DOI: 10.51790/2712-9942-2023-4-3-05

ИССЛЕДОВАНИЕ ПРИМЕНИМОСТИ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ЗАДАЧИ ИДЕНТИФИКАЦИИ ТИПА ЗАВИСИМОСТИ В НАБОРАХ ДАННЫХ

А. Д. Смородинов[1,2,а], **Т. В. Гавриленко**[1,2,б], **А. А. Рассадин**[1,в]

[1] *Сургутский филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», г. Сургут, Российская Федерация*

[2] *Сургутский государственный университет, г. Сургут, Российская Федерация*

[а] *ORCID: https://orcid.org/0000-0002-9324-1844,* ✍ *Sachenka_1998@mail.ru*

[б] *ORCID: https://orcid.org/0000-0002-3243-2751, taras.gavrilenko@gmail.com*

[в] *ORCID: https://orcid.org/0000-0001-5596-0891, rassadin_aa@office.niisi.tech*

Аннотация: в работе рассматривается возможность применения сверточных искусственных нейронных сетей для решения задачи идентификации типа зависимости в наборах данных. Для обучения сверточной нейронной сети использовались обучающие выборки, состоящие из графиков функций. В качестве базового набора функций для обучения искусственной нейронной сети были выбраны 6 функций, ключевым свойством которых является линеаризуемость. Сверточная нейронная сеть применялась для определения типа зависимостей в наборах данных, полученных из международной базы данных MNIST, предназначенной для тестирования статистического программного обеспечения. Приведен анализ результатов применения сверточной нейронной сети к наборам данных из базы данных MNIST, делается вывод о принципиальной возможности применения данного подхода для визуального корреляционного анализа данных и, как следствие, возможности идентификации типа зависимости по графическому образу представления данных.

Ключевые слова: искусственные нейронные сети, аппроксимация, сверточные нейронные сети, корреляционный анализ.

Благодарности: работа выполнена в рамках государственного задания ФГУ ФНЦ НИИСИ РАН (Выполнение фундаментальных научных исследований ГП 47) по теме №0580-2021-0007 «Развитие методов математического моделирования распределенных систем и соответствующих методов вычисления».

Для цитирования: Смородинов А. Д., Гавриленко Т. В., Рассадин А. А. Исследование применимости сверточных нейронных сетей для задачи идентификации типа зависимости в наборах данных. *Успехи кибернетики*. 2023;4(3):47–54. DOI: 10.51790/2712-9942-2023-4-3-05.

Поступила в редакцию: 18.08.2023. В окончательном варианте: 10.09.2023.

APPLICABILITY OF CONVOLUTED NEURAL NETWORKS TO THE DATASET FITTING PROBLEM

A. D. Smorodinov[1,2,a], **T. V. Gavrilenko**[1,2,b], **A. A. Rassadin**[1,c]

[1] *Surgut Branch of Federal State Institute "Scientific Research Institute for System Analysis of the Russian Academy of Sciences", Surgut, Russian Federation*

[2] *Surgut State University, Surgut, Russian Federation*

[a] *ORCID: https://orcid.org/0000-0002-9324-1844,* ✍ *Sachenka_1998@mail.ru*

[b] *ORCID: https://orcid.org/0000-0002-3243-2751, taras.gavrilenko@gmail.com*

[c] *ORCID: https://orcid.org/0000-0001-5596-0891, rassadin_aa@office.niisi.tech*

Abstract: we analyzed the applicability of convolutional neural networks to solving the dataset fitting problem. The convolutional neural network was trained with training datasets containing function curves. We selected 6 linearizable functions. The convolutional neural network detected the functional relations in the datasets taken from the MNIST database intended for statistical software testing. The results show that it is possible to use the proposed approach for visual correlation analysis and curve-based data fitting.

Keywords: artificial neural networks, data fitting, convolutional neural networks, correlation analysis.

Acknowledgements: this study is a part of the GP 47 government order contracted to the Scientific Research Institute for System Analysis of the Russian Academy of Sciences, phase No. 0580-2021-0007 Advanced Simulation of Distributed Systems.

Cite this article: Smorodinov A. D., Gavrilenko T. V., Rassadin A. A. Applicability of Convoluted Neural Networks to the Dataset Fitting Problem. *Russian Journal of Cybernetics*. 2023;4(3):47–54. DOI: 10.51790/2712-9942-2023-4-3-05.

Original article submitted: 18.08.2023. *Revision submitted: 10.09.2023*.

Введение

В 1637 году Рене Декарт в своей работе [1] первым применил такое понятие, как координаты, которое впоследствии получило название «прямоугольные декартовы координаты». Это позволило перейти к анализу функций, данных и многих других объектов математики на абсолютно новом уровне. Графическое представление стало играть решающую роль. Но именно благодаря его идеям, описанным в данном труде, появились новые направления развития целого ряда разделов современной математики.

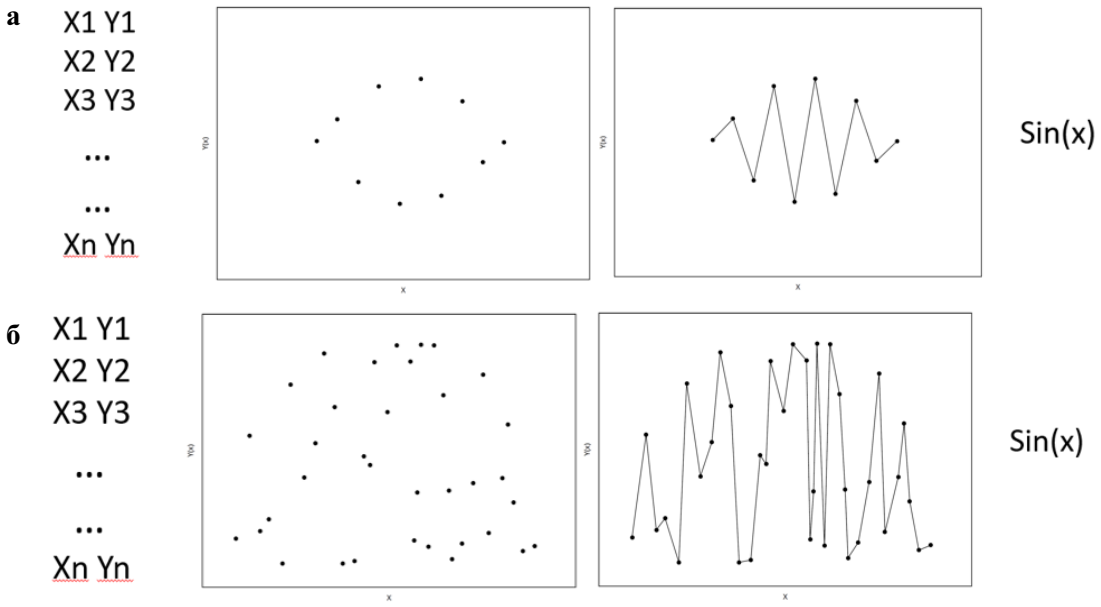

Рис. 1. *Иллюстрация последовательности действий исследователя*

Исследователь или аналитик данных начинает свою работу с того, что, используя различные графические отображения, формирует визуальные образы, после чего на основе полученного визуального образа данных пытается предугадать вид зависимости. Но чаще всего аналитику требуется перейти ко второму этапу, соединить последовательные точки прямой линией для формирования представления о законах развития процесса. На рис. 1а представлена типичная схема работы исследователя. Исследователь строит «фоторобот» некоторой зависимости по имеющимся у него описаниям — точкам. Потом он ищет совпадение с имеющимися в его голове «фотороботами» других зависимостей. Но такая ситуация получается не всегда. Так, на рис. 1б показан случай, когда исследователь сталкивается с существенными затруднениями.

Проблема выбора функции для аппроксимации данных

При решении подобных задач определяющим фактором является опыт и кругозор исследователя. Но даже опытный аналитик не всегда может сходу определить тип зависимости. Так, на рис. 2 приведен ряд примеров графиков, при определении типа зависимости которых могут возникнуть проблемы даже у опытного исследователя. Графики похожи, а типы зависимостей нет. Если в некоторых случаях ошибка выбора типа зависимости при аппроксимации может и не сказаться на результатах расчета, то в случаях, показанных на рис. 2, она может привести к значительной погрешности. Особенно критичны ошибки второго рода.

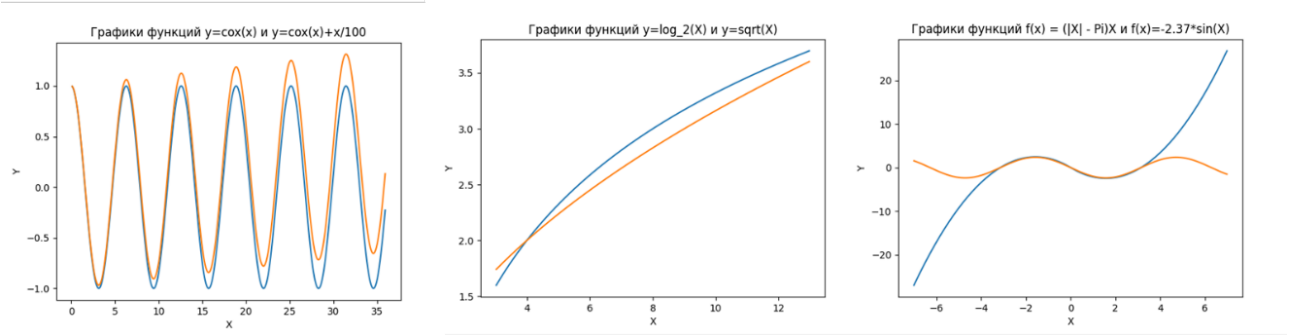

Рис. 2. *Проблема определения типа зависимости по графику функции*

Одни и те же данные можно аппроксимировать разными функциями. На рис. 3 красным нарисован сектор синуса, и, если у исследователя есть только часть данных (красные точки, соединённые прямой), то такие данные можно аппроксимировать целым набором данных, таким как:

$$y_1 = \frac{1.2}{exp\left(\left(x - \frac{\pi}{2.0}\right)^2\right)}, \quad y_2 = \frac{x^2 + \pi x}{2.5}, \quad y_3 = \frac{1}{\left(x - \frac{\pi}{2}\right)^2 + 1}$$

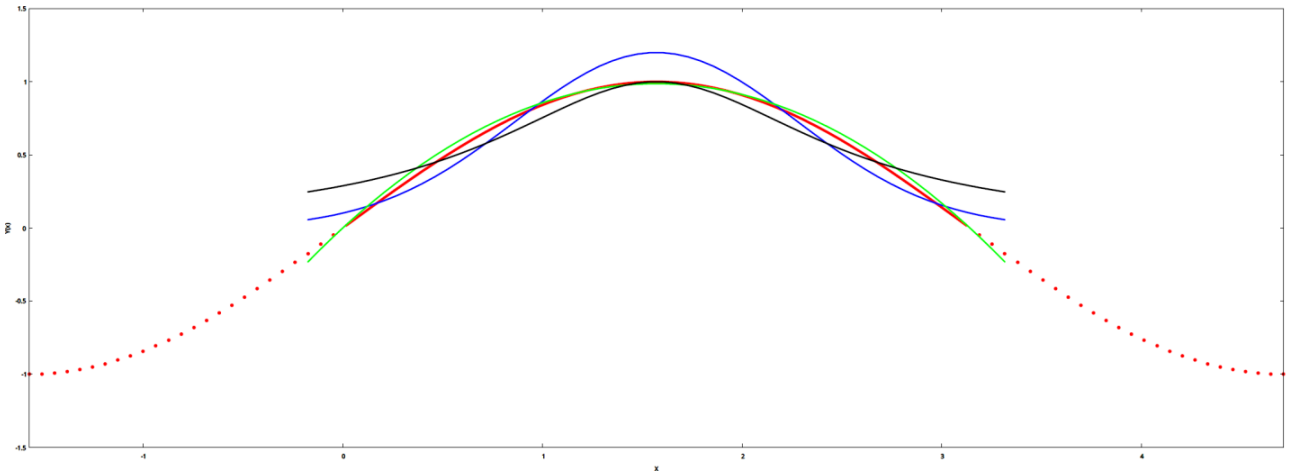

Рис. 3. *Пример аппроксимации одного набора данных серией различных функций*

Типов зависимостей бесконечное количество, например, в справочнике для студентов [2] насчитывается 1200 различных функций и их графиков. Также стоит отметить, что некоторые функции имеют различные графики функций в зависимости от значений коэффициентов. Упростить работу исследователя может применение искусственной нейронной сети (ИНС) для визуального корреляционного анализа данных.

Описание методики выбора аппроксимирующей функции

В одной из работ уже была продемонстрирована возможность применения ИНС для решения задачи аппроксимации и интерполяции [3], в работе представлено решение задачи аппроксимации — выявление зависимости как суммы функций с некоторым набором коэффициентов, где в качестве коэффициентов функций использованы весовые коэффициенты ИНС. Был разработан программный продукт [4], который позволяет упростить работу аналитика данных. Получаемый результат не является универсальным и требует повторного обучения ИНС для решения новой задачи.

В данной работе используются сверточные нейронные сети для определения класса зависимости по графическому образу исходных данных. На вход нейронной сети подаётся изображение графика, а на выходе получаем тип зависимости, функции. Затем исследователь может использовать любой известный ему метод для определения значений коэффициентов функций, например, метод наименьших

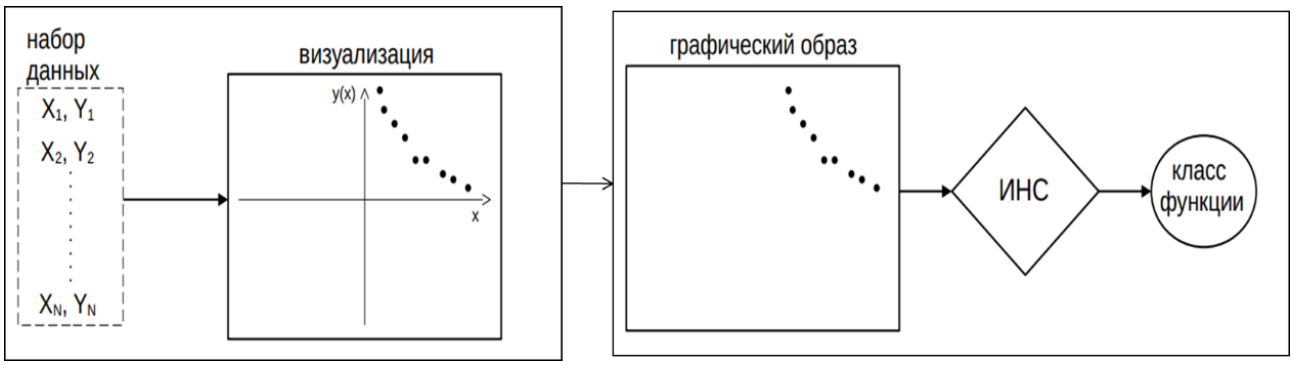

Рис. 4. *Схема работы исследователя с ИНС*

квадратов (МНК). В результате решается задача аппроксимации данных. На рис. 4 представлена схема работы разработанного инструмента.

На первом этапе разработки необходимо установить количество определяемых зависимостей. В рамках первичной проверки возможности применения данного подхода были выбраны простые и наиболее часто применяемые исследователями начального уровня зависимости:

$$y = ax^b \tag{1}$$

$$y = a\ln(x) + b \tag{2}$$

$$y = ax^b \tag{3}$$

$$y = \frac{a}{x+b} \tag{4}$$

$$y = \frac{1}{ax+b} \tag{5}$$

$$y = \frac{x}{ax+b} \tag{6}$$

Данные функции были выбраны, т. к. они легко линеаризуются, к ним применим МНК.

Важным этапом является подготовка обучающих данных, т. к. это напрямую влияет на работу ИНС. Обучающая выборка представляет собой набор изображений с графиками функций (1) - (6). Для каждого из 6 классов (функций) этот набор содержит 1200 изображений, которые делятся на три блока: 1000 изображений — обучающая выборка, 100 — контрольная и 100 — тестовые выборки данных.

Для формирования графиков были определены следующие требования:

- График должен быть представлен в виде множества точек, т. к. соединение их прямыми линиями уже является самой простой аппроксимацией.
- Стиль всех графиков должен быть одинаковым (вид точек, их размер и цвет и т. д.), т. к. стиль графика не влияет на тип зависимости.
- На изображениях не должно быть осей координат, подписей и т. д. — в данном случае все это является «шумом», который может помешать обучению. Координатная сетка исключена в связи с тем, что в первую очередь идентифицируется класс зависимости, а не конкретные значения коэффициентов функций.
- В пределах одного класса все графики должны быть разными (уникальными).

Описание эксперимента

Ввиду того, что готового набора данных, удовлетворяющего перечисленным выше требованиям, не существует, была создана программа для генерации обучающих данных [7]. Для получения уникальных графиков для каждого класса из формул (1) - (6) коэффициентам a, b задаётся диапазон допустимых значений, и при построении отдельного графика конкретные значения коэффициентов выбираются случайным образом из соответствующего диапазона. В данном случае последние три функции из (1) - (6) — это гиперболы, для них область определения и диапазон значений коэффициентов

выбраны так, чтобы получалась только одна ветвь. Для всей обучающей выборки задается шаг построения графиков, а также плотность точек. Плотность определяет, какой процент от всех расчетных точек будет нанесен на изображение. Расчетные значения точек графиков получаются с помощью программы на языке C, а изображение строится с помощью утилиты Gnuplot. Формат изображения растровый (png) с разрешением 640x480. Плотность точек была равна 100%. Пример обучающих изображений представлен на рис. 5.

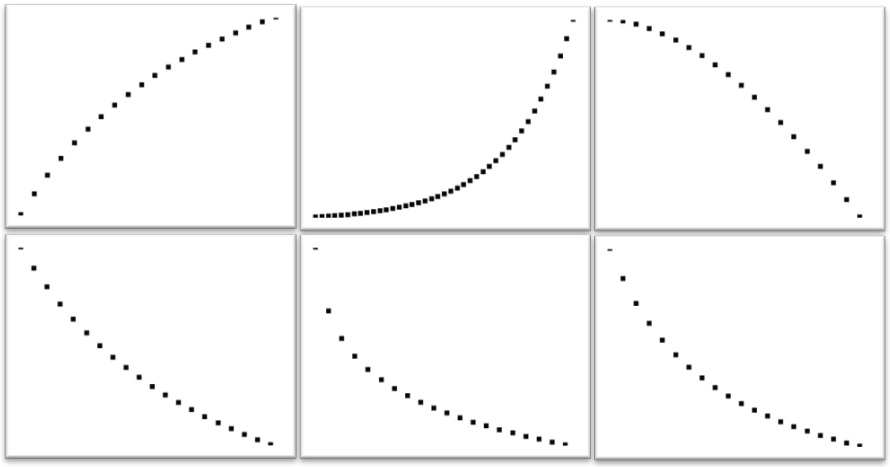

Рис. 5. *Пример обучающей выборки для каждого класса*

После подготовки обучающих данных была сконструирована сверточная ИНС следующей конфигурации: 5 сверточных слоев с окном (3, 3) и 32, 64, 128, 258 и 512 нейронами на слоях, функции активации relu, 1 полносвязный слой с 1024 нейронами, функции активации relu, 1 полносвязный слой с 6 нейронами, функции активации softmax. Схема обучаемой ИНС представлена рис. 6.

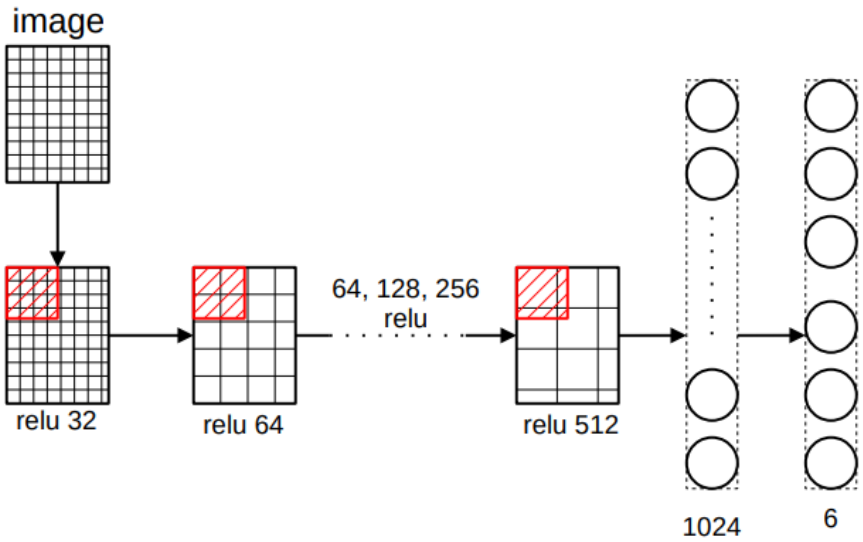

Рис. 6. *Схема сверточной ИНС для решения задачи классификации графика по его графическому образу*

При такой конфигурации ИНС точность на тестовых данных была максимальной. Обучение проходило в течение 72 эпох. Скорость обучения автоматически изменялась от 10^{-3} до 10^{-7}. В качестве алгоритма оптимизации был выбран алгоритм RMSProp, а в качестве функции потерь — categorical_crossentropy. Они были выбраны, т. к. именно их используют для решения задачи классификации во многих проектах.

По итогам обучения точность на контрольном наборе данных составила 86,7 %. Следующим шагом была проверка адекватности работы ИНС на данных, не относящихся к обучающей выборке. Задача по поиску таких данных не является тривиальной, т. к. задача нелинейной регрессии достаточно сложна, и чаще всего можно найти такой набор данных, на котором не будут работать даже самые эффективные алгоритмы. Также в работе [6] автор поднимает вопрос о сложностях в тестировании программного обеспечения для оценки качества программных средств. Но в настоящее время для этих целей существует специальная стандартная справочная база данных NIST 140, содержащая набор данных, на которых можно проводить тестирование математического программного кода [5]. Целью этой БД является повышение точности статистического программного обеспечения путем предоставления справочных наборов данных с сертифицированными результатами вычислений, которые позволяют объективно оценивать статистическое программное обеспечение. Поэтому для оценки качества работы ИНС была взята коллекция из 27 наборов данных из этой БД. В ней содержатся как реальные (наблюдаемые) данные, так и сгенерированные. В данной базе наборы разделены на 3 уровня сложности: легкие, средние и сложные. ИНС была протестирована на всех наборах данных из этой БД. На основе данных строились аналогичные изображения, примеры изображений представлены на рис. 7.

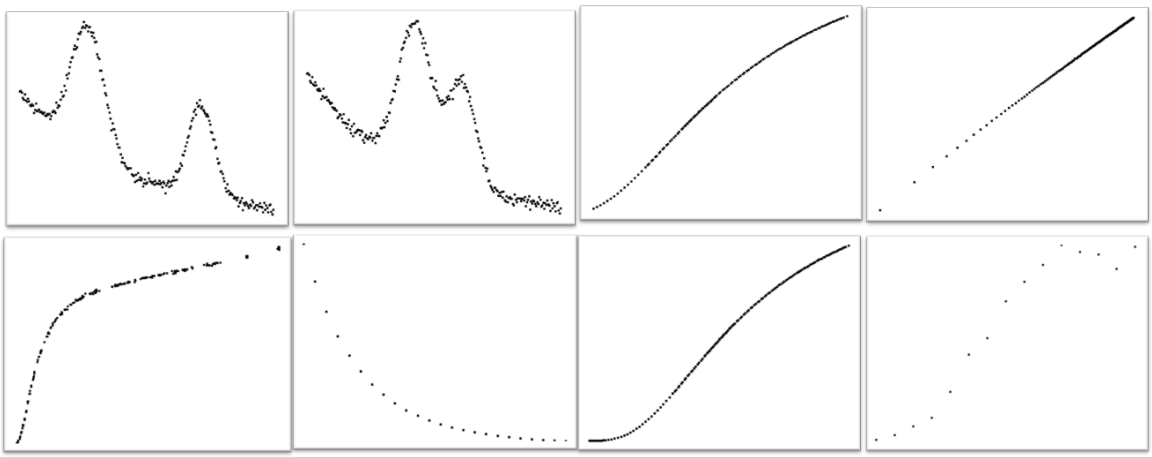

Рис. 7. *Примеры построенных изображений на данных из БД NIST 140*

После чего изображения подавались на вход в обученную ИНС. Затем на основе тех же данных строили аппроксимацию с использованием МНК с двумя параметрами, где в качестве аппроксимирующих функций использовались функции (1) - (6). В качестве типа функции выбирали ту функцию, которая имеет наименьшую погрешность. Схема проведенного эксперимента представлена на рис. 8.

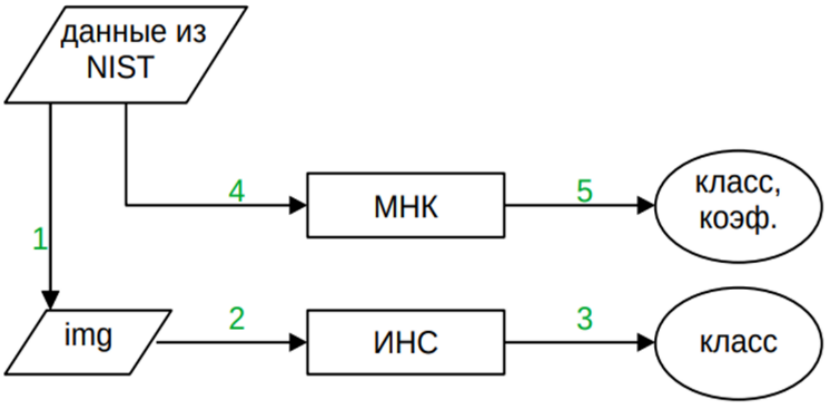

Рис. 8. *Схема проведения эксперимента для оценки адекватности обученной модели*

Результаты проведенного эксперимента

Был проведен анализ полученных результатов: изучали количество совпадений между ответами

ИНС, типом, предлагаемым МНК, и типом, указанным в БД. Полученные результаты представлены в таблице 1. Красным цветом выделены ячейки таблицы, если совпадения есть только между БД и типом, полученным из МНК, — 5 случаев. Синим цветом — если есть совпадения только между БД и ответом ИНС — 10 случаев. Зеленым цветом выделены совпадения между всеми тремя способами определения типа — 6 случаев. Белым — те случаи, в которых ни МНК, ни ИНС не смогли дать верный ответ, — 6 случаев.

Таблица 1

Результаты проведенного вычислительного эксперимента

Номер набора данных	**Число параметров**	**Число точек**	**Тип данных**	**Ответ ИНС**	**Ответ МНК**	**Ответ сайта**
1	2	14	Реальные	0	5	0
2	3	54	Реальные	0	2	0
3	3	214	Реальные	0	1	0
4	6	24	Сгенерированные	0	0	0
5	8	250	Сгенерированные	0	0	0
6	8	250	Сгенерированные	0	1	0
7	2	6	Реальные	0	2	3,4,5
8	2	14	Реальные	0	5	3,4,5
9	5	151	Реальные	0	1	3,4,5
10	7	236	Реальные	0	1	3,4,5
11	3	128	Реальные	0	Nan	0
12	5	33	Сгенерированные	0	0	0
13	6	24	Сгенерированные	0	0	0
14	6	24	Сгенерированные	0	0	0
15	8	250	Сгенерированные	0	1	0
16	2	14	Реальные	0	5	3,4,5
17	2	14	Реальные	0	5	3,4,5
18	4	25	Реальные	2	1	3,4,5
19	9	168	Реальные	0	1	3,4,5
20	4	11	Сгенерированные	2	5	3,4,5
21	7	37	Реальные	2	1	3,4,5
22	2	6	Реальные	0	1	0
23	3	9	Реальные	0	2	0
24	3	16	Сгенерированные	0	0	0
25	3	35	Реальные	0	1	0
26	4	15	Реальные	0	1	0
27	3	154	Реальные	0	4	3,4,5

Заключение

Установлено, что точность определения типа функций с помощью ИНС составляет почти 60 %, точность определения типа функции с помощью МНК и базовых функций составляет 40 %. В результате проведенного вычислительно эксперимента была показана принципиальная возможность применения ИНС для визуального корреляционного анализа наборов экспериментальных данных. В ряде случаев эксперимент демонстрирует высокую степень эффективности. Естественным образом повышение эффективности и качества визуального корреляционного анализа, проводимого с помощью ИНС, напрямую зависит от ее масштаба, и точности, и качества обучения. В настоящий момент проводится серия экспериментов, направленных на расширение обучающей выборки и количества классов функций.

ЛИТЕРАТУРА

1. Декарт Р. *Геометрия. С приложением избранных работ П. Ферма и переписки Декарта.* Пер., примеч. и статьи А. П. Юшкевича. М.-Л.: Гостехиздат; 1938. 297 с.

2. Старков С. Н. *Справочник по математическим формулам и графикам функций для студентов*. СПБ.: Питер; 2009. 235 с.
3. Галкин В. А., Гавриленко Т. В., Смородинов А. Д. Некоторые аспекты аппроксимации и интерполяции функций искусственными нейронными сетями. *Вестник КРАУНЦ. Физико-математические науки*. 2022;38(1):54–73. DOI: 10.26117/2079-6641-2022-38-1-54-73. EDN: JTZQQZ.
4. *Свидетельство о государственной регистрации программы для ЭВМ №2023611413 Российская Федерация. Автоматизированная система конструирования нейронных сетей: № 2022680250: заявл. 26.10.2022: опубл. 19.01.2023 /* А. Д. Смородинов, Т. В. Гавриленко, Н. Р. Урманцева. EDN: AWFDRA.
5. *Стандартная справочная база данных NIST 140*. Режим доступа: https://www.itl.nist.gov/div898/strd/nls/nls_main.shtml. DOI: 10.18434/T43G6C.
6. Hiebert K. L. An Evaluation of Mathematical Software That Solves Nonlinear Least Squares Problems. *ACM Transactions on Mathematical Software*. 1981;7(1):1–16. DOI: 10.1145/355934.355935.
7. *Свидетельство о государственной регистрации программы для ЭВМ №2023661479 Российская Федерация. Программа автоматической генерации изображений графиков функций для обучения искусственной нейронной сети распознавания образов: №2023619894: заявл. 18.05.2023: опубл. 31.05.2023* / А. А. Рассадин, А. Д. Смородинов, Т. В. Гавриленко. EDN: DLAXJE.