ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2021, № 6, с. 106–121

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 510.64–519.816

ОБЗОР ИССЛЕДОВАНИЙ В ОБЛАСТИ РАЗРАБОТКИ МЕТОДОВ ИЗВЛЕЧЕНИЯ ПРАВИЛ ИЗ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ[1]

© 2021 г. А. Н. Аверкин[а,б,*], С. А. Ярушев[б,]**

[а]ФИЦ ИУ РАН, Москва, Россия

[б]РЭУ им. Г.В. Плеханова, Москва, Россия

**e-mail: averkin2003@inbox.ru*

***e-mail: sergey.yarushev@icloud.com*

Поступила в редакцию 29.06.2021 г.
После доработки 03.07.2021 г.
Принята к публикации 26.07.2021 г.

Проводится масштабный обзор и анализ существующих методов и подходов к извлечению правил из искусственных нейронных сетей, в том числе из нейронных сетей глубокого обучения. Рассмотрены широкий спектр методов и подходов к извлечению правил, а также связанные с этим подходы к разработке систем объяснимого искусственного интеллекта. Исследуется таксономия и несколько направлений в изучении объяснимых нейронных сетей, связанных с извлечением правил из нейронных сетей, которые позволяют пользователю получить представление о том, как нейронная сеть использует входные данные, а также при помощи правил выявить скрытые взаимосвязи входных данных и найденных результатов. Таким образом в настоящем обзоре делается акцент на связи наиболее распространенных в искусственном интеллекте систем объяснений на основе правил с наиболее мощными алгоритмами машинного обучения с помощью нейросетей. Помимо извлечения правил рассматриваются и другие методы построения систем объяснимого искусственного интеллекта на базе построения специальных модулей, которые интерпретируют каждый шаг изменения весов нейронной сети. Комплексный анализ существующих исследований дает возможность сделать выводы о целесообразности применения тех или иных подходов. Результаты анализа позволят получить подробную картину состояния исследований в данной области и создать собственные приложения на основе нейронных сетей, результаты работы которых можно будет подробно изучать и оценивать их достоверность. Разработка подобных систем крайне необходима для развития цифровой экономики в России и создания приложений, позволяющих принимать ответственные и объяснимые управленческие решения в критических областях народного хозяйства.

DOI: 10.31857/S0002338821060044

Введение. Масштабное развитие систем искусственного интеллекта (ИИ), в том числе приложений на основе искусственных нейронных сетей, открывает широчайшие возможности их использования в различных областях, от систем распознавания эмоций до систем предиктивной аналитики, применения в медицине и военной области. Но в то же время существующие системы и приложения имеют один общий существенный недостаток – невозможность интерпретации полученных результатов и принятых решений. Широко известная проблема так называемого черного ящика накладывает существенные ограничения для использования подобных систем, в том числе законодательных, так как нельзя проследить ход принятия решения нейронной сетью.

В настоящее время эти проблемы решаются в рамках направления объяснимого ИИ (explainable artificial intelligence или XAI). Системы на базе объяснимого ИИ помогают понять пользователю принятые с помощью методов машинного обучения решения, что повышает доверие к этим системам и дает возможности принимать более эффективные решения на основе результатов работы системы. Все это позволяет разработчикам и пользователям исследовать факторы,

[1] Работа выполнена при финансовой поддержке РФФИ (грант № 20-17-50199) по программе “Экспансия”.

которые используются нейронной сетью при решении конкретной задачи и понять, какие параметры нейронной сети нужно поменять, чтобы повысить точность ее работы.

Кроме этого, изучение того, как нейронные сети извлекают, хранят и преобразуют знания, может быть полезно для будущего развития методов ИИ. Например, повышение объяснимости искусственных нейронных сетей позволит обнаружить так называемые скрытые зависимости, которые не присутствуют во входных данных, но появляются в результате их интеграции в нейронную сеть. Методы извлечения правил из нейронных сетей являются одним из связующих элементов между символьными и коннекционисткими моделями представления знаний в ИИ.

1. История развития объяснимых моделей ИИ. Исследования в данной области можно разделить на три этапа: первый этап (с 1970 г.) – разработка экспертных систем, второй этап (середина 1980-х годов) – переход от экспертных систем к системам, основанным на знаниях, и третий (с 2010 г.) – изучение глубоких архитектур искусственных нейронных сетей, потребовавший новых глобальных исследований по построению объяснимых систем.

Рассмотрим далее подробнее каждый из этих этапов.

1.1. Первый этап. Экспертные системы. Системы объяснения, основанные на правилах. Первое и второе поколения объяснимого ИИ связаны с экспертными системами, включавшими в себя принятие решений и постановку диагнозов. Главный этап развития экспертных систем пришелся на начало 1970-х годов и содержал представления, основанные на знаниях, использовании правил и отношений. Системы имели инструментарий вопрос-ответных интерфейсов для пользователей и могли давать рекомендации и ставить диагнозы, основанные на правилах. Исследования показывали, что объяснение, как компьютерные системы работают, позволяет повысить уровень доверия людей к рекомендательным системам [1]. При интерпретации результатов работы экспертной системы также различались объяснение принципа работы системы и обоснование выбора архитектуры системы [2].

На первом этапе, в эпоху разработки экспертных объяснительных систем, велись масштабные исследования, демонстрирующие, что экспертные системы, имеющие объяснительную составляющую, способны оказывать большее влияние на принятие решения, причем эффект был прямо пропорционален уровню навыков пользователя [3–6].

Но многие экспертные системы первого поколения не принесли ожидаемых преимуществ. При работе над медицинскими экспертными системами исследователи признали, что врачи будут игнорировать рекомендации экспертной системы, если не будет предоставлено обоснование (оправдание) того, почему система дала такие рекомендации. Системы первоначального объяснения пытались предоставить это обоснование, описывая основные цели и шаги, используемые для постановки диагноза. Этот подход, который Свартаут и Мур назвали “Резюме как миф объяснения” [7], также не полностью устраивал пользователей, и это привело к переосмыслению подходов, целей и задач систем XAI. Первое поколение появилось в конце 1970-х годов [8] и просуществовало около 10 лет.

Некоторое несовершенство экспертных систем первого поколения породила в свою очередь первое поколение систем, включающее блок объяснения как обязательный элемент, например, Mycin [9] и связанные с ним системы Digitalis Therapy Advisor [10], BLAH [11] и другие системы объяснения специального назначения. Применяя деревья вывода, эти системы работали, создавая логические и вероятностные правила для постановки диагноза или ответов на вопросы. В целом, поскольку знания и опыт были сформулированы в терминах правил, эти правила описывались на естественном языке. Простое объяснение выражалось правилом для принятия решения. Такие объяснения обычно писались на ограниченном естественном языке, но часто они были простым использованием продукционных правил “если-то” для текстовых описаний.

В [12] рассмотрены системы первого поколения, которые создавали объяснения, перефразируя правила для принятия решения. В целом, экспертные системы и системы объяснения первого поколения были сосредоточены на описании внутренних состояний интеллектуальной системы, ее целей и планов. Иногда это было довольно просто, потому что сами правила являлись формализацией правил, используемых экспертами. Иногда, однако, языковые описания мало походили на человеческое объяснение или вообще на естественный язык. Во-первых, они полагались только на формат логических и причинно-следственных правил “если-то”, а не на предоставление объяснений на более высоком уровне стратегии рассуждений (например, сбор базовой информации о пациенте, создание сети для альтернативных объяснений, попытки поддержать конкретную гипотезу). Во-вторых, некоторые правила были логически необходимы для того, чтобы система работала, не обязательно имели смысл для пользователей и не имели отношения к стандартным вопросам пользователя “как” и “почему”. В-третьих, знания предметной

области иногда компилировались (например, причинно-следственные связи между симптомами и диагнозами), поэтому они не включались в объяснение.

Когда был изначально разработан Mycin, неспособность полностью объяснить набор правил и их обоснований не считалась недостатком, потому что создание каких-либо объяснений в удобочитаемой форме на базе записанного дерева рассуждений программы уже было сложной задачей и значительным достижением в области ИИ. Более того, такие ассоциативные модели представляли собой практические правила, основанные на демографических и физиологических данных, которые обычно были знакомы пользователям. Последние, как предполагалось, просто следовали советам программы после того, как она была протестирована и сертифицирована экспертами. Однако попытки расширить или уточнить эти наборы правил, использовать их для обучения или объяснить высокоуровневые ассоциации были неудачными.

1.2. Второй этап. Системы объяснения, основанные на знаниях. К середине 1980-х годов ограничения систем объяснения первого поколения столкнулись с проблемой, которая заключалась в том, что недостаточно просто резюмировать внутреннюю работу системы. Созданный текст может быть верным, но это не обязательно то, что хотел узнать пользователь, или ему было непонятным. Системы второго поколения часто стирали грань между консультационной программой [12], наставником [13], советником [14] и системой ввода данных [15], но перед исследователями стояла цель выйти за рамки методов, которые не помогли улучшить понимание или принятие экспертных систем. Основной движущей силой систем второго поколения была необходимость разработки более абстрактных структур, которые способствовали бы повторному использованию и простоте разработки систем. NEOMYCIN сделал это, предоставив диагностическую процедуру на основе задач и метаправил и связанный с ней язык таксономических и причинно-следственных связей [16, 17].

На раннем этапе разработки объяснительных экспертных систем было признано, что их база знаний, правила и объяснения могут быть использованы для создания интеллектуальных наставников. Отличительной особенностью интеллектуальных систем обучения является то, что они выводят ментальную модель предметной области каждого учащегося (его базу знаний) на основе поведения учащегося. Эти системы сначала собирательно назывались интеллектуальными системами компьютерной аналитики и рассуждений (computational analytics and intelligence, CAI), чтобы отличать интеллектуальных наставников от простых обучающих машин 1960-х годов, и применялись сообществом ИИ в образовании, особенно в 1980-х и начале 1990-х годов.

Например, в экспертной системе GUIDON запрос студента о данных пациента и его заявленные гипотезы использовались для обратного поиска через сеть правил MYCIN (или любую другую сеть), чтобы определить, какие правила не учитывались. Объяснение представляло собой сеть умозаключений и иногда включало неопределенности (например, свидетельство того, что студент знает правило, базирующееся на предыдущих взаимодействиях, нужно отличать от его применения в конкретном случае). Вот почему системы иногда называли “учителями, основанными на знаниях”, в отличие от обучающих машин 1960-х годов. Главная идея заключалась в том, что предметные знания экспертной системы (например, правила медицинской диагностики) были отделены от базы знаний обучения (например, правил управления диалогом). Кроме того, процесс интерпретации правил предметной области, аналогичный объяснению поведения экспертной системы на основе модели (трассировки) ее внутренних процессов, использовался для объяснения поведения студента путем построения модели того, как студент рассуждает.

1.3. Третий этап. Новые методы объяснения на основе распознавания образов и обработки текстов. После некоторого застоя, связанного с замедлением развития направления систем, основанных на знаниях, возникает третий этап, относящийся к 2017–2021 гг. и отраженный в разд. 4 настоящей статьи, который описывает программу DARPA по объяснимому ИИ.

Возникновение третьего этапа можно наглядно продемонстрировать, проанализировав количество публикаций по темам извлечения правил и объяснимого ИИ. При анализе количества публикации за последние 20 лет можно отметить, что интерес к извлечению правил из нейросетей более или менее стабилен (по некоторым источникам даже периодичен, что связано с тремя рассмотренными поколениями систем объяснимого ИИ) и возрастает только с появлением третьего поколения систем объяснимого ИИ (рис. 1), что связано,в основном с созданием программы DARPA.

С появлением программы DARPA можно наблюдать взрывной рост количества публикаций непосредственно по теме объяснимого ИИ (начиная с 2018 г.) и его дальнейший рост, как показано на рис. 2.

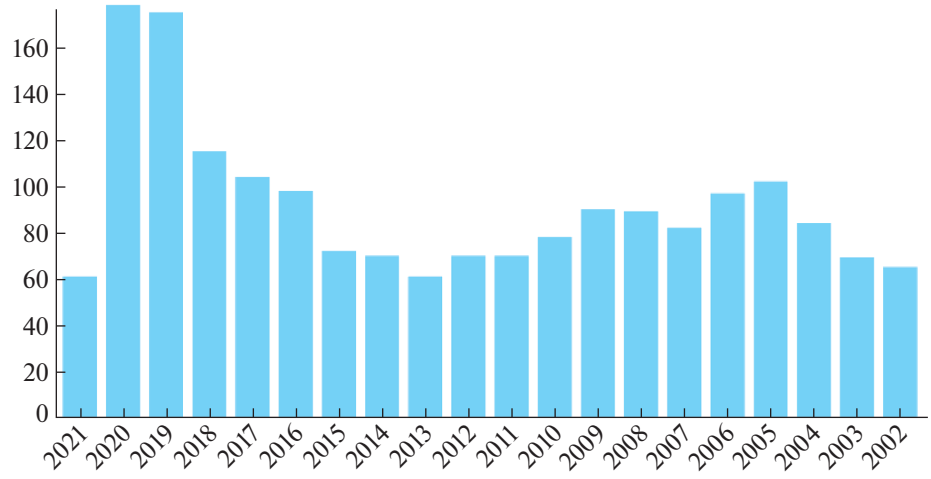

Рис. 1. Количество публикаций по годам в области извлечения правил из нейронных сетей

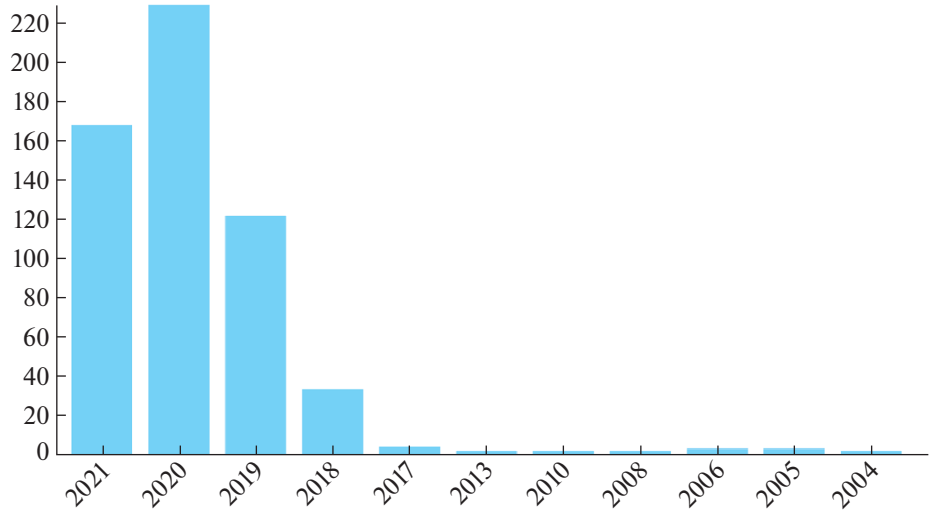

Рис. 2. Количество публикаций по годам по теме объяснимого ИИ. (XAI)

В системах третьего поколения, как и в системах первого поколения, осуществлялись попытки объяснить внутреннюю работу системы, что само по себе остается серьезной проблемой. Системы первого поколения встраивали экспертные знания в правила, часто получаемые непосредственно от экспертов, и пытались составить языковые описания на основе оценок экспертов. Эти правила часто трансформировались в выражения естественного языка, и большая часть этой работы была направлена на создание систем представления знаний. В системах третьего поколения эта задача оказалась значительно сложнее. Недостатки систем первого поколения, связанные со слабым уровнем детализации и непонятным языком, могут стать проблемой для систем третьего поколения. Со времени систем первого поколения компьютерные технологии в визуализации данных, анимации, видео и т.п. значительно продвинулись вперед, и много новых идей было предложено в качестве потенциальных методов для генерации объяснений. Хотя системы первого поколения поддерживали диалоги на естественном языке и интерактивность в вопросно-ответных системах, современные системы делают это на более высоком уровне, чем экспертные системы. Примером может служить использование аргументированных объяснительных диалогов человек-машина для устранения несоответствий в базах знаний в процессе приобретения знаний с помощью программного обеспечения для экспертных систем [18].

Некоторые из объяснений, представленных в системах первого поколения, было легко создать по сравнению с объяснениями в третьем поколении, поскольку они были прямым повторением закодированных вручную правил. Нынешнему поколению систем, возможно, будет труднее давать простые объяснения, производимые системами первого поколения. Но в некотором смысле системы третьего поколения, разрабатываемые в настоящее время, отражают достижения систем первого поколения. Ожидается, что многие из этих систем могут встретить проблемы, аналогичные тем, с которыми столкнулись системы первого поколения. И они могут быть решены методами, которые возникли в системах второго поколения.

2. Обзор исследований по извлечению правил из искусственных нейронных сетей. Повышение прозрачности нейронных сетей путем извлечения из них правил имеет два основных преимущества. Это дает пользователю некоторое представление о том, как нейронная сеть использует входные переменные, чтобы принять решение, и позволяет выявить скрытые функции в нейросетях, когда правила применяются для объяснения работы отдельных нейронов. Выявление особо важных атрибутов или причин ошибок нейронной сети может быть частью понимания. Пытаясь сделать непрозрачные нейронные сети более понятными, методы извлечения правил устраняют разрыв между точностью и ясностью [19, 20].

Для того чтобы, например, нейронная сеть использовалась в критически важных приложениях, к примеру в авиации или электроэнергетике, требуется объяснение не только принципов ее работы, но и того, каким образом нейронная сеть получила результат. В этих случаях крайне важно, чтобы у пользователя системы была возможность проверить выход искусственной нейронной сети при всех возможных условиях входа [21].

Для формализации задачи извлечения правил из нейронной сети можно использовать следующее определение: “При заданных параметрах обученной нейронной сети и данных, на которых она была обучена, создайте описание нейросетевой гипотезы, которая понятна, но приближена к поведению заданной сети”.

Чтобы различать разные подходы к извлечению правил из нейронных сетей, в [21] была введена многомерная таксономия. Первое измерение, которое в ней описывается, является мощностью извлекаемых правил (например, IF-THEN правила или нечеткие продукционные правила). Второе измерение называется прозрачностью и описывает стратегию, на основе которой работает алгоритм извлечения правил. Если метод использует нейронную сеть только как черный ящик, независимо от архитектуры нейросети, мы называем его педагогическим подходом. Если вместо этого алгоритм учитывает внутреннюю структуру нейронной сети, мы называем этот подход декомпозиционным. Если алгоритм применяет компоненты как педагогических, так и декомпозиционных методов, то этот подход называется эклектическим. Третьим измерением является качество извлеченных правил. Поскольку качество – широкий термин, оно делится на несколько критериев: качество, точность, непротиворечивость и понятность. В то время как качество измеряет способность правильно классифицировать ранее не видимые примеры, точность измеряет степень, с которой правила имитируют поведение нейронной сети [19]. Точность может рассматриваться как точность по отношению к выходу нейронной сети. Непротиворечивость может быть измерена только тогда, когда алгоритм извлечения правил включает процесс обучения нейронной сети вместо обработки уже обученной нейронной сети. Извлеченный набор правил считается непротиворечивым, когда нейронная сеть генерирует наборы правил, которые правильно классифицируют тестовые данные для различных эпох обучения. Понятность рассматривается как мера размера правил, т.е., короткие правила при их небольшом количестве считаются более понятными.

В данной работе остановимся только на трех описанных критериях. В соответствии с [22] ориентируемся на методы, которые не предъявляют особых требований к тому, как была обучена нейронная сеть до того, как были извлечены правила. Кроме того, анализируются только алгоритмы, способные извлекать правила из нейронных сетей прямого распространения, несмотря на любые другие характеристики архитектуры. В соответствии [23] необходимо, чтобы алгоритм обладал наибольшей степенью общности.

Проанализируем некоторые методы извлечения правил, которые отвечают вышеуказанным характеристикам. Начнем с декомпозиционного подхода. Как упоминалось ранее, декомпозиционные подходы для извлечения правил из нейронных сетей действуют на уровне нейронов. Как правило, декомпозиционный подход анализирует каждый нейрон, и формирует правила, которые имитируют поведение этого нейрона. Рассмотрим ниже послойный алгоритм извлечения правил KT, полиномиальный алгоритм Цукимото и экстрактор правил через индукцию дерева решений.

Алгоритм KT был одним из первых подходов для извлечения правил из нейронных сетей [24]. Он описывает каждый нейрон (слой за слоем) с правилами IF-THEN путем эвристического поиска комбинаций входных атрибутов, превышающих порог нейрона. Модуль записи используется для получения правил, которые относятся к исходным атрибутам ввода, а не к выводам предыдущего уровня. Чтобы найти подходящие комбинации, метод KT применяет поиск на дереве, т.е. правило (представленное в качестве узла в дереве) на этом уровне генерирует свои начальные узлы, добавляя дополнительный доступный атрибут. Кроме того, алгоритм использует ряд эвристик, чтобы остановить рост дерева в ситуациях, когда дальнейшее улучшение невозможно.

Полиномиальный алгоритм Цукимото для извлечения правил из нейронной сети очень похож на метод KT. Он использует многоуровневый декомпозиционный алгоритм для извлечения правил IF-THEN для каждого нейрона, а также отслеживает стратегию поиска входных конфигураций, превышающих порог нейрона. Основное преимущество алгоритма Цукимото – его вычислительная сложность, которая является полиномиальной, в то время как метод KT – экспоненциальный. Алгоритм достигает полиномиальной сложности, находя соответствующие термины, используя пространство многомерных функций. На втором этапе эти термины применяются для создания правил IF-THEN. Впоследствии обучающие данные используются для повышения точности правил. На последнем этапе алгоритм Цукимото пытается оптимизировать понятность, удаляя из правил несущественные атрибуты.

Другой метод извлечения правил путем индукции дерева решений был введен в [25]. Их алгоритм CRED преобразует каждую выходную единицу нейронной сети в решение, где узлы деревьев тестируются с помощью узлов скрытого слоя, а листья представляют класс. После этого шага извлекаются промежуточные правила. Затем для каждой точки разветвления, используемой в этих правилах, создается другое дерево решений с помощью точки разветвления на входном слое нейронной сети. В новых деревьях листья не выбирают непосредственно класс. Извлечение правил из второго дерева решения приводит нас к описанию состояния скрытых нейронов, состоящих из входных переменных. В качестве заключительного шага заменяются промежуточные правила, описывающие выходной слой через скрытый слой на те, которые описывают скрытый слой на основе входов нейронной сети. Затем они объединяются, чтобы построить правила, описывающие выход нейронной сети на базе ее входных данных.

Педагогические подходы не учитывают внутреннюю структуру нейронной сети. Целью педагогических подходов является рассмотрение обученных нейросетей как целостного объекта или как черного ящика [26]. Главная идея состоит в том, чтобы извлечь правила, непосредственно сопоставляя входные данные с выходными данными [27].

Педагогические подходы работают с нейронной сетью, как с функцией. Эта функция задает выход нейронной сети для произвольного входа, но не дает понимания внутренней структуры нейронной сети или ее весов. Для нейронной сети этот класс алгоритмов пытается найти связь между возможными вариациями входа и выходами, созданными нейронной сетью, причем некоторые из них используют заранее определенные обучающие данные, а некоторые нет.

Извлечение правил, основанное на интервальном анализе, применяет анализ доверительных интервалов (VIA) для извлечения правил, имитирующих поведение нейронных сетей. Главная идея этого метода заключается в том, чтобы найти входные интервалы, в которых выходной сигнал нейронной сети стабилен, т.е. прогнозируемый класс одинаков для незначительно меняющихся конфигураций входа. В результате чего анализ доверительных интервалов обеспечивает основу для надежных корректных правил.

Получение правил с использованием выборки представляет собой несколько методов, которые придерживаются более или менее той же стратегии для извлечения правил из нейронной сети с помощью выборки, т.е. они создают обширный набор данных в качестве основы для извлечения правил. После этого выбранный набор данных передается в стандартный алгоритм обучения для генерации правил, имитирующих поведение сети. В [28] доказано, что применение выборочных данных эффективнее, чем использование только обучающих данных в проблемах извлечения правил.

Одним из первых методов, последовавших за этой стратегией, был алгоритм Трепана [29]. Он работает аналогично алгоритму “разделяй и властвуй”, выполняя поиск точек разветвления на обучающих данных для отдельных экземпляров разных классов. Основными отличиями от метода “разделяй и властвуй” являются лучшая стратегия расширения древовидной структуры, дополнительные точки разветвления и возможность выбора дополнительных примеров обучения в более глубоких узлах дерева. В результате алгоритм также создает дерево решений, которое, однако, может быть преобразовано в набор правил, если это необходимо.

Алгоритм извлечения правил двоичного входа и выхода (BIO-RE) способен обрабатывать только нейронную сеть с двоичными или бинаризированными входными атрибутами. BIO-RE создает все возможные комбинации входных данных и запрашивает их у нейронной сети. С помощью вывода нейронной сети для каждого примера создается таблица истинности. От таблицы истинности также легко перейти к правилам, если это необходимо.

ANN-DT является еще одним методом выборки, основанным на принятии решений для описания поведения нейронной сети. Общий алгоритм базируется на алгоритме CART с некоторыми вариациями в первоначальной реализации. ANN-DT использует метод выборки для расширения, с тем чтобы большая часть обучающей выборки была репрезентативной. Это достигается с помощью метода ближайшего соседа, в котором рассчитывается расстояние от точки выборки до ближайшей точки в наборе обучающих данных и сравнивается с эталонным значением.

Идея создания большого набора примеров на первом этапе также реализуется алгоритмом STARE. Как и BIO-RE, STARE формирует большие таблицы истинности для обучения. Преимущество STARE заключается в его способности не только обрабатывать двоичные и дискретные атрибуты, но и работать с непрерывными входными данными. Для формирования таблиц истинности алгоритм перестраивает входные данные. Примером педагогического подхода с помощью обучающей выборки является KDRuleEx. Аналогично алгоритму Trepan, этот алгоритм также генерирует дополнительные обучающие примеры в случае, когда данных для следующих точек разделения слишком мало. KDRuleEx использует генетический алгоритм для создания новых примеров обучения. Эта техника приводит к таблице принятия решений, которая может быть преобразована, например, в правила IF-THEN.

Эклектический подход – это набор методов извлечения правил, включающих элементы как педагогического, так и декомпозионного подхода. В частности, эклектический подход использует знания о внутренней архитектуре и вектора весов в нейронной сети в дополнение к символьному алгоритму обучения.

Подход быстрого поиска правил в нейронной сети включает в себя подход FERNN, который сначала пытается определить соответствующие скрытые нейроны, а также входы в сеть. Для этого шага строится дерево решений с помощью алгоритма C4.5. Процесс извлечения правил приводит к генерации правил M-of-N и IF-THEN. Имея набор правильно классифицированных примеров обучения, FERNN анализирует значения активации каждой скрытой вершины, для которой значения активации сортируются в порядке возрастания. Затем используется алгоритм C4.5, чтобы найти наилучшую точку разделения для формирования дерева решений. Ниже в таблице приводится сравнение алгоритмов извлечения правил из нейронных сетей по типу нейросети, типу алгоритма и виду извлекаемого правила [30].

Таблица. Алгоритмы извлечения правил из нейронных сетей

Алгоритм	Используемый тип сети	Тип алгоритма	Вид извлекаемого правила
DIFACON-miner	Многослойный персептрон	Декомпозиционный	IF-THEN
CRED	То же	>>	Дерево решений
FERNN	>>	>>	M-of-N, IF-THEN
KT	>>	>>	IF-THEN
Tsukamoto's algorithm	Многослойный персептрон, рекуррентная нейронная сеть	>>	IF-THEN
TREPAN	Многослойный персептрон	Педагогический	M-of-N, дерево решений
HYPINV	>>	>>	Правило гиперплоскости
BIO-RE	>>	>>	Бинарное правило
KDRuleEX	>>	>>	Дерево решений
RxREN	>>	>>	IF-THEN
ANN-DT	>>	>>	Бинарное дерево решений
RX	>>	Эклектический	IF-THEN
Kahramanli and Allahverdi's algorithm	>>	>>	IF-THEN
DeepRED	Глубокая нейронная сеть	Декомпозиционный	IF-THEN

3. Нейронечеткие модели в задачах извлечения правил из искусственных нейронных сетей. Наиболее интересным в рамках данного исследования является извлечение правил с использованием нейронечетких моделей. Системы, основанные на нечетких правилах (FRBS), разработанные с помощью нечеткой логики, стали полем активных исследований за последние несколько лет. Эти алгоритмы доказали свои сильные стороны в таких задачах, как управление сложными системами, создание нечетких элементов управления. Взаимоотношения между обоими подходами (ANN и FRBS) были тщательно изучены и показана их эквивалентность. Это позволяет сделать два важных вывода. Во-первых, можно применить то, что было обнаружено для одной из моделей, к другой. Во-вторых, мы можем перевести знания, встроенные в нейронную сеть, на более когнитивно-приемлемый язык – нечеткие правила. Другими словами, получаем семантическую интерпретацию нейронных сетей [31–33].

Для того, чтобы получить семантическую интерпретацию черного ящика глубокого обучения, нейронечеткие сети могут быть использованы вместо последнего полносвязного слоя. Например, ANFIS (адаптивная нейронечеткая система) является многослойной сетью прямого распространения. Эта архитектура имеет пять слоев, таких как нечеткий слой, продукционный слой, слой нормализации, слой дефаззификации и выходной слой. ANFIS сочетает преимущества нейросети и нечеткой логики. Далее приведем классификацию наиболее известных нейронечетких подходов.

Рассматривая архитектуры нейронечетких моделей, можно выделить три методики объединения искусственных нейронных сетей (ИНС) и нечетких моделей [34, 35]:

neuro-FIS, в которых ИНС используется как инструмент в нечетких моделях;

нечеткие ИНС, в которых классические модели ИНС фаззифицированы;

нейронечеткие гибридные системы, в которых нечеткие системы и ИНС объединены в гибридные системы [36, 37].

Исходя из данных методик, нейронечеткие модели можно разделить на три класса [38–40].

Кооперативные нейронечеткие модели. В данном случае часть ИНС изначально используется для определения нечетких множеств и / или нечетких правил, где впоследствии выполняется только полученная нечеткая система. В процессе обучения определяются функции принадлежности, а также формируются нечеткие правила на основе обучающей выборки. Здесь основная задача нейронной сети заключается в подборе параметров нечеткой системы.

Параллельные нейронечеткие модели. Нейронная сеть в данном типе модели работает параллельно с нечеткой системой, предоставляя входные данные в нечеткую систему или изменяя выходные данные нечеткой системы. Нейронная сеть может являться также и постпроцессором выходных данных из нечеткой системы.

Гибридные нейронечеткие модели. Нечеткая система использует метод обучения, как это делает и ИНС, чтобы настроить свои параметры на основе обучающих данных. Среди представленных классов моделей наибольшей популярностью пользуются модели именно данного класса, доказательством тому служит их применение в широком спектре реальных задач [41–44].

Среди наиболее популярных гибридных моделей можно выделить следующие архитектуры.

Сеть управления нечетким адаптивным обучением (FALCON) [45], которая имеет пятислойную архитектуру. На одну выходную переменную приходится по два лингвистических узла. Первый узел работает с обучающей выборкой (паттерном обучения), второй является входным для всей системы. Первый скрытый слой размечает входную выборку в соответствии с функциями принадлежности. Второй слой задает правила и их параметры. Обучение происходит на основе гибридного алгоритма без учителя для определения функции принадлежности, базы правил и использует алгоритм градиентного спуска для оптимизации и подбора итоговых параметров функции принадлежности.

Адаптивная нейронечеткая система вывода ANFIS [46] — это хорошо известная нейронечеткая модель, которая применялась во многих приложениях и исследовательских областях [47]. Более того, сравнение архитектур нейронечетких сетей показало, что ANFIS показывает минимальную ошибку в задаче прогнозирования. Основным недостатком модели ANFIS является то, что она предъявляет серьезные требования к вычислительной мощности [48].

Система обобщенного приближенного интеллектуального управления на основе рассуждений (GARIC) [49] представляет собой нейронечеткую систему, использующую два нейросетевых модуля, модуль выбора действия и модуль оценки состояния, который отвечает за оценку качества выбора действий предыдущим модулем. GARIC – пятислойная сеть прямого распространения.

Нейронный нечеткий регулятор (NEFCON) [50] был разработан для реализации системы нечеткого вывода типа Мамдани. Связи определяются с помощью нечетких правил. Входной слой является фаззификатором, а выходной решает задачу дефаззификации. Обучается сеть на основе гибридного алгоритма обучения с подкреплением и алгоритма обратного распространения ошибки.

Система нечеткого вывода и нейронной сети в программном обеспечении нечеткого вывода (FINEST) [51] представляет собой систему настройки параметров. Производится настройка нечетких предикатов, функции импликации и комбинаторной функции.

Система для автоматического построения нейронной сети нечеткого вывода (SONFIN) [52] по своей сути аналогична NEFCON контроллеру, но вместо реализации нечеткого вывода типа Мамдани реализует вывод типа Такаги-Сугено. В данной сети входная выборка обрабатывается с помощью алгоритма выровненной кластеризации. При идентификации структуры части предварительного условия входное пространство разделяется гибким образом в соответствии с алгоритмом, основанным на выровненной кластеризации. Настройка параметров системы частично реализована на базе метода наименьших квадратов, предварительные условия настраиваются с помощью метода обратного распространения ошибки.

Динамически развивающаяся нечеткая нейронная сеть (dmEfuNN) и (EFuNN) [53]. В EFuNN все узлы создаются в процессе обучения. Первый слой передает обучающие данные на второй, который вычисляет степень соответствия с заранее определенной функцией принадлежности. Третий слой содержит в себе наборы нечетких правил, являющихся прототипами входных-выходных данных, которые можно представить в качестве гиперсфер нечеткого входного и выходного пространств. Четвертый слой рассчитывает степень, с которой выходная функция принадлежности разметила входные данные, а пятый слой производит дефаззификацию и подсчитывает числовые значения выходной переменной. DmEfuNN представляет собой модифицированную версию EFuNN. Основная идея состоит в том, что для всех входных векторов динамически подбирается набор правил, значения активации которых используются для расчета динамических параметров выходной функции. В то время как EFuNN реализует нечеткие правила типа Мамдани, dmEFuNN применяет тип Такаги-Сугено.

4. Обзор подходов к разработке систем объяснимого ИИ. Проведем исследование различных моделей объяснимого ИИ. Практически все они связаны с системами объяснений третьего поколения и начавшейся в 2018 г. программой DARPA [54]. Программа DARPA по созданию систем объяснимого ИИ (XAI) стремится создать такие системы ИИ, чьи модели обучения и решения могут быть понятны и должным образом проверены конечными пользователями. Достижение этой цели требует методов построения более объяснимых моделей, разработки эффективных объяснимых интерфейсов и понимания психологических требований для эффективного объяснения. Объяснимый ИИ нужен для того, чтобы пользователи понимали, должным образом доверяли и эффективно управляли своими умными партнерами. DARPA рассматривает XAI как системы ИИ, которые могут объяснить свое решение человеку-пользователю, охарактеризовать свои сильные и слабые стороны, и то, как они будут вести себя в будущем. Целью DARPA является создание более понятных для человека систем ИИ с помощью эффективных объяснений. Команды разработчиков XAI решают первые две проблемы путем создания и развития технологий объяснимого машинного обучения (ML), разрабатывая принципы, стратегии и методы взаимодействия человека и компьютера для получения эффективных объяснений. Еще одна команда разработчиков XAI решает третью задачу путем объединения, расширения и применения психологических теорий объяснения, которые команды разработчиков будут использовать для тестирования своих систем. Команды разработчиков оценивают, насколько хорошо объяснения XAI-систем улучшают работу пользователей, их доверие и производительность.

В России также уделяется большое внимание направлению объяснимого ИИ. Так, Нижегородский государственный университет в 2020 г. стал победителем в конкурсе крупных научных проектов от Минобрнауки РФ с проектом “Надежный и логически прозрачный искусственный интеллект: технология, верификация и применение при социально-значимых и инфекционных заболеваниях” [55]. Главным результатом проекта должна стать разработка новых методов и технологий, позволяющих преодолеть два основных барьера систем машинного обучения и ИИ: проблему ошибок и проблему явного объяснения решений. Руководитель проекта профессор Александр Горбань так объяснил основную идею проекта: “Эти проблемы тесно связаны: без возможности логического прочтения ошибки ИИ будут оставаться необъяснимыми. Дообучение системы в рамках существующих методов может повредить имеющиеся навыки и, с другой стороны, может потребовать огромных ресурсов, что в серьезных задачах непрактично. К при-

меру, широко известная система когнитивных вычислений IBM Watson потерпела неудачу на рынке персонализированной медицины вследствие систематически совершаемых ошибок в диагностике и рекомендации лечения рака, найти и устранить источники которых не удалось".

Далее приведено краткое описание моделей объяснимого ИИ и научных центров, которые занимаются данными исследованиями в рамках программы DARPA по объяснимому ИИ [54].

1. Глубокий объяснимый ИИ (DEXAI) в UCB. Команда Калифорнийского университета в Беркли (UCB) (включая исследователей из Бостонского университета, Амстердамского университета и Kitware) разрабатывает систему ИИ, понятную человеку благодаря явной структурной интерпретации и интроспективному объяснению, которая обладает предсказуемым поведением и обеспечивает соответствующую степень доверия [56]. Ключевые проблемы глубокого объяснимого ИИ (DEXAI) состоят в том, чтобы генерировать точные объяснения поведения модели и выбирать те, которые наиболее полезны для пользователя. UCB обращается к первой проблеме, создавая неявные или явные модели объяснения: они могут неявно представлять сложные скрытые представления понятным образом или строить явные структуры, которые по своей сути понятны. Эти модели DEXAI создают набор возможных объяснительных действий. Для второй проблемы UCB предлагает рациональные объяснения, которые используют модель убеждений пользователя при принятии решения, какие объяснительные действия выбрать. UCB также создает интерфейс объяснения, основанный на этих нововведениях и на принципах интерактивной разработки. Автономные модели DEXAI применяются для управления транспортными средствами (с помощью набора данных Berkeley Deep Drive и симулятора CARLA) [57], а также в сценариях стратегической игры (StarCraft II). Для данных аналитики DEXAI использует визуальные ответы на вопросы (VQA) и методы фильтрации (например, с помощью больших наборов данных, таких, как VQA-X и ACT-X для задач VQA и задач распознавания активности), xView и Distinct [58].

2. Причинно-следственные модели для объяснения машинного обучения (CRA). Цель команды Charles River Analytics (CRA) (включая исследователей из Массачусетского университета и Университета Брауна) – создать и предоставить каузальные объяснения работы машинного обучения с помощью причинно-следственных моделей (CAMEL). Объяснения CAMEL представлены пользователю в виде рассказов в интерактивном, интуитивно понятном интерфейсе. CAMEL включает в себя структуру каузального вероятностного программирования, которая объединяет представления и методы обучения из каузального моделирования [59] с вероятностными языками программирования [60]. Генеративные вероятностные модели, представленные на языке вероятностного программирования, естественным образом выражают причинно-следственные связи; они хорошо подходят для задачи по объяснению систем машинного обучения. CAMEL исследует внутреннее представление системы машинного обучения, чтобы выявить, как оно представляет определенные пользователем концепции естественной области. Затем он строит причинно-следственную модель их влияния на работу системы машинного обучения, проводя эксперименты, в которых области согласования систематически включаются или удаляются. После изучения он использует причинно-вероятностные модели для вывода объяснений предсказаний или действий системы. В области анализа данных CAMEL решает задачу обнаружения пешеходов (с помощью набора данных о пешеходах INRIA) [61], а CRA работает над задачами распознавания активности (с использованием ActivityNet). Свойство автономности CAMEL демонстрируется в игре Atari Amidar, и CRA работает в StarCraft II.

3. Изучение и передача объяснимых представлений для аналитики и автономности. Команда Калифорнийского университета в Лос-Анджелесе (UCLA) (совместно с исследователями из Университета штата Орегон и Университета штата Мичиган) разрабатывает интерпретируемые модели, сочетающие репрезентативные парадигмы, включая интерпретируемые глубокие нейронные сети, композиционные графические модели вида И/ИЛИ графики и модели, которые производят объяснения на трех уровнях (композиционности, причинности и полезности). Система UCLA содержит модуль исполнения, который выполняет задачи с мультимодальными входными данными, и модуль объяснения, который объясняет пользователю свое восприятие, когнитивные рассуждения и решения. Модуль исполнения выводит интерпретируемые представления в виде графа пространственного, временного и причинного анализа (STC-PG) для трехмерного восприятия сцены (для аналитики) и планирования задач (для автономности). STC-PG являются композиционными, вероятностными, интерпретируемыми и основанными на принципах глубоких нейронных сетей и используются для анализа изображений и видео. Модуль объяснения выводит поясняющий синтаксический граф в виде диалога [62], локализует соответствующий подграф в STC-PG и определяет намерения пользователя. UCLA охватывает обе

проблемные области XAI, применяя общую структуру представления и вывода. В области анализа данных UCLA продемонстрировал свою систему с помощью сети видеокамер для понимания сцены и анализа событий. Автономность UCLA показана в сценариях с использованием роботов, выполняющих задачи на платформах виртуальной реальности с реалистичной физикой, и в игре с вождением автономного транспортного средства.

4. Тестирование глубоких адаптивных программ с обоснованной информацией. Университет штата Орегон (OSU) разрабатывает инструменты для объяснения действий обученных агентов, которые выполняют последовательное принятие решений и определяют лучшие принципы разработки пользовательских интерфейсов с объяснениями. Модель объяснимого агента OSU использует объяснимую глубокую адаптивную программу (xDAP), сочетающую в себе адаптивные программы, глубокое обучение с подкреплением (RL) и объяснимость. С помощью xDAP программисты могут создавать агентов, представляющих решения, которые автоматически оптимизируются посредством глубокого RL при взаимодействии с симулятором. Для каждой точки выбора глубокое RL подключает обученную глубокую нейронную сеть для принятия решений (dNN), которая может обеспечить высокую производительность, но по своей сути не является объяснимой. После начального обучения xDAP программа xACT обучает объяснительную нейронную сеть [63] для каждой dNN. Они предоставляют разреженный набор функций объяснения (x-функций), которые кодируют свойства логики принятия решений dNN. Такие x-функции, которые являются нейронными сетями, изначально не интерпретируемыми человеком. Чтобы решить эту проблему xACT позволяет экспертам в предметной области прикреплять интерпретируемые описания к x-функциям, а программистам xDAP – аннотировать типы вознаграждений среды и другие концепции, которые автоматически встраиваются в dNN как “концепции аннотаций” во время обучения.

Пользовательский интерфейс объяснения OSU позволяет пользователям ориентироваться в тысячах решений агента и получить объяснения визуально и на естественном языке. Его конструкция основана на теории сбора информации, которая дает возможность пользователю в любой момент эффективно перейти к наиболее полезной пояснительной информации. OSU занимается проблемой автономности и продемонстрировала xACT в сценариях с использованием специально созданного игрового движка для стратегии в реальном времени. Пилотные исследования предоставили информацию для объяснения дизайна пользовательского интерфейса, охарактеризовав то, как пользователи ориентируются в игре с ИИ-агентом и стремятся объяснить игровые решения [64].

5. Общее обучение и объяснение. Команда Исследовательского центра Пало-Альто (PARC) (включая исследователей из Университета Карнеги-Меллона, Армейского кибер-института, Эдинбургского университета и Университета Мичигана) разрабатывает интерактивную систему объяснений, которая может объяснить возможности системы XAI, управляющей смоделированной беспилотной воздушной системой. Объяснения системы XAI должны сообщать, какую информацию она использует для принятия решений, понимает ли она, как все работает. Чтобы решить эту проблему, система общего изучения и объяснения PARC (COGLE) и ее пользователи устанавливают общую основу для определения того, какие термины применять в объяснениях и их значения. Это обеспечивается интроспективной моделью дискурса PARC, которая чередует процессы обучения и объяснения.

Многоуровневая архитектура COGLE разделяет обработку информации на осмысление, когнитивное моделирование и обучение. Уровень обучения использует повторяющиеся и иерархические глубокие нейронные сети с ограниченной пропускной способностью для создания абстракций и композиций по состояниям и действиям беспилотных воздушных систем для поддержки понимания обобщенных закономерностей.

Интерфейсы пояснений COGLE поддерживают анализ производительности, оценку рисков и обучение. Первый представляет собой карту, которая отслеживает действия беспилотных воздушных систем и разделяет путь действия или решения (полета) на объяснимые сегменты. Инструменты второго интерфейса позволяют пользователям изучать и оценивать компетенции системы и делать прогнозы относительно эффективности миссии. COGLE демонстрируется на программном симуляторе ArduPilot Software-in-the-Loop Simulator и на испытательном стенде дискретного абстрактного моделирования. Его качество оценивают операторы дронов и аналитики. Оценка на основе компетенций поможет PARC определить, как лучше всего разработать подходящие модели, понятные для предметной области.

6. Объяснимое обучение с подкреплением (RL) в Университете Карнеги-Меллон. Университет Карнеги-Меллон создает новую дисциплину объяснимого RL, чтобы обеспечить динамиче-

ское взаимодействие человека с машиной и адаптацию для максимальной производительности команды. Ученые преследуют две цели: разработать новые методы изучения объяснимых алгоритмов RL и создавать стратегии, которые могут объяснить существующие проблемы черного ящика. Для достижения первой цели Карнеги-Меллон разрабатывает методы улучшения обучения моделей для агентов RL, чтобы использовать преимущества подходов, основанных на моделях (способность визуализировать планы во внутреннем пространстве модели), в то же время объединяя их с преимуществами подходов без моделей (простотой и максимальной производительностью). К ним относятся методы, которые постепенно добавляют состояния и действия к моделям мира после обнаружения соответствующей скрытой информации, изучают модели посредством сквозного обучения комплексными алгоритмам оптимального управления, исследуют общие модели DL, которые используют физику твердого тела [65], и изучают предсказания состояний с помощью повторяющихся архитектур [66].

Университет Карнеги-Меллон также разрабатывает методы, которые могут объяснить действия и планы агентов RL черного ящика. Методы включают в себя ответы на такие вопросы, как, например, “Почему агент выбрал определенное действие”, или “Какие данные обучения больше всего повлияли на этот выбор”. Для этого университет разработал методы, которые генерируют описания агентов из журналов поведения и обнаруживают выбросы или аномалии. Университет Карнеги-Меллон занимается проблемой автономности и продемонстрировал объяснимый RL в нескольких сценариях, включая OpenAI Gym, игры Atari, моделирование автономных транспортных средств, мобильных сервисных роботов.

7. Объяснимые генеративные состязательные сети. Команда SRI International (включая исследователей из Университета Торонто, Университета Гвельфа и Калифорнийского университета в Сан-Диего) разрабатывает объяснимую структуру машинного обучения для анализа мультимодальных данных, которая генерирует понятные объяснения с обоснованием решений, сопровождаемых визуализацией входных данных, используемых для создания выводов. Система представлений на основе глубокого внимания для объяснимых генеративных состязательных сетей (DARE/X-GANS) применяет архитектуры DNN, аналогичные моделям внимания в визуальной нейробиологии. Она идентифицирует, извлекает и представляет доказательства пользователю как часть объяснения. Механизмы внимания предоставляют пользователю средства для исследования системы и совместной работы. DARE/X-GANS использует генеративные состязательные сети (GAN), которые учатся понимать данные, создавая их, одновременно изучая представления с объяснительными возможностями. Сети GAN становятся объяснимыми с помощью интерпретируемых декодеров. Это включает в себя создание визуальных доказательств по заданным текстовым запросам с помощью генерации текста по частям [67], причем части являются интерпретируемыми функциями, такими, как человеческие позы или ограничивающие рамки. Это свидетельство затем применяется для поиска запрошенных визуальных данных.

8. Система ответов на объяснимые вопросы. Команда Raytheon BBN Technologies (включая исследователей из Технологического института Джорджии, Массачусетского технологического института и Техасского университета в Остине) разрабатывает систему, которая отвечает на любые вопросы на естественном языке (NL), задаваемые пользователями о мультимедийных данных, и обеспечивает интерактивные возможные объяснения того, почему он получил такой ответ. Система объяснимых ответов на вопросы (EQUAS) изучает объяснимые модели DNN, в которых внутренние структуры (например, отдельные нейроны) согласованы с семантическими концепциями (например, колеса и руль) [68]. Это позволяет преобразовывать нейронные активации в сети в процессе принятия решения в объяснения NL (например, “этот объект является велосипедом, потому что у него два колеса и руль”). EQUAS также использует методы нейронной визуализации, чтобы выделить входные области, связанные с нейронами, которые больше всего повлияли на его решения. Чтобы выразить объяснения на основе случаев, EQUAS сохраняет индексы и извлекает случаи из своих обучающих данных, которые поддерживают его выбор. Отклоненные альтернативы распознаются и исключаются с помощью контрастного языка, визуализации и примеров. Четыре способа объяснения соответствуют ключевым элементам построения аргументов и интерактивного обучения: дидактическим утверждениям, визуализациям, случаям и отказу от альтернативных вариантов.

9. Управляемые вероятностные логические модели. Команда Техасского университета в Далласе (UTD) (включая исследователей из Калифорнийского университета в Лос-Анджелесе, Texas A&M и Индийского технологического института в Дели) разрабатывает единый подход к XAI с помощью управляемых вероятностных логических моделей (TPLM). TPLM – это семейство представлений, которое включает в себя деревья решений, диаграммы двоичных решений, сети

сечений, диаграммы сентенциальных решений, арифметические схемы первого порядка и управляемую логику Маркова [69]. Система UTD расширяет TPLM для генерации объяснений результатов запроса. Для масштабируемого вывода система применяет новые алгоритмы для ответа на сложные запросы с объяснением, используя такие методы, как обобщенный вывод, вариационный вывод и их комбинации.

10. Техасский университет A&M (TAMU). Команда Техасского университета A&M (TAMU) (включая исследователей из Вашингтонского государственного университета) разрабатывает интерпретируемую структуру DL, которая использует имитационное обучение для применения объяснимых неглубоких моделей и облегчает интерпретацию предметной области с визуализацией и взаимодействием. Интерпретируемые алгоритмы обучения системы извлекают знания из DNN для соответствующих объяснений. Его модуль DL подключается к модулю генерации шаблонов, используя интерпретируемость неглубоких моделей. Результаты обучения отображаются пользователям с визуализацией, включая скоординированные и интегрированные представления. Система TAMU обрабатывает данные изображений [70] и текст [71] и применяется в проблемной области аналитики XAI. Он обеспечивает эффективную интерпретацию обнаруженных неточностей из различных источников, сохраняя при этом конкурентоспособные характеристики обнаружения. Система TAMU сочетает интерпретируемость на уровне модели и на уровне экземпляра для генерации объяснений, которые легче понять пользователям. Эта система была развернута для решения множества задач с использованием данных из Twitter, Facebook, ImageNet, CIFAR-10, онлайн-форумов по вопросам здравоохранения и новостных веб-сайтов.

11. Объяснение модели с помощью оптимального выбора обучающих примеров (Rugers University). Университет Руджерс расширяет возможности байесовского обучения, чтобы сделать возможным автоматическое объяснение, выбирая подмножество данных, которое является наиболее репрезентативным для вывода модели. Этот подход также позволяет объяснить выводы любой вероятностной генеративной и дискриминативной модели, а также моделей глубокого обучения [72]. Университет Руджерс также разрабатывает формальную теорию взаимодействия человека и машины и поддерживает интерактивное объяснение сложных композиционных моделей. Распространенным среди них является подход, основанный на моделях человеческого обучения, которые способствуют объяснимости, и тщательно контролируемых поведенческих экспериментах для количественной оценки объяснимости. Объяснение с помощью байесовского обучения вводит набор данных, вероятностную модель и метод вывода и возвращает небольшое подмножество примеров, которые лучше всего объясняют вывод модели. Было продемонстрировано, что данный подход облегчает понимание больших корпусов текстов, что оценивается способностью человека точно составить резюме корпуса после коротких, управляемых объяснений. Университет Руджерс занимается проблемной областью анализа данных и продемонстрировал свой подход на изображениях, тексте, их комбинациях (например, VQA) и структурированном моделировании с использованием временной причинно-следственной структуры.

Заключение. В статье делается попытка обзора существующих алгоритмов извлечения правил из ИНС сетей и моделей машинного обучения. Некоторые из современных алгоритмов делятся на три категории – декомпозиционные, педагогические и эклектические. Особое внимание уделяется извлечению правил из нейронечетких сетей. Изучение нечеткой логики достигло кульминации в конце XX в., и с тех пор начало уменьшаться. Это снижение может быть частично связано с отсутствием результатов в машинном обучении. Извлечение правил является одним из способов помочь понять нейронные сети. Эти исследования проложат путь для исследователей нечеткой логики для разработки приложений в области ИИ и решения сложных проблем, которые также представляют интерес для сообщества машинного обучения. Опыт и знания в области нечеткой логики хорошо подходят для моделирования неоднозначностей в больших данных, моделирования неопределенности в представлении знаний и обеспечения обучения передаче с неиндуктивным выводом. Рассматривается также извлечение правил из сетей глубокого обучения, которые в настоящее время обеспечивают приемлемое решение для множества проблем ИИ. Это новая область машинного обучения, которая, как считается, продвигает машинное обучение на шаг вперед в области распознавания образов и понимании текстов. Но с точки зрения объяснений – это все еще модель черного ящика. В последние несколько лет проблема была расширена на общую концепцию и извлечения знаний из алгоритмов машинного обучения – объяснимый ИИ. Достижения в машинном обучении и рост вычислительных мощностей привели к разработке интеллектуальных систем, которые применяются, чтобы рекомендовать фильм, диагностировать злокачественную опухоль, принимать инвестиционные решения или вести машину без во-

дителя. Однако эффективность этих систем ограничена невозможностью объяснить решения и действия пользователю. Программа объяснимого ИИ DARPA разрабатывает и оценивает широкий спектр новых методов машинного обучения: модифицированные методы глубокого обучения, которые изучают объяснимые функции; методы, которые исследуют более структурированные, интерпретируемые причинные модели; методы индуктивных моделей, которые выводят объяснимую модель из любой модели черного ящика. Полученные технологии и результаты показывают, что эти три стратегии заслуживают дальнейшего изучения и предоставят будущим разработчикам варианты дизайна, увеличивающие производительность и объяснимость.

СПИСОК ЛИТЕРАТУРЫ

1. *Bilgic M., Mooney R.J.* Explaining Recommendations: Satisfaction vs. Promotion // Beyond Personalization Workshop. 2005. V. 5. P. 153.
2. *Swartout W.R., Moore J.D.* Explanation in Second Generation Expert Systems. Second Generation Expert Systems. Berlin, Heidelberg: Springer-Verlag, 1993. P. 543–585.
3. *Chandrasekaran B., Tanner M.C., Josephson J.R.* Explaining Control Strategies in Problem Solving // IEEE Expert. 1989. V. 4 (1). P. 9–15.
4. *Dhaliwal J.S., Benbasat I.* The Use and Effects of Knowledge-Based System Explanations: Theoretical Foundations and a Framework for Empirical Evaluation // Information Systems Research. 1996. V. 7 (3). P. 342–362.
5. *Eining M.M., Dorr P.B.* The Impact of Expert System Usage on Experiential Learning in an Auditing Setting // Information Systems. 1991. V. 5 (1). P. 1–16.
6. *Murphy D.S.* Expert System Use and the Development of Expertise in Auditing: A preliminary investigation // Information System, 1990. V. 4. P. 18–35.
7. *Lamberti D.M., Walace W.A.* Intelligent Interface Design: An Empirical Assessment of Knowledge Presentation in Expert Systems // MIS Quarterly. 1990. V. 14. P. 279–311.
8. Искусственный интеллект. Справочник в 3-х т. / Под ред. В.Н. Захарова, Э.В. Попова, Д.А. Поспелова, В.Ф. Хорошевского. М.: Радио и связь, 1990.
9. *Попов Э.В.* Экспертные системы: Решение неформализованных задач в диалоге с ЭВМ. М.: Наука, 1987. 288 с.
10. *Swartout W.R.* A Digitalis Therapy Advisor with Explanations // Proc. 5th International Joint Conf. on Artificial Intelligence. Cambridge, 1977. V. 2. P. 819–825.
11. *Weiner J.L.* BLAH, A System that Explains its Reasoning // Artificial Intelligence, 1980. V. 15. P. 19–48.
12. *Swartout W.R., Paris C., Moore J.* Explanations in Knowledge Systems: Design for Explainable Expert Systems // IEEE Expert. 1991. V. 6 (3). P. 58–64.
13. *Clancey W.J.* Intelligent Tutoring Systems: A Tutorial Survey. Stanford: Stanford University Department of Computer Science, 1986. 56 p.
14. *Sinha R., Swearingen K.* The Role of Transparency in Recommender Systems // CHI'02 Extended Abstracts on Human Factors in Computing Systems. Minneapolis, 2002. P. 830–831.
15. *Gruber T.* Learning Why by Being Told What // IEEE Expert. 1991. V. 6 (4). P. 65–75.
16. *Clancey W.J.* Details of the Revised Therapy Algorithm. Rule-Based Expert Systems. The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, MA: Addison-Wesley, 1984. P. 133–146.
17. *Clancey W.J.* From GUIDON to NEOMYCIN and HERACLES in Twenty Short Lessons // AI Magazine. 1986. V. 7 (3). P. 40.
18. *Arioua A., Buche P., Croitoru M.* Explanatory Dialogs with Argumentative Faculties Over Inconsistent Knowledge Bases // Expert Systems with Applications. 2017. V. 80. P. 244–262.
19. *Johansson U., Lofstrom T., Konig R., Sonstro C., Nilsson L.* Rule Extraction from Opaque Models–a Slightly Different Perspective // 5th Intern. Conf. on Machine Learning and Applications (ICMLA'06). Orlando, FL, USA, 2006. P. 22–27.
20. *Craven M., Shavlik J.* Rule extraction: Where Do We Go from Here. University of Wisconsin Machine Learning Research Group Working Paper. Wisconsin, 1999. P. 99–108.
21. *Andrew, R., Diederich J., Tickle A.B.* Survey and Critique of Techniques for Extracting Rules from Trained Artificial Networks // Knowledge-based Systems. 1995. V. 8 (6). P. 373–389.
22. *Thrum S.* Extracting Provably Correct Rules from Artificial Neural Networks. Technical report. Bonn: Institut für Informatik III, 1993.
23. *Craven M., Shavlik J.W.* Using Sampling and Queries to Extract Rules from Trained Neural Networks // Proc. Eleventh Intern. Conf. Rutgers University. New Brunswick, USA, 1994. P. 37–45.
24. *Fu L.* Rule Generation from Neural Networks // IEEE Transactions on Systems, Man and Cybernetics. 1994. V. 24 (8). P. 1114–1124.

25. *Sato M., Tsukimoto H.* Rule Extraction from Neural Networks via Decision Tree Induction // Proc. Intern. Joint Conf. on Neural Networks (IJCNN'01). Washington, DC, 2001. V. 3. P. 1870–1875.
26. *Tickle A.B., Andrew R., Golea M., Diederich J.* The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded within Trained Artificial Neural Networks // IEEE Transactions on Neural Networks. 1998. V. 9 (6). P. 1057–1068.
27. *Sethi K.K., Mishra D.K., Mishra B.* KDRuleEx: A novel approach for enhancing user comprehensibility using rule extraction // Intelligent Systems, Modelling and Simulation (ISMS), Third Intern. Conf. Kota Kinabalu, Malaysia, 2012. P. 55–60.
28. *Johansson U., Lofstrom T., Konig R., Sonstrod C., Niklasson L.* Rule extraction from opaque models–a slightly different perspective // 5th International Conference on Machine Learning and Applications, ICMLA'06. Orlando, FL, USA, 2006. P. 22–27.
29. *Rangwala M., Weckman G.R.* Extracting Rules from Artificial Neural Networks Utilizing TREPAN // Proceedings of IIE Annual Conference. Orlando, Florida, 2006. 7 p.
30. *Hailesilassie T.* Extraction Algorithm for Deep Neural Networks: A Review// International Journal of Computer Science and Information Security, 2016. V. 14. № 7. P. 376–381.
31. *Averkin A., Yarushev S.* Hybrid Neural Networks and Time Series Forecasting // Artificial Intelligence. Communication in Computer and Information Sciences, 2018. V. 934. P. 230–239.
32. *Pilato G., Yarushev S.A., Averkin A.N.* Prediction and Detection of User Emotions Based on Neuro-Fuzzy Neural Networks in Social Networks // Proc. of the Third International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'18), Advances in Intelligent Systems and Computing. Sochi, Russia. 2018. V. 875. P. 118–126.
33. *Averkin A.N., Pilato G., Yarushev S.A.* An Approach for Prediction of User Emotions Based on ANFIS in Social Networks // Second Intern. Scientific and Practical Conf. Fuzzy Technologies in the Industry. FTI 2018–CEUR Workshop Proceedings. Ostrava–Prague, Czech Republic, 2018. P. 126–134.
34. *Jin X.-H.* Neurofuzzy Decision Support System for Efficient Risk Allocation in Public-private Partnership Infrastructure Projects // J. Comput. Civ. Eng. 2010. V. 24 (6). P. 525–538.
35. *Jin X.-H.* Model for Efficient Risk Allocation in Privately Financed Public Infrastructure Projects Using Neuro-Fuzzy Techniques // J. Constr. Eng. Manag. 2011. P. 1003–1014.
36. *Борисов В.В., Федулов А.С., Зернов М.М.* Основы гибридизации нечетких моделей. Серия "Основы нечеткой математики". Книга 9. М.: Горячая линия–Телеком, 2017. 100 с.
37. *Рудковская Д.* Нейронные сети, генетические алгоритмы и нечеткие системы / Пер. с пол. И.Д. Рудинского. М.: Горячая линия–Телеком, 2008. 452 с.
38. *Rajab S., Sharma V.* A Review on the Applications of Neuro-Fuzzy Systems in Business // Artif. Intell. Rev. 2018. V. 49. P. 481–510.
39. *Mitra S., Hayashi Y.* Neuro-Fuzzy Rule Generation: Survey in Soft Computing Framework // IEEE Trans. Neural Netw. 2000. V. 11 (3). P. 748–768.
40. *Vieira J., Morgado-Dias F., Mota A.* Neuro-Fuzzy Systems: a Survey // WSEAS Transactions on Systems, 2004. V. 3 (2). P. 414–419.
41. *Kim J., Kasabov N.* HyFIS: Adaptive Neuro-Fuzzy Inference Systems and Their Application to Nonlinear Dynamical Systems // Neural Netw., 1999. V. 12 (9). P. 1301–1319.
42. *Shihabudheen K.V., Pillai G.N.* Recent Advances in Neuro-Fuzzy System: A Survey // Knowl.-Based Syst., 2018. V. 152. P. 136–162.
43. *Батыршин И.З., Недосекин А.О., Стецко А.А., Тарасов В.Б., Язенин А.В., Ярушкина Н.Г.* Нечеткие гибридные системы. Теория и практика / Под ред. Н.Г. Ярушкиной. – М.: Физматлит, 2007. – 208 с.
44. *Viharos Z.J., Kis K.B.* Survey on Neuro-Fuzzy Systems and Their Applications in Technical Diagnostics and Measurement // Measurement. 2015. V. 67. P. 126–136.
45. *Lin C.T., Lee C.S.G.* Neural Network based Fuzzy Logic Control and Decision System // IEEE Trans Comput. 1991. V. 40 (12). P. 1320–1336.
46. *Jang J.-S. R.* ANFIS: Adaptive-Network-Based Fuzzy Inference System // IEEE Trans. Systems & Cybernetics. 1993. V. 23. P. 665–685.
47. *Naderpour H., Mirrashid M.* Shear Failure Capacity Prediction of Concrete Beam-Column Joints in Terms of ANFIS and GMDH // Pract. Period. Struct. Des. Constr., 2019. V. 24 (2).
48. *Fan L.* Revisit Fuzzy Neural Network: Demystifying Batch Normalization and Relu with Generalized Hamming Network // Proc. 31st International Conference on Neural Information Processing Systems. Long Beach California USA, 2017. P. 1920–1929.
49. *Bherenji H.R., Khedkar P.* Learning and Tuning Fuzzy Logic Controllers through Reinforcements // IEEE Trans Neural Networks, 1992. V. (3). P. 724–740.
50. *Nauck D., Kruse R.* Neuro-Fuzzy Systems for Function Approximation // Fuzzy Sets and Systems, 1999. V. 101 (2). P. 261–271.

51. *Tano S., Oyama T., Arnould T.* Deep Combination of Fuzzy Inference and Neural Network in Fuzzy Inference // Fuzzy Sets and Systems. 1996. V. 82 (2). P. 151–160.

52. *Juang Chia Feng, Lin Chin Teng.* An Online Self Constructing Neural Fuzzy Inference Network and its Applications // IEEE Transactions on Fuzzy Systems, 1998. V. 6. № 1. P. 12–32.

53. *Kasabov N., Qun Song* Dynamic Evolving Fuzzy Neural Networks with 'm-out-of-n' Activation Nodes for Online Adaptive Systems. Technical Report TR99/04, Department of information science. Otago. University of Otago: 1999.

54. *Gunning D., Aha D.* DARPA's Explainable Artificial Intelligence (XAI) Program // AI Magazine. 2019. V. 40 (2). P. 44–58.

55. *Горбань А.Н.* Ошибки интеллекта, основанного на данных // Интеллектуальные системы в науке и технике. Искусственный интеллект в решении актуальных социальных и экономических проблем XXI в. // Сб. ст. по матер. Междунар. конф. "Интеллектуальные системы в науке и технике" и Шестой всероссийской научно-практической конф. "Искусственный интеллект в решении актуальных социальных и экономических проблем XXI века". Пермь, 2020. С. 11–13.

56. *Hu R., Andreas J., Rohrbach M., Darrell T., Saenko K.* Learning to Reason: End-to-End Module Networks for Visual Question Answering // Proceedings of the IEEE International Conference on Computer Vision. N. Y.: IEEE, 2017. P. 804–813.

57. *Kim J., Canny J.* Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention // Proc. Intern. Conf. on Computer Vision. N. Y.: IEEE, 2017. P. 2942–295.

58. *Hendricks L.A., Darrell T., Akata Z.* Grounding Visual Explanations // European Conf. of Computer Vision (ECCV). Munich, Germany: Springer, 2018.

59. *Marazopoulou K., Maier M., Jensen D.* Learning the Structure of Causal Models with Relational and Temporal Dependence // Proc. Thirty-First Conference on Uncertainty in Artificial Intelligence, Association for Uncertainty in Artificial Intelligence. Amsterdam, Netherlands, 2015. P. 572–581.

60. *Pfeffer A.* Practical Probabilistic Programming. Greenwich, CT: Manning Publications, 2016.

61. *Harradon M., Druce J., Ruttenberg B.* Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations. arXiv preprint. arXiv:1802.00541v1 [cs.AI]. Ithaca, N.Y.: Cornell University Library, 2018.

62. *She L., Chai J. Y.* Interactive Learning for Acquisition of Grounded Verb Semantics towards Human-Robot Communication // Proc. 55th Annual Meeting of the Association for Computational Linguistics. 2017. V. 1. P. 1634–1644.

63. *Qi Z., Li F.* Learning Explainable Embeddings for Deep Networks // NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning. Long Beach, 2017. 4 p.

64. *Dodge J., Penney S., Hilderbrand C., Anderson A., Burnett M.* How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games // Proc. CHI Conference on Human Factors in Computing Systems. N. Y.: Association for Computing Machinery, 2018. P. 1–12.

65. *Belbute-Peres F., Kolter J. Z.* A Modular Differentiable Rigid Body Physics Engine // Neural Information Processing Systems. Deep Reinforcement Learning Sympos. Long Beach, CA, 2017. 7 p.

66. *Hefny A., Marinho Z., Sun W., Srinivasa S., Gordon G.* Recurrent Predictive State Policy Networks // Proc. 35th Intern. Conf. on Machine Learning. International Machine Learning Society. Stockholm, Sweden, 2018. P. 1949–1958.

67. *Vicol P., Tapaswi M., Castrejon L. Fidler S.* MovieGraphs: Towards Understanding Human-Centric Situations from Videos // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. N. Y.: IEEE, 2018. P. 4631–4640.

68. *Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A.* Object Detectors Emerge in Deep Scene CNNs // Paper Presented at the Intern. Conf. on Learning Representations. San Diego, CA, 2015.

69. *Gogate V., Domingos P.* Probabilistic Theorem Proving // Communications of the ACM. 2016. V. 59(7). P. 107–15.

70. *Du M., Liu N. Song Q., Hu X.* Towards Ex Planation of DNN-Based Prediction and Guided Feature Inversion // Proc. 24th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining. N. Y.: Association for Computing Machinery, 2018. P. 1358–67.

71. *Gao J., Liu N., Lawley M., Hu X.* An Interpretable Classification Framework for Information Extraction from Online Healthcare Forums // J. Healthcare Engineering, 2017. V. 2017. 12 p.

72. *Yang S.C.-H., Shafto P.* Explainable Artificial Intelligence via Bayesian Teaching // Paper Presented at the 31st Conf. on Neural Information Processing Systems Workshop on Teaching Machines, Robots and Humans. Long Beach, CA, 2017.