(19) Евразийское патентное ведомство

(21) 201891425 (13) A1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2020.01.31

(22) Дата подачи заявки
2018.07.13

(51) Int. Cl. ***G06N 3/02*** (2006.01)
G06N 3/04 (2006.01)

(54) СПОСОБ ИНТЕРПРЕТАЦИИ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

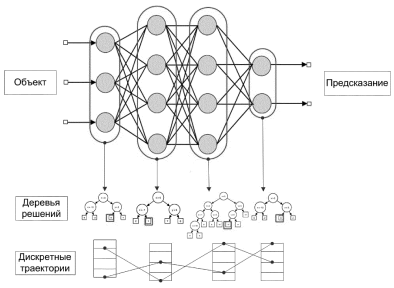
(96) **2018000088 (RU) 2018.07.13**

(71) Заявитель:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Жаров Ярослав Максимович, Корженков Денис Михайлович, Швечиков Павел Дмитриевич (RU)

(74) Представитель:
Герасин Б.В. (RU)

(57) Данное техническое решение, в общем, относится к области вычислительной техники, а в частности к способам и системам интерпретации работы моделей искусственных нейронных сетей. Способ интерпретации искусственных нейронных сетей, в котором получают по меньшей мере одну предварительно обученную на наборе объектов искусственную нейронную сеть; формируют для каждого слоя обученной нейронной сети по меньшей мере одно дерево решений, причем дерево решений получает в качестве входных данных активации соответствующего слоя, полученного при прохождении по нейронной сети объекта, из имеющегося набора данных; предсказывают посредством деревьев решений тот же ответ, который выдает на этом объекте обученная искусственная нейронная сеть; затем получают для каждого объекта упорядоченную последовательность номеров листьев сформированных на предыдущем шаге деревьев решений; далее формируют набор правил, предсказывающий последовательность номеров листьев по объекту. Технический результат - повышение качества и точности интерпретации работы искусственной нейронной сети.



201891425 A1

201891425 A1

МПК G06N 3/008

СПОСОБ ИНТЕРПРЕТАЦИИ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

ОБЛАСТЬ ТЕХНИКИ

[001] Данное техническое решение, в общем, относится к области вычислительной техники, а в частности к способам и системам интерпретации работы моделей искусственных нейронных сетей.

УРОВЕНЬ ТЕХНИКИ

[002] В настоящее время искусственные нейронные сети являются важным инструментом для решения многих прикладных задач. Они уже позволили справиться с рядом непростых проблем и обещают создание новых изобретений, способных решать задачи, которые пока под силу только человеку. Искусственные нейронные сети, также как и биологические, представляют собой системы, состоящие из огромного количества функционирующих процессоров-нейронов, каждый из которых выполняет какой-либо небольшой объём работ, возложенный на него, при этом обладая большим числом связей с остальными, что и характеризует силу вычислений сети.

[003] К одному из недостатков искусственных нейронных сетей относят сложность содержательной интерпретации, и в том числе сложность обоснования активации нейронов. Проблемы интерпретируемости приводят к снижению ценности полученных результатов работы искусственной нейронной сети. Внутреннее представление результатов обучения зачастую настолько сложно, что его невозможно проанализировать даже эксперту в уровне техники, за исключением некоторых простейших случаев, обычно не представляющих интереса.

[004] В настоящее время искусственные нейронные сети используются во многих областях техники, но прежде чем их можно будет применять там, где под угрозу могут быть поставлены человеческие жизни или значительные материальные ресурсы, должны быть решены важные вопросы, касающиеся надежности их работы, в связи с чем интерпретация искусственных нейронных сетей приобретает дополнительную важность.

СУЩНОСТЬ ИЗОБРЕТЕНИЯ

[005] Данное техническое решение направлено на устранение недостатков, присущих существующим решениям из известного уровня техники.

[006] Технической задачей, поставленной в данном техническом решении, является представление правил принятия решения нейронной сетью в виде легко интерпретируемых логических выражений.

[007] Техническим результатом, достигающимся при решении вышеуказанной задачи, является повышение качества и точности интерпретации работы искусственной нейронной сети.

[008] Указанный технический результат достигается благодаря осуществлению способа интерпретации искусственных нейронных сетей, в котором получают по меньшей мере одну предварительно обученную на наборе объектов искусственную нейронную сеть; далее формируют для каждого слоя обученной нейронной сети по меньшей мере одно дерево решений, причем дерево решений получает в качестве входных данных активации соответствующего слоя, полученного при прохождении по нейронной сети объекта, из имеющегося набора данных; предсказывают посредством деревьев решений ответ, который выдает на этом объекте обученная искусственная нейронная сеть; затем получают для каждого объекта упорядоченную последовательность номеров листьев сформированных на предыдущем шаге деревьев решений; после чего формируют набор правил, предсказывающий последовательность номеров листьев по объекту.

[009] В некоторых вариантах осуществления изобретения в качестве меры качества дерева решений используют кросс-энтропию между распределением классов, предсказанным им, и распределением, возвращенным классификатором.

[0010] В некоторых вариантах осуществления изобретения в качестве меры качества дерева решений используют среднюю квадратичную либо абсолютную ошибку между ответом, предсказанным им, и ответом, возвращенным классификатором.

[0011] В некоторых вариантах осуществления изобретения деревья решений строятся независимо друга от друга.

[0012] В некоторых вариантах осуществления изобретения деревья решений строятся зависимо друг от друга на основании алгоритма бустинга.

[0013] В некоторых вариантах осуществления изобретения деревья решений строятся зависимо друг от друга посредством добавления информации о номерах

листьев из деревьев, построенных на предыдущих слоях, на вход дерева следующего слоя.

[0014] В некоторых вариантах осуществления изобретения в качестве алгоритма бустинга используют XGBoost, или AdaBoost, или LPBoost, или TotalBoost, или BrownBoost, или MadaBoost, или LogitBoost.

[0015] В некоторых вариантах осуществления изобретения формируют для каждого слоя обученной нейронной сети дерево решений на основании алгоритма CLS, или ID3, или C4.5, или CART, или IndCART, или DB-CART, или CHAID, или MARS.

[0016] В некоторых вариантах осуществления изобретения формируют дерево решений исходя из минимизации функционала ошибки на все объекты, которые ему подаются на вход.

[0017] В некоторых вариантах осуществления изобретения на вход дереву решений подается исходное признаковое описание для искусственной нейронной сети или описание объекта с измененным набором признаков.

[0018] В некоторых вариантах осуществления изобретения нумеруют листья в каждом дереве решений.

[0019] В некоторых вариантах осуществления изобретения упорядоченная последовательность номеров листьев формируется по номеру слоя в нейронной сети.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

[0020] Признаки и преимущества настоящего технического решения станут очевидными из приводимого ниже подробного описания изобретения и прилагаемых чертежей, на которых:

[0021] На Фиг. 1 приведена блок-схема способа интерпретации искусственных нейронных сетей.

[0022] На Фиг. 2 показан вариант реализации попадания объекта в дерево решений, когда объект попадает в один из листьев дерева;

[0023] На Фиг. 3 показан вариант реализации формирования деревьев решений для каждого слоя ИНС и получающиеся в итоге дискретные траектории по листьям деревьев решений, куда попадает объект.

[0024] На Фиг. 4 показан вариант реализации формирования дерева решений, предсказывающего последовательность номеров листьев по объекту.

[0025] На Фиг. 5 показан пример реализации, согласно которому выбрано из общего набора минимальное количество построенных дискретных траекторий, которые перекрывают не менее 90% примеров.

[0026] На Фиг. 6 показано усреднение объектов в разных дискретных траекториях, приводящее к высокой вероятности символа 2.

[0027] На Фиг. 7 показан пример реализации изображений дискретной траектории с только пятью примерами различного написания символа 3.

[0028] На Фиг. 8 показан вариант полученной дискретной траектории, ведущей к высокой вероятности символа 7, и бинаризованные, и взвешенные правила, приводящие к этой траектории.

ПОДРОБНОЕ ОПИСАНИЕ ИЗОБРЕТЕНИЯ

[0029] Ниже будут описаны понятия и термины, необходимые для понимания данного технического решения.

[0030] В данном техническом решении под системой подразумевается, в том числе компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, четко определенную последовательность операций (действий, инструкций).

[0031] Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы).

[0032] Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройств хранения данных. В роли устройства хранения данных могут выступать, но не ограничиваясь, жесткие диски (HDD), флеш-память, ПЗУ (постоянное запоминающее устройство), твердотельные накопители (SSD), оптические приводы.

[0033] Программа - последовательность инструкций, предназначенных для исполнения устройством управления вычислительной машины или устройством обработки команд.

[0034] Искусственная нейронная сеть (далее - ИНС) - вычислительная или логическая схема, построенная из однородных процессорных элементов, являющихся упрощенными функциональными моделями нейронов.

[0035] Дерево решений (англ. decision tree) — граф, схема, отражающая структуру задачи оптимизации для многошагового процесса принятия решений. Применяется в задачах классификации и в других областях для анализа решений, структуризации проблем. Ветви дерева отображают различные события, которые могут иметь место, а узлы (вершины) — состояния, в которых возникает необходимость выбора. В данном техническом решении это способ деления пространства, так как каждое условие разделяет пространство гиперплоскостью перпендикулярной оси, по которой осуществлен выбор.

[0036] Boosting — это мета-алгоритм машинного обучения для выполнения обучение с учителем.

[0037] Нейронная сеть полносвязная (англ. Fully connected neural network) - нейронная сеть, в которой каждый нейрон передает свой выходной сигнал на вход нейронов следующего слоя.

[0038] Слой нейронной сети (англ. layer) - совокупность нейронов сети, объединяемых по особенностям их функционирования.

[0039] Функция активации нейронной сети - функция, которая используется для преобразования уровня активации элемента (нейрона) в выходной сигнал. Обычно функция активации имеет «сжимающее» действие.

[0040] Предикат - элемент атрибутивного суждения, обозначающий какой-либо признак (свойство) его субъекта, или то, что говорится о субъекте.

[0041] Кортеж — в математике последовательность конечного числа элементов. Например, граф определяется как кортеж (V,E), где V — это набор вершин, а E — подмножество V × V, обозначающее ребра. В теории множеств, кортеж обычно определяется индуктивно.

[0042] Способ интерпретации искусственных нейронных сетей, показанный на Фиг. 1 в виде-блок-схемы, может включать следующие шаги.

[0043] Задача состоит в классификации и регрессии набора данных с помощью искусственной нейронной сети и последующего анализа полученной сети с целью нахождения правил, характеризующих процесс трансформации входных данных сетью для получения ответа.

[0044] **Шаг 110**: получают по меньшей мере одну предварительно обученную на наборе объектов искусственную нейронную сеть.

[0045] Пусть в данном техническом решении получают предварительно обученную нейронную сеть с D слоями (каждый слой содержит S_i нейронов) с размерами

слоев $S_1, \ldots, S_D$, и обучающую выборку $\{x_i, y_i\}_{i=1}^{N}$. На данном шаге собирают дополнительные данные для дальнейшей обработки, а именно для каждого слоя нейронной сети $d = 1, \ldots, D$ получают все выходы нейронов (определенные как активации) для каждого обучающего набора объектов как D матриц A_d размера $N \times S_d$. Также формируют результаты полученных вероятностей классов как матрицу P размером $N \times C$, где C – размерность выхода (количество классов или размерность переменных регрессии).

[0046] Количество слоев искусственной нейронной сети не ограничено вариантами реализации. В качестве обученной нейронной сети могут использовать полносвязную нейронную сеть, или сверточную нейронную сеть, или рекуррентную нейронную сеть или их комбинацию, не ограничиваясь.

[0047] В качестве функции активации могут использовать выпрямленную линейную функцию активации (rectified linear unit, ReLU), которая выражается следующей формулой:

$$f(s) = \max(0, s)$$

[0048] В некоторых вариантах реализации могут использоваться функции активации Sigmoid, tanh, LeakyReLU, PReLU и другие известные из уровня техники, не ограничиваясь.

[0049] Искусственная нейронная сеть может быть обучена на наборе объектов, которые, например, представляют собой изображения. Обучающая выборка может состоять из положительных и отрицательных примеров. В некоторых вариантах реализации обучающая выборка может состоять из примеров произвольного количества классов, может быть вообще не размечена, а может быть и задачей регрессии, где классы – неприменимое понятие. Например, если ИНС обучается для детекции лиц людей, обучающая выборка состоит из изображений, которые содержат лица и изображений, на которых отсутствуют лица. Соотношение положительных примеров к отрицательным примерам может быть выбрано как N:M, например, 4:1, например 8000 положительных и 2000 отрицательных.

[0050] В конкретном примере реализации в качестве положительной обучающей выборки может использоваться база данных LFW3D. Она содержит цветные изображения фронтальных лиц типа JPEG, размером 90x90 пикселей, в количестве 13000.

[0051] В качестве отрицательных обучающих примеров может использоваться для обучения база данных SUN397, которая содержит огромное количество всевозможных сцен, которые разбиты по категориям. Всего данная база содержит 130000 изображений, 908 сцен, 313000 объектов сцены. Общий вес этой базы составляет 37 GB.

[0052] В случае использования полносвязной или сверточной нейронной сети и когда объектами являются рукописные цифры, может использоваться для обучения база данных MNIST, а именно объёмная база данных образцов рукописного написания цифр.

[0053] Искусственная нейронная сеть передается и хранится как архитектура и значения весов в ее слоях. В некоторых вариантах реализации архитектура сети и значения весов в слоях хранятся отдельно для того, чтобы можно было загрузить значения весов в сеть с другой архитектурой. Такой подход используется, например, при совмещении обучения без учителя и с учителем. На первом этапе выполняется обучение без учителя с использованием автокодировщика, глубокой сети доверия или другого метода. Затем полученные веса загружаются в сеть другой архитектуры, которая дообучается стандартным подходом обучения с учителем с помощью метода обратного распространения ошибки. Совмещение двух способов позволяет обучать сеть в случае, когда мало размеченных данных для обучения. Значения весов в слоях могут хранить, например, в формате данных HDF5. Содержимое файлов HDF5 организовано подобно иерархической файловой системе, и для доступа к данным применяются пути, сходные с POSIX-синтаксисом, например, /path/to/resource. Метаданные хранятся в виде набора именованных атрибутов объектов.

[0054] В некоторых вариантах реализации осуществляют нормировку множества объектов для ИНС, поскольку нейронные сети лучше работают с данными, представленными числами, нормально распределенными вокруг 0, а исходные данные могут иметь произвольный диапазон или вообще быть не числовыми данными. При этом возможны различные способы, начиная от простого линейного преобразования в требуемый диапазон и заканчивая многомерным анализом параметров, и нелинейной нормировкой, в зависимости от влияния параметров друг на друга.

[0055] **Шаг 120**: формируют для каждого слоя обученной нейронной сети по меньшей мере одно дерево решений.

[0056] Дерево решений состоит из вершин двух типов, как показано на Фиг. 2. Вершины решений, содержащие условия, обозначаются окружностями. Цели или логические выводы обозначаются прямоугольниками. Вершины нумеруются и на дугах задаются условия. Каждая вершина может иметь не более одного входа. Пути движения по дереву с верхнего уровня на самые нижние определяют логические правила в виде цепочек конъюнкций. Правила, выражающие закономерности, формулируются в виде продукций: «ЕСЛИ A ТО B» или в случае множества условий: «ЕСЛИ (условие 1) ∧ (условие 2) ∧ … ∧ (условие N) ТО (Значение вершины вывода)».

[0057] Формирование деревьев решений может осуществляться на основе экспертных оценок или с использованием алгоритмов обработки примеров (CLS, ID3 - Interactive Dichotomizer, C4.5, CART - classification and regression trees, IndCART, DB-CART, автоматический детектор взаимодействия Хи-квадрат (CHAID), MARS, не ограничиваясь и др.).

[0058] В одном варианте реализации деревья решений для каждого слоя нейронной сети могут строиться независимо друг от друга. Дерево решений представляет собой список конъюнкций как показано на Фиг. 2, где в каждой вершине дерева решений находятся предикаты или правила (показано как x и y), которые включает в себя два аргумента: конкретный признак объекта, который подается на вход дереву решений, а также частное значение данного признака. Непосредственно признаки объекта подаются в дерево решений только на нулевом (входном) слое. На всех остальных слоях деревья решений принимают активации, полученные при прохождении объекта через соответствующий слой. Функция активации — функция, принимающая взвешенную сумму как аргумент. Значение этой функции и является выходом нейрона в ИНС.

[0059] На данном шаге получают D независимых деревьев решений T_d (для прогнозирования конечных вероятностей P из A_d), обученных на матрицах A_d как вход и конечные вероятности P как выход. Этот вариант реализации можно рассматривать как особый тип создания множества деревьев с усреднением деревьев, где каждое дерево решений строится на отдельных входных данных.

[0060] Если предикат выполняется для конкретного поданного значения признака объекта в это дерево решений, объект дерева попадают в правую или левую ветку в зависимости от выполнения условия предиката. Дерево решений строится исходя из минимизации функционала ошибки на все объекты, которые ему подаются на вход. Целью дерева решений является выполнение как можно более

точного прогноза. Для оценки точности используют функционал ошибки (или, если взять его с обратным знаком, функционал качества.) Чем ниже значения функционала ошибки – тем лучше дерево решений решает задачу. Каждое новое условие в вершине дерева решений выбирается так, чтобы в результате как можно сильнее уменьшить общий функционал ошибки.

[0061] В некоторых вариантах реализации в процедуре формирования дерева решений принимают участие все признаки всех объектов обучающей выборки для ИНС. В качестве входных признаков для первого дерева решений на нулевом слое выступают признаки объектов, а для очередного дерева решений выступают значения активаций соответствующего слоя на объектах выборки.

[0062] В некоторых вариантах реализации формируют регрессионные деревья решений, которые предсказывают не вероятность той или иной классификации на последнем слое ИНС, а предсказывают вещественные значения (например, баллы классов), которые могут быть однозначно преобразованы в значения вероятности. При этом обратное заключение неверно, так как одним и тем же значениям вероятности на выходе ИНС могут соответствовать разные вещественные значения. Несколько разных наборов таких значений при применении детерминированной процедуры преобразования могут давать одно и то же значение вероятности на последнем слое ИНС.

[0063] В другом варианте реализации деревья решений строятся не независимо друга от друга, а зависимо на основании способа бустинга. Бустинг — это процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. Бустинг представляет собой жадный алгоритм построения композиции алгоритмов. В данном варианте реализации последовательно формируются деревья решений, причем каждое из последующих деревьев добавляет нечто в ансамбль, исправляет или улучшает предыдущий результат. В зависимости от номера дерева, которое формируется в данный момент времени, меняются и данные, на которых это дерево строится. То есть если сформирован ансамбль деревьев решений до N-го слоя, то (N+1)-е дерево решений формируется на активациях (N+1)-го слоя ИНС, таким образом, каждое дерево решений ставится в соответствие слою ИНС. В конкретном варианте реализации формируют первую модель (имеется в виду дерево решений), затем вторую, которая пытается исправить ошибки первой, затем третью, которая пытается исправить ошибку первых двух, и т.д. При этом

исправление ошибки обычно происходит следующим образом: берется функционал ошибки уже готовой композиции деревьев решений, вычисляется его градиент – то есть направление скорейшего возрастания ошибки, и следующая модель пытается приблизить этот градиент, чтобы затем ее предсказание, будучи вычтенным, было направлено уже в сторону скорейшего убывания ошибки.

[0064] На данном шаге формируют $D-1$ деревьев посредством бустинга. Обучение происходит следующим образом. Выбирают баллы A_D как цель для оценки. Формируют базовую оценку, прогнозирующую средние баллы от всех обучающих наборов и принимают эту оценку как текущую наилучшую оценку C. Для каждого слоя $d = 1, \dots, D-1$ формируют новое дерево решений по следующим шагам:

a) Подготавливают новую цель как $\mu = A_D - C$, чтобы позволить следующей оценке исправить ошибку предыдущего;
b) Формируют дерево решений T_d, которое получает на вход A_dи дает на выходе μ как цель;
c) Собирают обучающий набор предсказаний текущего дерева в качестве предсказания;
d) Выбирают коэффициент β посредством минимизации функционала кросс-энтропии (Softmax ($C - \beta * prediction$), P) по β;
e) Выбирают новую $C \leftarrow C - \beta * prediction$.

[0065] Таким образом, эти деревья решений можно рассматривать как особый случай алгоритма бустинга, где используются разные данные для каждого нового дерева, а также выбирается β с целью, отличной от цели дерева решений. Эта стратегия состоит в том, что в данной реализации нет множества независимых деревьев, но деревья, которые пытаются получить новую информацию из каждого слоя, не забывая о предыдущих разбиениях.

[0066] В некоторых вариантах реализации в алгоритме бустинга над деревьями решений для построения каждого дерева решений используется один и тот же набор признаков, из которого делается случайная подвыборка.

[0067] В качестве алгоритма бустинга может использоваться XGBoost, AdaBoost, LPBoost, TotalBoost, BrownBoost, MadaBoost, LogitBoost, не ограничиваясь.

[0068] В еще одном варианте реализации для формирования деревьев решений используют следующий алгоритм. Признаки каждого объекта обучающей выборки для i-го дерева пополняются номером листа, в который он попал в предыдущем $i-1$ дереве или во всех предыдущих деревьях от 1 до $i-1$.

[0069] **Шаг 120.1**: дерево решений получает в качестве входных данных активации соответствующего слоя, полученного при прохождении по нейронной сети объекта, из имеющегося набора.

[0070] **Шаг 120.2**: предсказывают посредством деревьев решений тот же ответ, который выдает на этом объекте обученная искусственная нейронная сеть.

[0071] Ансамблем деревьев решений осуществляют дистилляцию ответа нейронной сети, т.е. формируют такие же ответы посредством дерева решений на объектах, какие на них выдает сеть.

[0072] Чем глубже деревья решений, тем больше количество потенциально возможных дискретных траекторий, что плохо. Однако чем менее глубоки деревья решений, тем хуже качество предсказания итогового ансамбля решений.

[0073] **Шаг 130**: получают для каждого объекта упорядоченную последовательность номеров листьев сформированных на предыдущем шаге деревьев решений.

[0074] Упорядоченная последовательность номеров листьев деревьев решений формируется по номеру слоя в нейронной сети. Имеется в виду, что если в нулевом дереве решений объект был в листе a, в первом - в листе b, а в третьем – c листе, то в итоге эта «упорядоченная последовательность номеров листьев» сводится к кортежу «a, b, c». Необходимым условием является то, чтобы порядок был одинаковым для всех объектов и в некоторых вариантах реализации изобретения может иметь другой принцип формирования.

[0075] Так для одного входа X получают дискретный набор чисел (дискретную траекторию) следующим образом:

a) Получают соответствующие активации $a_1, \ldots, a_D$;
b) Пропускают активации через связанные деревья T и собирают идентификаторы прогнозируемых листьев $L = L_1, \ldots, L_D$.

[0076] После формирования набора траекторий активаций как декартовых произведений индикаторов листьев деревьев решений, можно выбрать наиболее интересующие. В некоторых вариантах реализации листья в каждом дереве решений нумеруют, чтобы иметь возможность однозначно к ним обращаться. По сути, дискретная траектория является кортежем, где подряд находятся листы деревьев, в которые попал объект по слоям ИНС. Например, присутствуют два дерева решений, причем у первого дерева решений имеется пять листьев и у второго пять листьев, которые сформированы для разных слоев, таким образом, общее количество потенциально возможных траекторий достигает двадцати пяти.

Получается дискретная версия той траектории, по которой прошел объект как совокупность его признаков, которые подаются на вход ИНС, при активации нейронов ИНС.

[0077] **Шаг 140**: формируют набор правил, предсказывающий последовательность номеров листьев по объекту.

[0078] На основании всего ансамбля деревьев решений по объектам и их активациям анализируют, что выполняет с объектом ИНС в том или ином слое. На данном шаге необходимо понять, по каким признакам исходного описания нейронная сеть определяет объект в ту или иную дискретную траекторию.

[0079] Таким образом решается задача предсказания с помощью набора правил номера дискретной траектории по исходному набору интерпретируемых признаков объекта. В других вариантах реализации набор интерпретируемых признаков не является исходным, а построенным на основании исходных признаков. Необходимо определить правила, описывающие, в какую траекторию попадет какой объект, т.е. номер дискретной траектории. Например, при анализе текста в нейронную сеть подается сам образец текста целиком. В качестве упомянутого преобразования может служить формирование «мешка слов» (англ. «bag of words»), когда записывают в хранилище данных, сколько раз встретилось каждое слово в тексте, при этом все слова приводятся к своей начальной форме, в результате чего появляется цифровое описание текста. Данное преобразование однозначно выполняется в одну сторону (от текста к мешку слов) и неоднозначно в обратную сторону, так как по мешку слов невозможно однозначно восстановить исходный текст.

[0080] Исходное признаковое пространство, на котором была обучена ИНС, подается на вход набору правил. В качестве алгоритма построения правил можно использовать дерево решений. В некоторых вариантах реализации на вход дереву решений подается не исходное признаковое описание, а описание данного объекта с измененным набором признаков. Данный набор может изменяться таким образом, чтобы помогать строить экспрессивные правила. Правила будут выглядеть как «ЕСЛИ интерпретируемый_признак_1 < значение_1 И … И интерпретируемый_признак_n < значение_n, ТО объект проследует по дискретной траектории x». Например, ИНС обучают предварительно на фотографиях людей, а дереву решений подают на вход цвет волос, расстояние между зрачками и т.д., сформированные в виде таблицы. В другом примере обучают нейронную сеть на последовательности транзакций пользователя, однако данный набор данных не

является интерпретируемым. Измененным интерпретируемым признаком будет следующее: какую долю суммы пользователь тратит по понедельникам, вторникам, средам и т.д. Таким образом, дерево решений строят уже на наборе измененных признаков.

[0081] В некоторых вариантах осуществления формируют коэффициент для каждого объекта, который определяет его важность. Во время процедуры построения финальных правил, разные примеры могут иметь разную важность, то есть степень влияния на получающиеся правила. Например, коэффициент может формироваться на основании того, насколько предсказание искусственной нейронной сети на объекте отличается от предсказания ансамбля послойных деревьев решений.

[0082] В качестве глобального критерия завершения работы технического решения могут быть использованы максимальный размер дерева и общая оценка качества классификации примеров деревом. Разумеется, чем глубже дерево (длиннее набор правил), тем точнее оно предскажет дискретную траекторию. Но слишком большое количество правил осложняет интерпретацию. После построения дерева решений необходимо еще выполнить постобработку правил по одному из известных алгоритмов. В некоторых вариантах реализации построение набора правил также может происходить и напрямую, без участия дерева решений, например посредством алгоритма RuleFit.

[0083] Основным преимуществом данного подхода является дискретизация последовательности активаций ИНС, что позволяет получать интерпретацию с помощью правил. Причем сервер включает в себя различные аппаратные компоненты, включая один или несколько одно- или многоядерных процессоров, которые представлены процессором, графическим процессором (GPU), твердотельным накопителем, ОЗУ, интерфейсом монитора и интерфейсом ввода/вывода.

[0084] Связь между различными компонентами сервера может осуществляться с помощью одной или нескольких внутренних и/или внешних шин (например, шины PCI, универсальной последовательной шины, высокоскоростной шины IEEE 1394, шины SCSI, шины Serial ATA и так далее), с которыми электронными средствами соединены различные аппаратные компоненты. Интерфейс монитора может быть соединен с монитором (например, через HDMI-кабель), видимым оператору, интерфейс ввода/вывода может быть соединен с сенсорным экраном,

клавиатурой (например, через USB-кабель) и мышью (например, через USB-кабель), причем как клавиатура, так и мышь используются оператором.

[0085] В соответствии с вариантами осуществления настоящей технологии твердотельный накопитель хранит программные инструкции, подходящие для загрузки в ОЗУ и использующиеся процессором и/или графическим процессором GPU для отбора данного целевого признака процесса из множества признаков и данного типа выходных значений, как будет описано ниже. Например, программные инструкции могут представлять собой часть библиотеки или приложение.

[0086] Сервер может быть настольным компьютером, ноутбуком, планшетом, смартфоном, персональным цифровым органайзером (PDA) или другим устройством, которое может быть выполнено с возможностью реализовать настоящую технологию, как будет понятно специалисту в данной области техники.

[0087] Сервер может быть выполнен с возможностью осуществлять алгоритм машинного обучения (MLA) и выполнять различные способы для обучения MLA. В некоторых вариантах осуществления настоящей технологии MLA может быть либо искусственной нейронной сетью, Байесовой сетью, машиной опорных векторов и т.д. В другом варианте осуществления настоящей технологии MLA может быть моделью прогнозирования, которая включает в себя набор деревьев решений для решения, среди прочего, задач регрессии и классификации. В этом случае MLA может быть обучен с помощью способов машинного обучения, например, градиентного бустинга (gradient boosting).

[0088] Сервер может быть выполнен с возможностью осуществлять множество процедур, причем по меньшей мере одна из множества процедур является созданием обучающей выборки для обучения MLA. В общем случае MLA может быть обучен прогнозировать расчетные погрешности, присущие способам расчетов. То, как может быть выполнен сервер для создания обучающей выборки для обучения MLA, будет описано ниже.

[0089] В некоторых вариантах осуществления настоящей технологии сервер может быть выполнен с возможностью получать доступ к данным истории, связанной с финансовыми транзакциями пользователя или другими данными. Данные истории могут быть сохранены локально на твердотельном накопителе сервера. В других вариантах осуществления настоящей технологии данные истории могут быть сохранены удаленно на носителе информации, который

функционально соединен с сервером по сети. В этом случае сервер может извлекать данные истории из носителя информации по сети.

[0090] Основное преимущество данного технического решения заключается в обобщающей способности искусственных нейронных сетей, что позволяет получать более простые деревья решений, т.е. выполняется так называемая дистилляция. Дистилляция осуществляется не стандартная и известная из уровня техники, а позволяющая определять какие объекты ИНС склонна обрабатывать схожим образом, т.е. какие паттерны в данных различает именно имеющаяся обученная нейронная сеть. Таким образом, изобретение позволяет извлекать структурированные знания не только из чрезвычайно упрощенных нейронных сетей, но и из нейронных сетей, которые интерпретируются в готовом виде, не упрощая их структуру в процессе анализа, что делает возможным его применение в широком круге практических задач.

[0091] Модули, описанные выше и используемые в данном техническом решении, могут быть реализованы с помощью электронных компонентов, используемых для создания цифровых интегральных схем. Не ограничиваюсь, могут использоваться микросхемы, логика работы которых определяется при изготовлении, или программируемые логические интегральные схемы (ПЛИС), логика работы которых задается посредством программирования. Для программирования используются программаторы и отладочные среды, позволяющие задать желаемую структуру цифрового устройства в виде принципиальной электрической схемы или программы на специальных языках описания аппаратуры: Verilog, VHDL, AHDL и др. Альтернативой ПЛИС являются: программируемые логические контроллеры (ПЛК), базовые матричные кристаллы (БМК), требующие заводского производственного процесса для программирования; ASIC - специализированные заказные большие интегральные схемы (БИС), которые при мелкосерийном и единичном производстве существенно дороже.

[0092] Также модули могут быть реализованы с помощью постоянных запоминающих устройств (см. Лебедев О.Н. Микросхемы памяти и их применение. - М.: Радио и связь, 1990. - 160 с; Большие интегральные схемы запоминающих устройств: Справочник/ А.Ю. Горденов и др. - М.: Радио и связь, 1990. - 288 с).

[0093] Таким образом, реализация всех используемых блоков достигается стандартными средствами, базирующимися на классических принципах реализации основ вычислительной техники, известных из уровня техники.

ПРИМЕРЫ РЕАЛИЗАЦИИ

[0094] Данное техническое решение может быть реализовано с ипсользованием выборки данных MNIST и полносвязной ИНС.

[0095] Модель сети будет объяснена ниже. В данном техническом решении обучили сеть с прямой связью, сконфигурированную следующим образом: (784, 200, ReLU) - (200, 200, ReLU) - (200, 200, ReLU) - (200, 10, LogSoftmax). Тройка данных (784, 200, ReLU) подразумевает, что есть слой, у которого входная размерность 784, выходная размерность 200, а функция активации ReLU.

[0096] Модель была натренирована с оптимизатором Adam с шагом обучения 3e-4. Сеть обучалась в течение 10 эпох и достигла точности 97,5% на тестовом наборе. Активации получали после каждого использования функции активации ReLU, а также перед функцией LogSoftmax (оценка для каждого класса), таким образом, получили три множества активаций, размером в 200 значений, и одно множество активаций, размером 10 значений.

[0097] Далее в конкретном варианте осуществления используется стратегия формирования деревьев решений посредством бустинга. Затем ограничили максимальную глубину деревьев решений значением три и минимальным количеством объектов в листе – 3% от размера выборки. В качестве деревьев принятия решений используется библиотека Scikit-learn. Оценки классов были выбраны в качестве цели для бустинга. Кросс энтропия ансамбля деревьев решений оказалась равно 0,54. Объекты с верхней 5%-ой ошибкой при бустинге были исключены из обучающего набора для конечного дерева.

[0098] Далее выбрали минимальное количество дискретных траекторий, которые включают не менее 90% примеров обучающей выборки, как показано на Фиг. 5. Выбранные дискретные траектории сохраняют свой исходный идентификатор, пока все другие траектории отмечаются как -1. Различные дискретные траектории, ведущие к одному и тому же номеру, показаны на Фиг. 6.

[0099] Затем формируется окончательное дерево решений, в котором исходное изображение является входными данными, и номер дискретной траектории является целью для определения. Таким образом, у нас есть задача с несколькими N + 1 дискретными траекториями, где выбраны N дискретных траекторий с предыдущего шага и 1 траектория содержит все остальные образцы. На Фиг. 8 показана дискретная траектория, которая с высокой вероятностью относится к символу 7 и правила, которые определяют эту траекторию. В двоичных правилах изображения пиксели имеют красный цвет (в ч/б варианте

фигуры не показано), если он был включен в окончательное дерево решений с отрицательным значением (значение пикселя должно быть ниже, чем порог установленный в узле дерева), зеленый цвет имеют пиксели с положительным значением, в ином случае – желтый (в ч/б варианте фигуры цвета не показаны). Во взвешенной части указывается каждый пиксель о значении яркости, пропорциональным количеству примеров, на которые повлияло это правило.

[00100] Для специалиста в данном уровне техники очевидно, что в настоящем описании выражение «получение данных» от пользователя подразумевает получение электронным устройством данных от вычислительной системы, сервера и т.д. в виде электронного (или другого) сигнала. Кроме того, специалисты в данной области техники поймут, что отображение данных пользователю через графический интерфейс пользователя (например, экран электронного устройства и тому подобное) может включать в себя передачу сигнала графическому интерфейсу пользователя, этот сигнал содержит данные, которые могут быть обработаны, и по меньшей мере часть этих данных может отображаться пользователю через графический интерфейс пользователя.

[00101] Некоторые из этих этапов, а также передача-получение сигнала хорошо известны в данной области техники и поэтому для упрощения были опущены в конкретных частях данного описания. Сигналы могут быть переданы-получены с помощью оптических средств (например, оптоволоконного соединения), электронных средств (например, проводного или беспроводного соединения) и механических средств (например, на основе давления, температуры или другого подходящего параметра).

[00102] Модификации и улучшения вышеописанных вариантов осуществления настоящей технологии будут ясны специалистам в данной области техники. Предшествующее описание представлено только в качестве примера и не несет никаких ограничений. Таким образом, объем настоящей технологии ограничен только объемом прилагаемой формулы изобретения.

ПАТЕНТНАЯ ФОРМУЛА

1. Способ интерпретации искусственных нейронных сетей, включающий следующие шаги:

    - получают по меньшей мере одну предварительно обученную на наборе объектов искусственную нейронную сеть;

    - формируют для каждого слоя обученной нейронной сети по меньшей мере одно дерево решений, причем

        - дерево решений получает в качестве входных данных активации соответствующего слоя, полученного при прохождении по нейронной сети объекта, из имеющегося набора данных;

        - предсказывают посредством деревьев решений тот же ответ, который выдает на этом объекте обученная искусственная нейронная сеть.

    - получают для каждого объекта упорядоченную последовательность номеров листьев сформированных на предыдущем шаге деревьев решений;

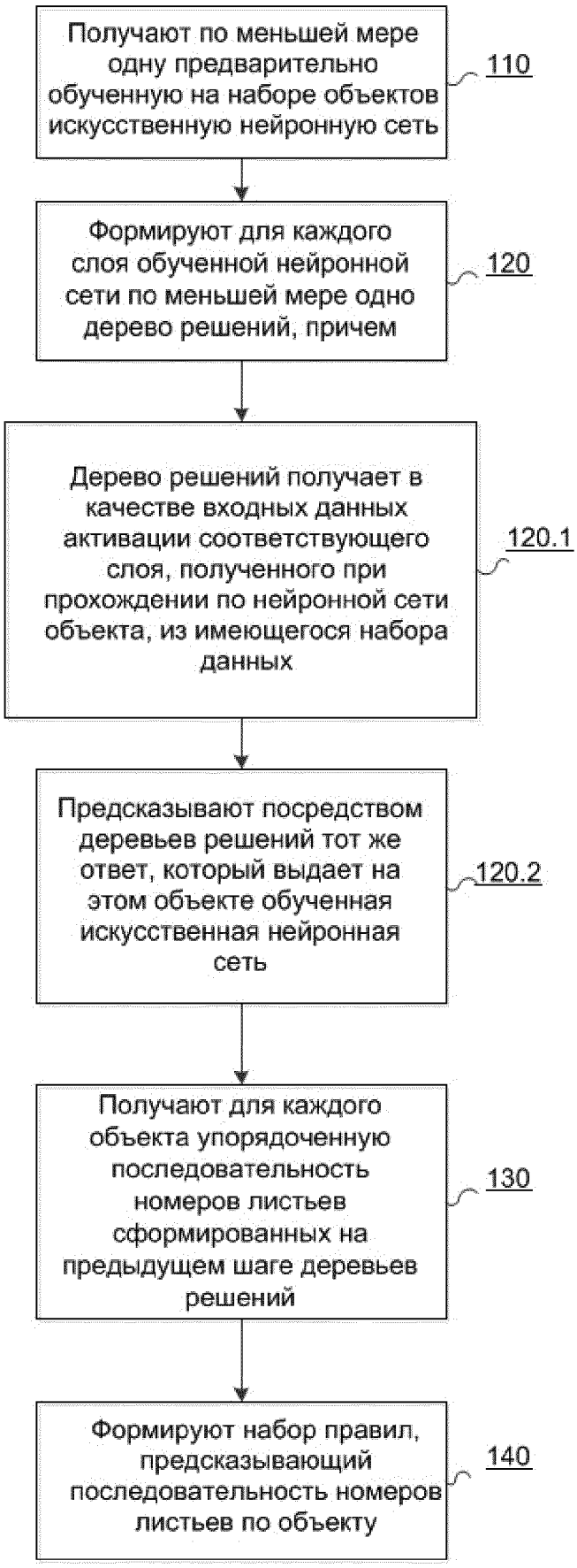
    - формируют набор правил, предсказывающий последовательность номеров листьев по объекту.

2. Способ по п.1, характеризующийся тем, что в качестве меры качества дерева решений используют кросс-энтропию между распределением классов, предсказанным им, и распределением, возвращенным ИНС.
3. Способ по п.1, характеризующийся тем, что в качестве меры качества дерева решений используют среднюю квадратичную либо абсолютную ошибку между ответом, предсказанным им, и ответом, возвращенным ИНС.
4. Способ по п.1, характеризующийся тем, что деревья решений строятся независимо друга от друга.
5. Способ по п.1, характеризующийся тем, что деревья решений строятся зависимо друг от друга на основании алгоритма бустинга.
6. Способ по п.1, характеризующийся тем, что деревья решений строятся зависимо друг от друга посредством добавления информации о номерах

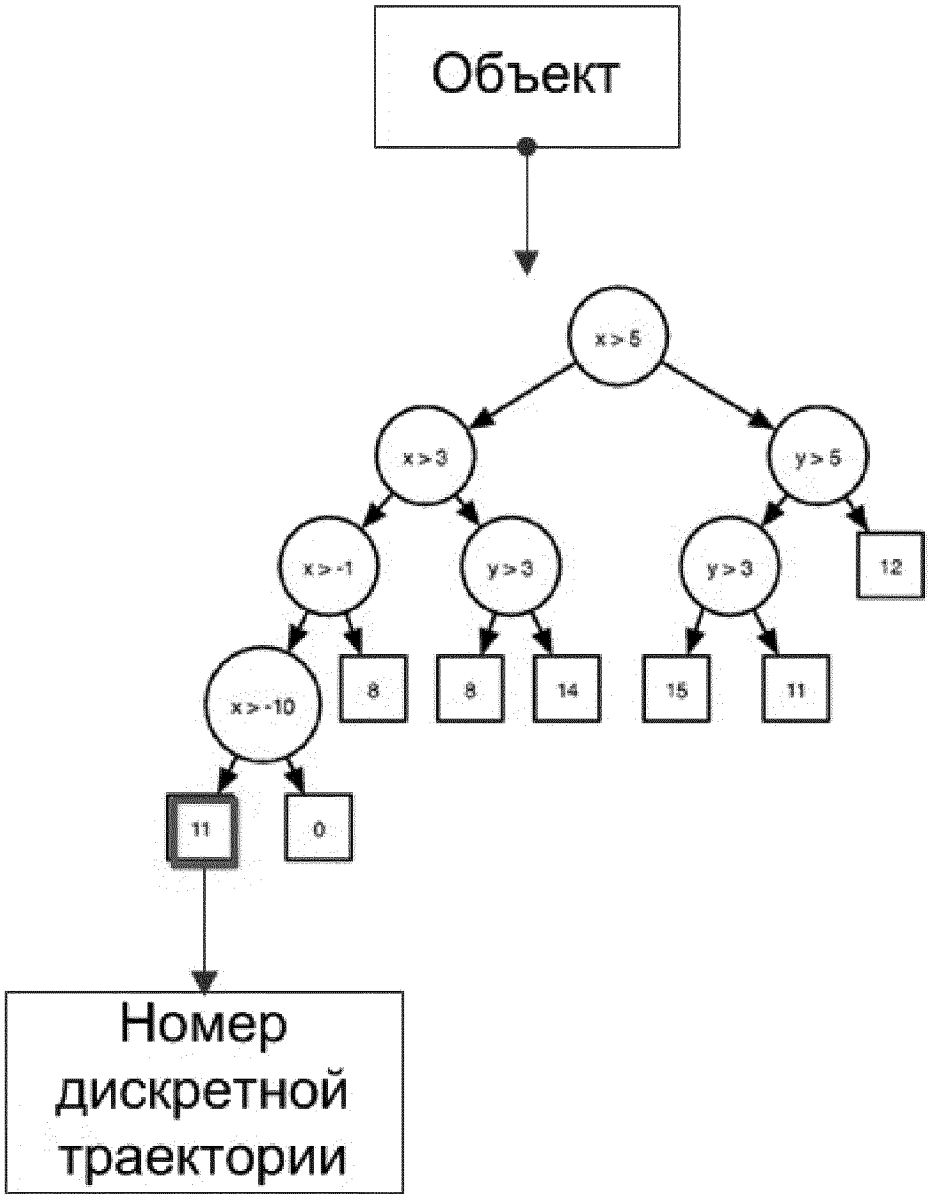
листьев из деревьев, построенных на предыдущих слоях, на вход дерева следующего слоя.

7. Способ по п.5 и п.6, характеризующийся тем, что в качестве алгоритма бустинга используют XGBoost, или AdaBoost, или LPBoost, или TotalBoost, или BrownBoost, или MadaBoost, или LogitBoost.
8. Способ по п.1, характеризующийся тем, что формируют для каждого слоя обученной нейронной сети дерево решений на основании алгоритма CLS, или ID3, или C4.5, или CART, или IndCART, или DB-CART, или CHAID, или MARS.
9. Способ по п.1, характеризующийся тем, что формируют дерево решений исходя из минимизации функционала ошибки на все объекты, которые ему подаются на вход.
10. Способ по п.1, характеризующийся тем, что на вход дереву решений подается исходное признаковое описание для искусственной нейронной сети или описание объекта с измененным набором признаков.
11. Способ по п.1, характеризующийся тем, что нумеруют листья в каждом дереве решений.
12. Способ по п.1, характеризующийся тем, что упорядоченная последовательность номеров листьев формируется по номеру слоя в нейронной сети.

ЧЕРТЕЖИ К ОПИСАНИЮ

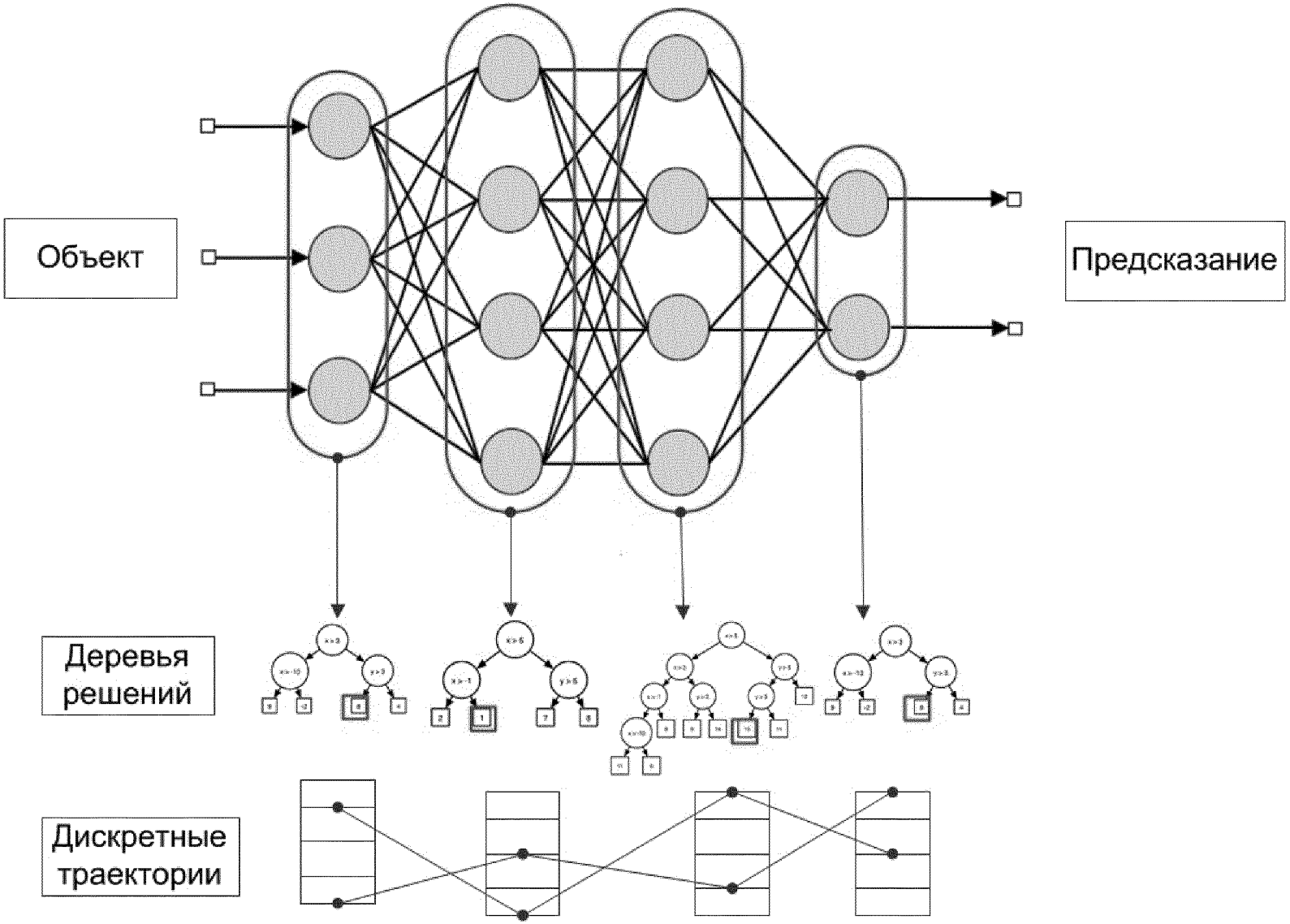


Фиг. 1

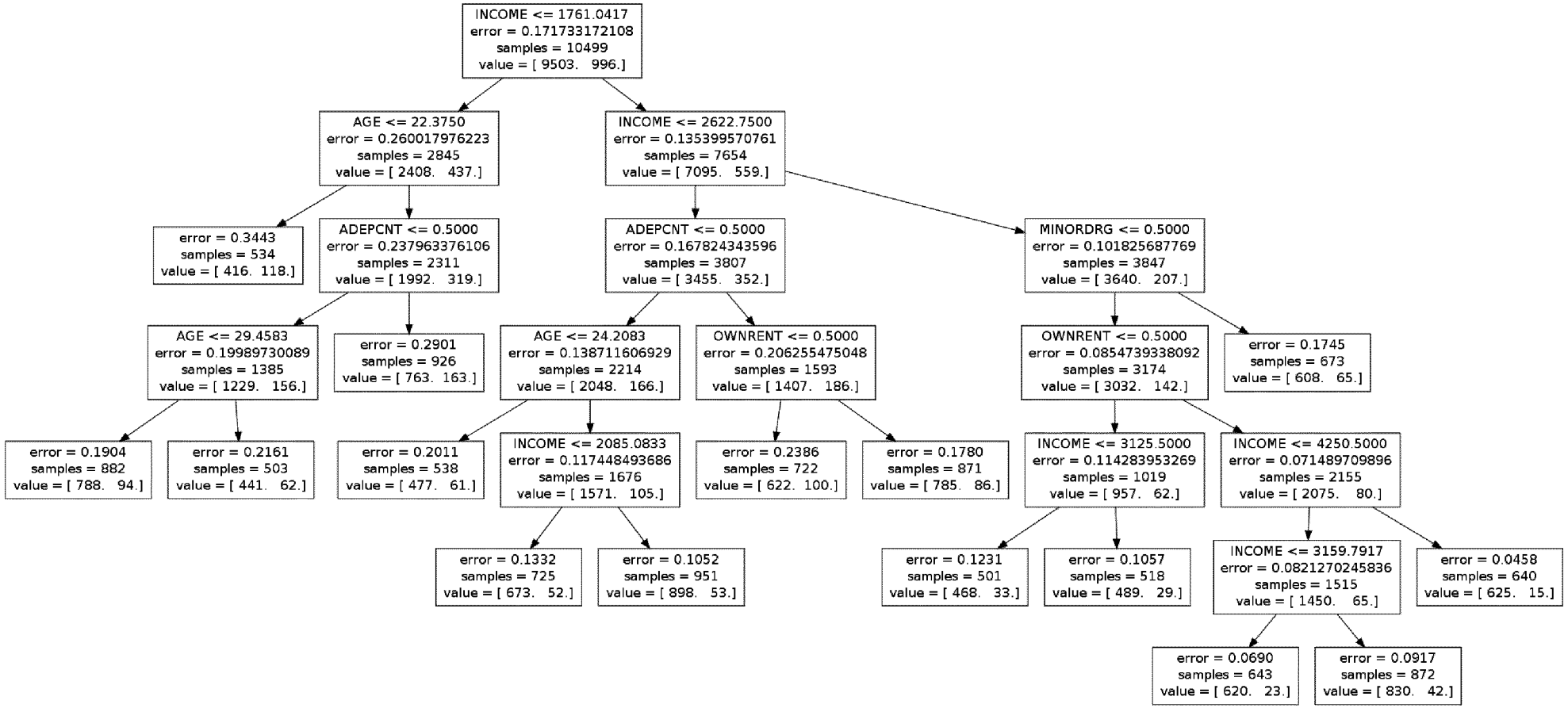

Объект
x>5
x>3
y>5
x>-1
y>3
y>3
12
x>-10
8
8
14
15
11
11
0
Номер дискретной траектории

Фиг. 2

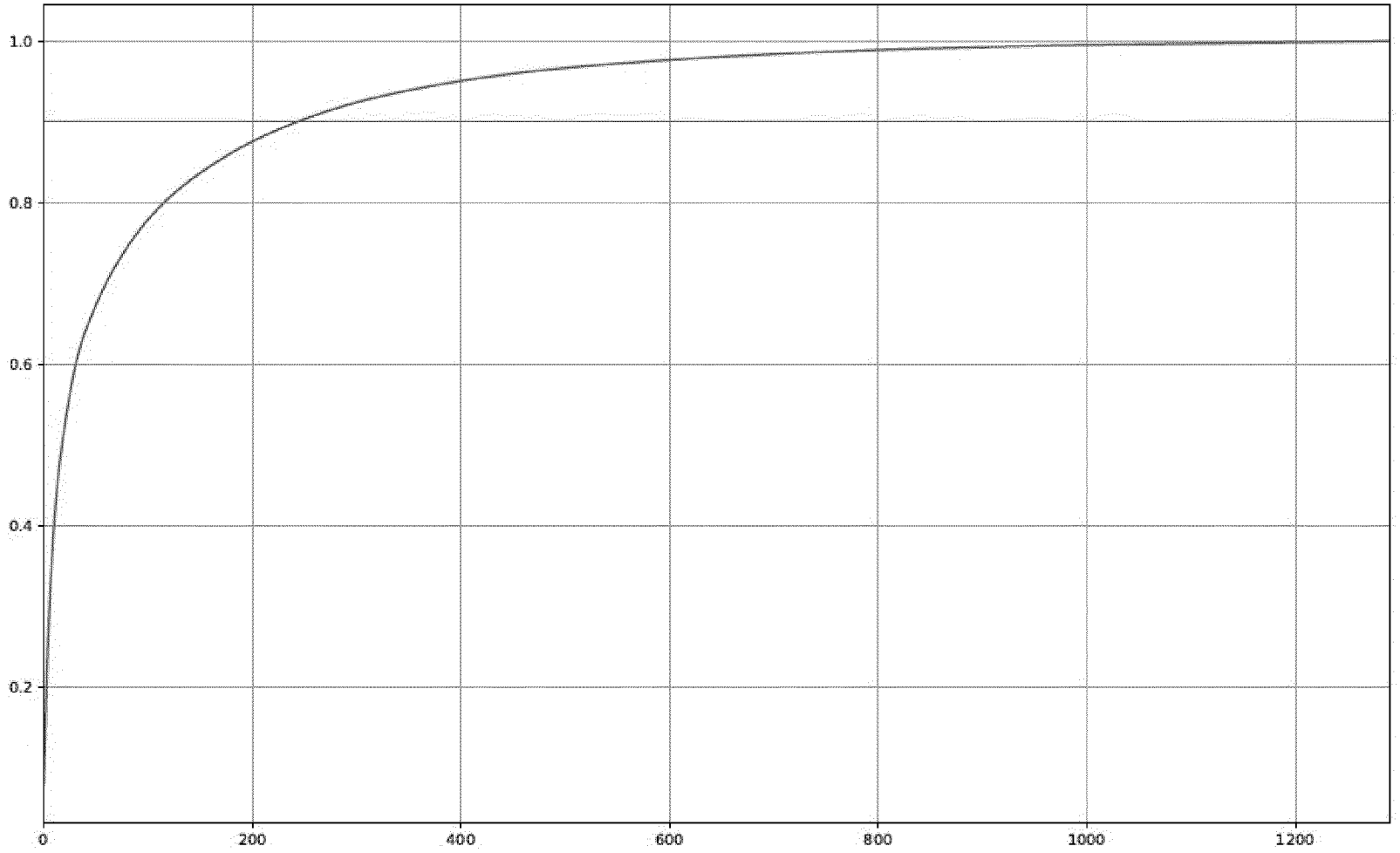
Фиг. 3


Объект
Предсказание
Деревья решений
Дискретные траектории

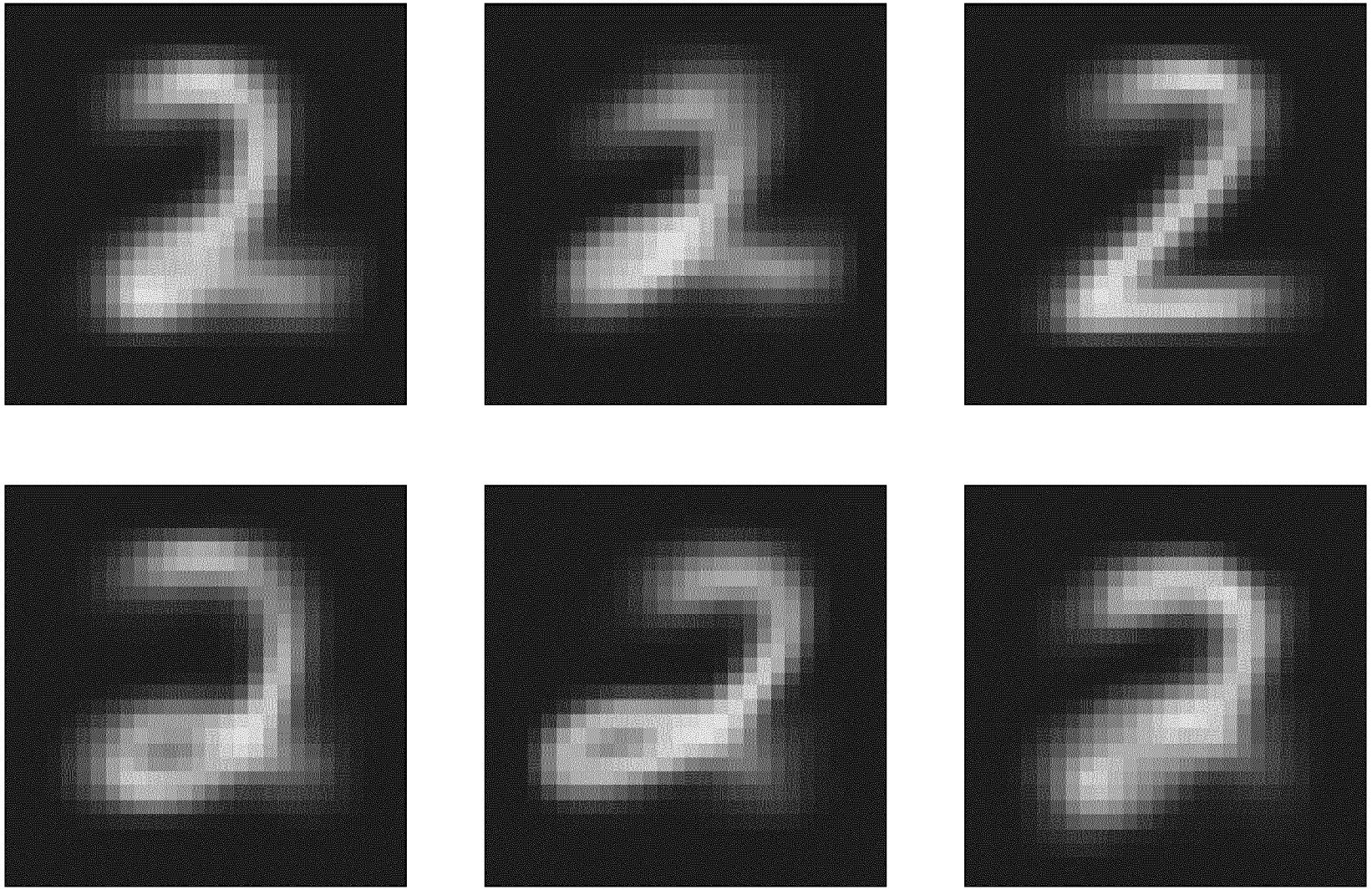
Фиг. 4


INCOME <= 1761.0417
error = 0.171733172108
samples = 10499
value = [9503. 996.]
AGE <= 22.3750
error = 0.260017976223
samples = 2845
value = [2408. 437.]
INCOME <= 2622.7500
error = 0.135399570761
samples = 7654
value = [7095. 559.]
error = 0.3443
samples = 534
value = [416. 118.]
ADEPCNT <= 0.5000
error = 0.237963376106
samples = 2311
value = [1992. 319.]
ADEPCNT <= 0.5000
error = 0.167824343596
samples = 3807
value = [3455. 352.]
MINORDRG <= 0.5000
error = 0.101825687769
samples = 3847
value = [3640. 207.]
AGE <= 29.4583
error = 0.19989730089
samples = 1385
value = [1229. 156.]
error = 0.2901
samples = 926
value = [763. 163.]
AGE <= 24.2083
error = 0.138711606929
samples = 2214
value = [2048. 166.]
OWNRENT <= 0.5000
error = 0.206255475048
samples = 1593
value = [1407. 186.]
OWNRENT <= 0.5000
error = 0.0854739338092
samples = 3174
value = [3032. 142.]
error = 0.1745
samples = 673
value = [608. 65.]
error = 0.1904
samples = 882
value = [788. 94.]
error = 0.2161
samples = 503
value = [441. 62.]
error = 0.2011
samples = 538
value = [477. 61.]
INCOME <= 2085.0833
error = 0.117448493686
samples = 1676
value = [1571. 105.]
error = 0.2386
samples = 722
value = [622. 100.]
error = 0.1780
samples = 871
value = [785. 86.]
INCOME <= 3125.5000
error = 0.114283953269
samples = 1019
value = [957. 62.]
INCOME <= 4250.5000
error = 0.071489709896
samples = 2155
value = [2075. 80.]
error = 0.1332
samples = 725
value = [673. 52.]
error = 0.1052
samples = 951
value = [898. 53.]
error = 0.1231
samples = 501
value = [468. 33.]
error = 0.1057
samples = 518
value = [489. 29.]
INCOME <= 3159.7917
error = 0.0821270245836
samples = 1515
value = [1450. 65.]
error = 0.0458
samples = 640
value = [625. 15.]
error = 0.0690
samples = 643
value = [620. 23.]
error = 0.0917
samples = 872
value = [830. 42.]

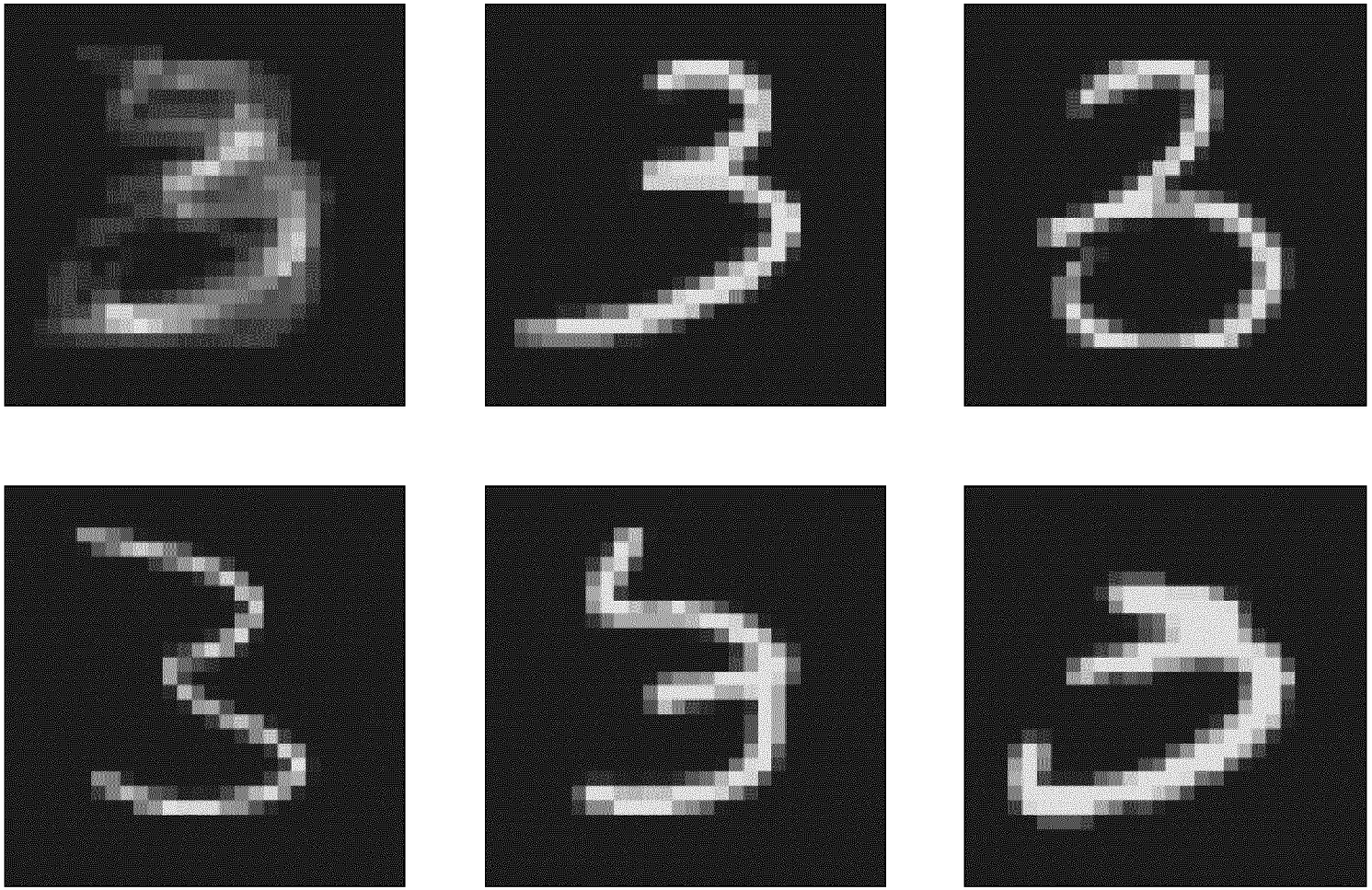
Фиг. 5



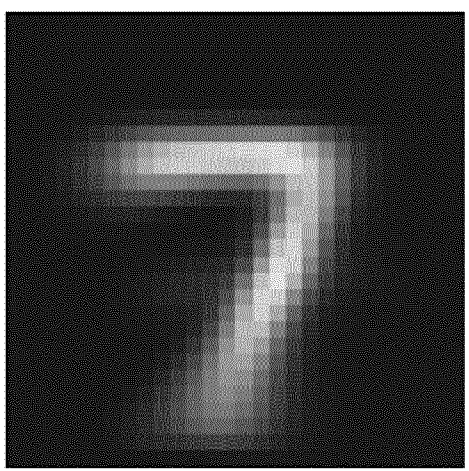
Фиг. 6



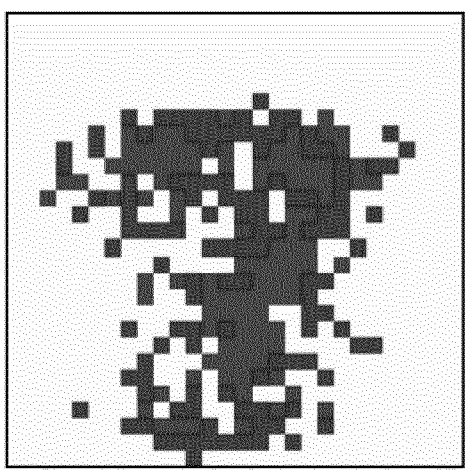
Фиг. 7



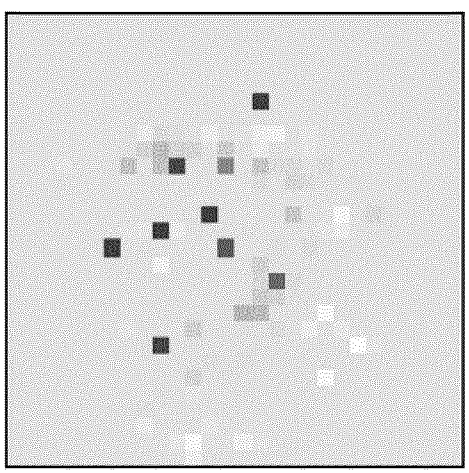
Фиг. 8



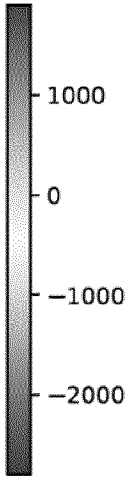
Среднее изображение траектории



Двоичные правила траекторий



Взвешенные правила траекторий



ЕВРАЗИЙСКОЕ ПАТЕНТНОЕ ВЕДОМСТВО

ОТЧЕТ О ПАТЕНТНОМ ПОИСКЕ

(статья 15(3) ЕАПК и правило 42 Патентной инструкции к ЕАПК)

Номер евразийской заявки: 201891425

Дата подачи: 13/07/2018 | Дата испрашиваемого приоритета:

Название изобретения: СПОСОБ ИНТЕРПРЕТАЦИИ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

Заявитель: ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК)

☐ **Некоторые пункты формулы не подлежат поиску (см. раздел I дополнительного листа).**

☐ **Единство изобретения не соблюдено (см. раздел II дополнительного листа)**

А. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ:

G06N 3/02 *(2006.01)*
G06N 3/04 *(2006.01)*

Б. ОБЛАСТЬ ПОИСКА:

Минимум просмотренной документации (система классификации и индексы МПК):
G06N 3/00-3/04

Другая проверенная документация в той мере, в какой она включена в область поиска:

В. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ

Категория*	Ссылки на документы с указанием, где это возможно, релевантных частей	Относится к пункту №
X	Gregoire Montavon et al, «Methods for interpreting and understanding deep neural networks», Digital Signal Processing, Volume 73, February 2018, с. 1-15, размещено в Интернет: https://www.sciencedirect.com/science/article/pii/S1051200417302385 Название, реферат, разделы 1-9	1-12
A	Vasile Palade et al, «Interpretation of Trained Neural Networks by Rule Extraction», Springer-Verlag Berlin Heidelberg, 2001	1-12
A	WO2005/024718 A1, (SEMEION), 17.03.2005	1-12
A	RU2573766 C1, (Открытое акционерное общество «Центральное научно-производственное объединение «Ленинец»), 27.01.2016	1-12

☐ последующие документы указаны в продолжении графы В ☐ данные о патентах-аналогах указаны в приложении

* Особые категории ссылочных документов:
"A" документ, определяющий общий уровень техники
"E" более ранний документ, но опубликованный на дату подачи евразийской заявки или после нее
"O" документ, относящийся к устному раскрытию, экспонированию и т.д.
"P" документ, опубликованный до даты подачи евразийской заявки, но после даты испрашиваемого приоритета
"D" документ, приведенный в евразийской заявке
"T" более поздний документ, опубликованный после даты приоритета и приведенный для понимания изобретения
"X" документ, имеющий наиболее близкое отношение к предмету поиска, порочащий новизну или изобретательский уровень, взятый в отдельности
"Y" документ, имеющий наиболее близкое отношение к предмету поиска, порочащий изобретательский уровень в сочетании с другими документами той же категории
"&" документ, являющийся патентом-аналогом
"L" документ, приведенный в других целях

Дата действительного завершения патентного поиска: 16/05/2019

Уполномоченное лицо:

Ведущий эксперт
Отдела механики, физики и электротехники

Д.А. Гудилин

Телефон: +7(495)411-61-61*310