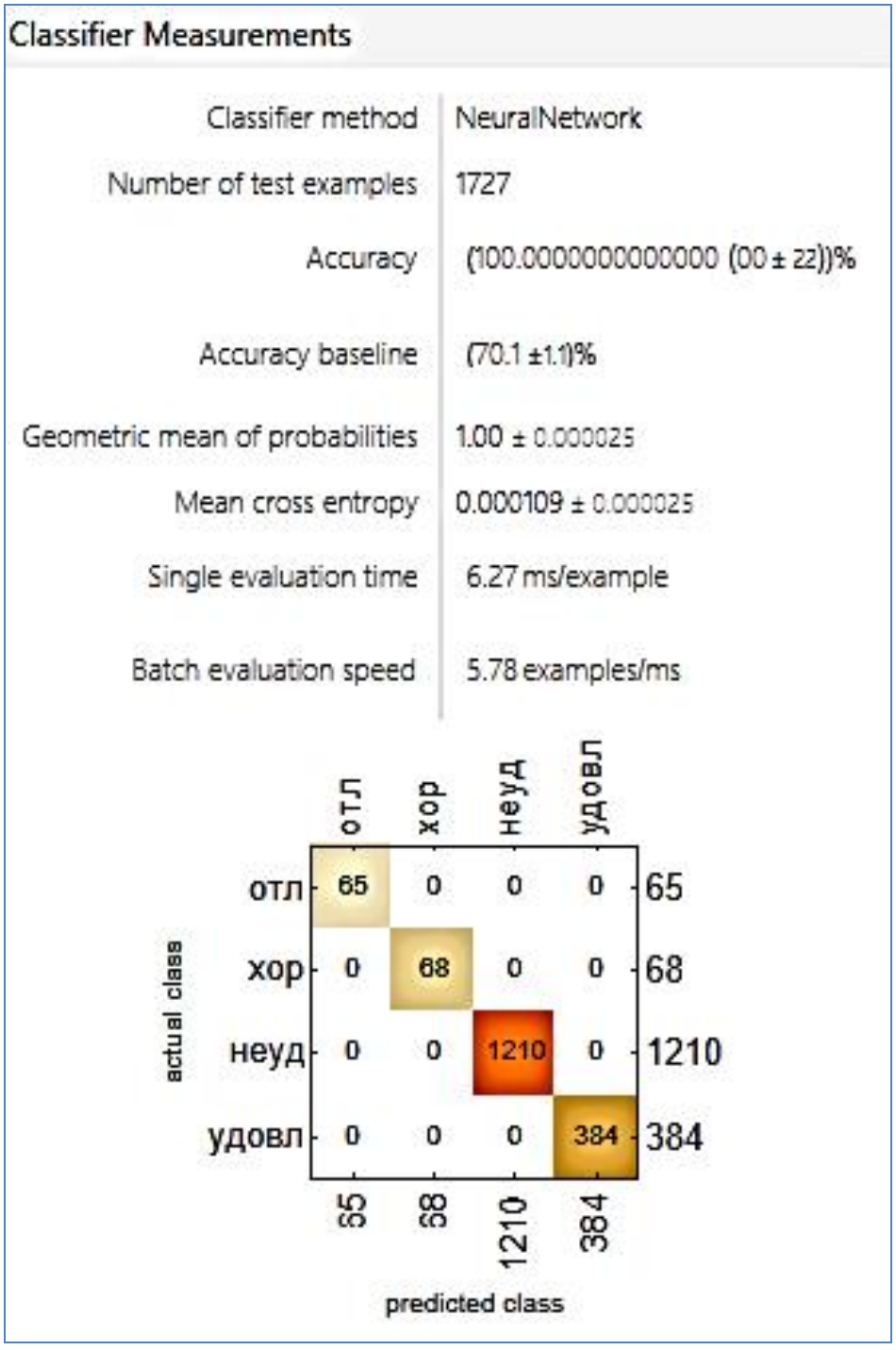

Рис 5 Результаты тестирования обученного классификатора

В таблице 2 представлена информация о точности классификации для всех девяти исследуемых методов машинного обучения для размерности задачи от исходной 6 до 4. Здесь же представлено относительное время обучения моделей, где за 100% взято время обучения при исходной размерности равной 6.

Таблица 2

Сравнение методов классификации

№	ID метода	Точность классификации %			Относительное время обучение, %		
		6	5	4	6	5	4
1	LGR	94	71	70	100	82	81
2	DTR	81	92	90		54	56
3	GBT	98	95	95		60	72
4	MM	92	84	78		74	71
5	NB	87	86	83		54	52
6	NN	100	89	76		23	27
7	NNB	80	95	95		74	74
8	RF	95	93	91		76	78
9	SVM	93	84	74		93	84

Выводы и заключение

Анализ полученных решений позволяет сделать следующие выводы:

1. При синтезе нейросетевого классификатора для исходной размерности обучающей выборки равной 6 точность классификации при тестировании составляет 100% (ошибок нет). При снижении размерности до 5 точность снижается только на 11% (до 89%) в то время, как относительное время построения модели снижается существенно со 100% до 23%.

2. Для метода градиентного бустинга показательно то, что снижение размерности практически не приводит к потере полезной информации, а время построения модели снижается значительно (со 100% до 60%).

3. Отметим, что метод дерева решений демонстрирует увеличение точности решения при снижении размерности задачи одновременно снижая время построения модели почти на 50%

4. В качестве примера на рисунке 6 представлены диаграммы точности и продолжительности построения модели методом случайного леса. Для данного метода характерно плавное незначительное снижение количества полезной информации при уменьшении размерности обучающей выборки и сокращение времени обучения практически на 25%.

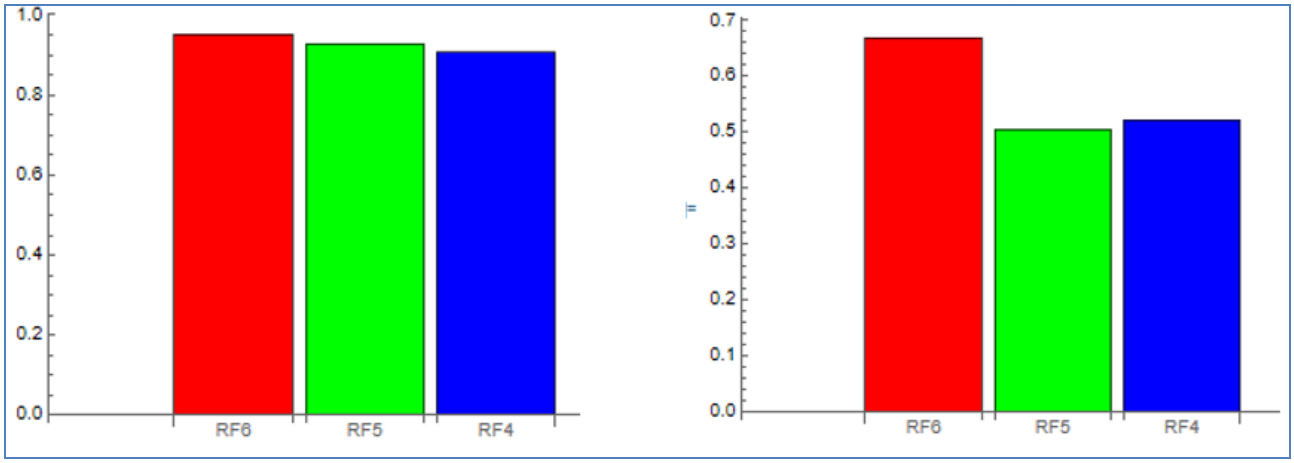

Рис. 6 Показатели обучения методом случайного леса

Предложенная методологии снижения размерности обучающей выборки в задаче синтеза классификаторов методами машинного обучения является унифицированной и может быть использована в задачах минимизации потерь полезной информации для обработки и редукции больших массивов исходных данных.

Список литературы:

1. Stephen Wolfram. An Elementary Introduction to the Wolfram Language. URL: https://www.wolfram.com/language/elementary-introduction/2nd-ed/ (Дата обращения 22.04.2023).

2. Классификация объектов, характеризуемых показателями различных типов / Г. С. Осипов, Н. С. Вашакидзе, Г. В. Филиппова, Н. Л. Рауш // Научные исследования по перспективным направлениям как основа инновационного совершенствования: сборник статей международной научной конференции,

Санкт-Петербург, 17 января 2023 года. – Санкт-Петербург: Общество с ограниченной ответственностью «Международный институт перспективных исследований имени Ломоносова», 2023. – С. 19-24. – DOI 10.58351/230117.2023.74.86.002. – EDN YRSPNP.

3. Введение в сравнительный анализ методов машинного обучения для решения задачи прогнозирования / П. В. Витковская, С. А. Рогаль, В. С. Шибанов // Молодые исследователи в ответ на современные вызовы: сборник статей II Международного научно-исследовательского конкурса, Петрозаводск, 09 ноября 2022 года. – Петрозаводск: Международный центр научного партнерства «Новая Наука», 2022. – С. 164-170. – EDN ENCXXD.