(19) Евразийское патентное ведомство

(11) 047784 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2024.09.10

(21) Номер заявки
202393317

(22) Дата подачи заявки
2023.12.19

(51) Int. Cl. ***G06N 3/08*** (2023.01)
G06N 20/20 (2019.01)
G06N 3/02 (2006.01)
G06F 18/211 (2023.01)

(54) СПОСОБ И СИСТЕМА АВТОМАТИЧЕСКОГО СОЗДАНИЯ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ

(31) **2023115731**
(32) **2023.06.15**
(33) **RU**
(43) **2024.09.09**

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

(72) Изобретатель:
Ахундов Зохраб Сейфуллаевич (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-A1-20200184380
US-B2-10417528
US-A1-2021304056
US-B2-10592386

(57) Заявленное техническое решение в общем относится к области машинного обучения, а в частности - к способу и системе автоматического создания модели машинного обучения для решения задач классификации, направленных на проблемы реального мира. Техническим результатом от реализации заявленного способа является автоматическое создание модели машинного обучения с низкой ресурсоемкостью и с сохранением высокой обобщающей способности и, как следствие, высокого качества прогнозирования целевого события. Указанный технический результат достигается благодаря осуществлению способа автоматического создания модели машинного обучения для решения задачи классификации, выполняющейся по меньшей мере одним вычислительным устройством, содержащего этапы, на которых задают значения параметров; создают скрипты; получают набор данных для обучения; определяют архитектуру модели и список признаков модели; генерируют модель машинного обучения; рассчитывают метрики классификации модели машинного обучения.

047784 B1

047784 B1

Область техники

Заявленное техническое решение в общем относится к области машинного обучения, а в частности - к способу и системе автоматического создания модели машинного обучения для решения задач классификации, направленных на проблемы реального мира.

Уровень техники

В настоящее время создание и внедрение моделей машинного обучения в различных отраслях жизнедеятельности стало неотъемлемой частью этих отраслей. Алгоритмы машинного обучения применяются для распознавания голосов, изображений, обработки текстов на естественном языке и многого другого. Внедрение моделей машинного обучения совершило настоящую революцию во многих отраслях человеческой деятельности.

Существует множество способов создания и обучения моделей машинного обучения для выполнения поставленных задач, например прогнозирования наступления целевого события и т.д. Общий принцип процесса обучения моделей заключается в предварительной обработке данных для подготовки к процессу обучения, определении характерных особенностей сформированного обучающего набора данных, конструировании новых признаков из указанного набора, оптимизации гиперпараметров модели машинного обучения и самого процесса обучения сформированной модели при оптимальных параметрах.

Однако указанный принцип сильно подвержен влиянию человеческого фактора, так как указанные этапы процесса обучения выполняет специалист в данной области техники. Кроме того, ручное обучение моделей является длительным процессом и, при определенных ситуациях, даже неэффективным.

Из уровня техники также известны подходы, основанные на принципе автоматического машинного обучения (AutoML) - процесса автоматизации сквозного процесса применения машинного обучения к задачам реального мира. Так, из уровня техники известно решение, раскрытое в заявке № US 2020159690 A1 (SAP SE [DE]), опубл. 21.05.2020. Указанное решение, в частности, раскрывает систему оценки показателей с использованием подхода классификации с автоматическим машинным обучением, выполненную с возможностью получения и предобработки набора данных, формирования файла конфигурации для создания модели машинного обучения, выбора требуемой модели из шаблонных моделей машинного обучения путем перебора указанных шаблонов под тип данных и подсчета их точности.

Недостатком указанного решения является низкая скорость генерации модели машинного обучения ввиду длительной процедуры перебора шаблонных моделей под решаемую задачу классификации. Кроме того, указанный процесс определения модели и выбора подходящей модели из представленных шаблонов также увеличивает ресурсоемкость модели машинного обучения, так как выбирается лучшая из представленных моделей, которая, очевидно, может обладать крайне высокой ресурсоемкостью (количество параметров), что, как следствие, приводит к недостаточному качеству (точностью) для получения корректных результатов.

Общим недостатком существующих решений является отсутствие эффективного способа автоматического создания модели машинного обучения для решения задач реального мира, обеспечивающего низкую ресурсоемкость и высокую скорость генерации модели при сохранении высокого качества (точности) для получения корректных результатов.

Раскрытие изобретения

В заявленном техническом решении предлагается новый подход к автоматическому созданию модели машинного обучения для решения задачи классификации. В данном решении используется автоматическая генерация модели машинного обучения на основе автоматического подбора оптимальной архитектуры и признакового пространства, использующихся для обучения модели. Таким образом, решается техническая проблема уменьшения ресурсоемкости и увеличения скорости генерации модели при сохранении высокой точности модели. Основным техническим результатом, проявляющимся при решении вышеуказанной проблемы, является повышение точности прогнозирования целевого события за счет повышения точности генерируемой модели машинного обучения. Дополнительным техническим результатом, проявляющимся при решении вышеуказанной проблемы, является уменьшение ресурсоемкости и увеличение скорости генерации модели.

Указанные технические результаты достигаются благодаря осуществлению способа автоматического создания модели машинного обучения для решения задачи классификации, выполняющегося по меньшей мере одним вычислительным устройством содержащего этапы, на которых:

a) получают первый набор данных для обучения в соответствии с решаемой задачей классификации, причем указанный набор данных содержит по меньшей мере период сбора данных и набор данных признакового пространства, включающий категориальные и числовые признаки для решаемой задачи классификации;

b) получают второй набор данных для обучения, содержащий данные о наступлении целевого события решаемой задачи классификации;

c) формируют репрезентативную выборку для обучения модели на основе первого и второго наборов данных;

d) определяют архитектуру модели машинного обучения в ходе которой:

i) задают параметр качества модели;

ii) получают начальное количество деревьев модели;

iii) выполняют обучение модели на данных, полученных на этапе a) и b), при заданном начальном количестве деревьев и сравнивают параметр качества модели с заданным параметром качества модели;

iv) повторяют шаг iii), рекуррентно уменьшая количество деревьев, до критерия останова;

v. формируют файл с архитектурой модели на основе количества деревьев, полученных на шаге iv);

e) определяют количество существенных признаков для модели машинного обучения, причем в ходе определения указанных признаков:

i) получают архитектуру модели машинного обучения, определенную на этапе d) и параметр качества модели;

ii) получают набор данных признакового пространства;

iii) исключают по меньшей мере один признак из набора и выполняют обучение модели;

iv) определяют качество модели и сравнивают его с заданным параметром качества модели;

v) повторяют шаги ii)-iii) до критерия останова;

vi) формируют набор существенных признаков на основе данных, полученных на шаге v);

f) генерируют модель машинного обучения на основе данных, полученных на этапах d) и e), и выполняют обучение указанной модели с помощью репрезентативной выборки, полученной на этапе c);

g) выполняют калибровку модели, полученной на этапе f);

h) осуществляют обработку данных с помощью модели машинного обучения, полученной на этапе f), в ходе которой получают вероятность наступления целевого события для решаемой задачи классификации;

i) выполняют действие, соответствующее результатам, полученным на этапе h).

В еще одном частном варианте реализации способа первый набор данных является репрезентативным набором данных.

В другом частном варианте реализации способа категориальные признаки представляют собой по меньшей мере следующие признаки: пол, семейное положение, город проживания.

В другом частном варианте реализации способа числовые признаки представляют собой по меньшей мере следующие признаки: возраст, количество предыдущих обращений.

В другом частном варианте реализации способа критерий останова определения архитектуры модели машинного обучения представляет собой по меньшей мере параметр, характеризующий минимальное количество деревьев модели машинного обучения, при сохранении параметра качества модели машинного обучения выше порогового параметра качества.

В другом частном варианте реализации способа критерий останова определения существенных признаков представляет собой по меньшей мере пороговый параметр качестве модели.

В другом частном варианте реализации способа решаемая задача классификации представляет собой по меньшей мере вероятность звонка клиента в центр обработки звонков.

В другом частном варианте реализации способа действие, соответствующее решаемой задаче классификации, представляет собой распределение нагрузки между оборудованием, обслуживающим центр обработки звонков.

Кроме того, заявленные технические результаты достигаются за счет системы автоматического создания модели машинного обучения для решения задачи классификации, содержащей:

по меньшей мере один процессор;

по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа автоматического создания модели машинного обучения для решения задачи классификации.

Краткое описание чертежей

Признаки и преимущества настоящего изобретения станут очевидными из приводимого ниже подробного описания изобретения и прилагаемых чертежей.

Фиг. 1 иллюстрирует блок-схему выполнения заявленного способа.

Фиг. 2 иллюстрирует пример общего вида вычислительной системы, которая обеспечивает реализацию заявленного решения.

Осуществление изобретения

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

Модель в машинном обучении (МО) - совокупность методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач.

Задача классификации в машинном обучении - это задача отнесения объекта к одному из заранее определенных классов на основании его формализованных признаков. Градиентный бустинг - это метод машинного обучения, используемый, среди прочего, в задачах регрессии и классификации. Он дает модель прогнозирования в виде совокупности слабых моделей прогнозирования, которые обычно пред-

ставляют собой деревья решений.

Коэффициент Джини (Gini coefficient) - метрика качества, которая часто используется при оценке предсказательных моделей в задачах бинарной классификации. Заявленное техническое решение предлагает новый подход, обеспечивающий понижение ресурсоемкости и повышение скорости генерации моделей машинного обучения для решения задач классификации без существенного уменьшения обобщающей способности, направленных на проблемы реального мира, например задач прогнозирования наступления целевого события. Кроме того, заявленное решение предлагает подход, обеспечивающий возможность автоматической генерации моделей машинного обучения, что, соответственно, исключает влияние человеческого фактора на процесс обучения. Еще одним преимуществом, проявляющимся при использовании заявленного изобретения, является снижение вычислительных мощностей, требуемых для создания и обучения моделей за счет эффективного процесса настройки архитектуры и признаков, используемых в моделях, решающих поставленные задачи. Данное техническое решение может быть реализовано на компьютере, в виде автоматической системы, содержащей интерфейсы ввода/вывода, или машиночитаемого носителя, содержащего инструкции для выполнения вышеупомянутого способа. Техническое решение может быть реализовано в виде распределенной компьютерной системы.

В данном решении под системой подразумевается компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, чётко определённую последовательность вычислительных операций (действий, инструкций). Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы). Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройства хранения данных, например таких устройств, как оперативно запоминающие устройства (ОЗУ) и/или постоянные запоминающие устройства (ПЗУ). В качестве ПЗУ могут выступать, но не ограничиваясь, жесткие диски (HDD), флэш-память, твердотельные накопители (SSD), оптические носители данных (CD, DVD, BD, MD и т.п.) и др.

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

Термин "инструкции", используемый в этом документе, может относиться, в общем, к программным инструкциям или программным командам, которые написаны на заданном языке программирования для осуществления конкретной функции, такой как, например, получение артефактов программно-аппаратного решения, формирование цифрового стандарта программно-аппаратного решения, формирование результатов проверки программно-аппаратного решения, анализ данных и т.п. Инструкции могут быть осуществлены множеством способов, включающих в себя, например, объектно-ориентированные методы. Например, инструкции могут быть реализованы, посредством языка программирования C++, Java, Python, различных библиотек (например, MFC; Microsoft Foundation Classes) и т.д. Инструкции, осуществляющие процессы, описанные в этом решении, могут передаваться как по проводным, так и по беспроводным каналам передачи данных, например Wi-Fi, Bluetooth, USB, WLAN, LAN и т.п.

На фиг. 1 представлена блок-схема способа 100 автоматического создания модели машинного обучения для решения задачи классификации, который раскрыт поэтапно более подробно ниже. Указанный способ 100 заключается в выполнении этапов, направленных на обработку различных цифровых данных. Обработка, как правило, выполняется с помощью системы, которая может представлять, например, сервер, компьютер, мобильное устройство, вычислительное устройство и т.д. Более подробно элементы системы раскрыты на фиг. 2.

На этапе 110 получают первый набор данных для обучения в соответствии с решаемой задачей классификации, причем указанный набор данных содержит по меньшей мере период сбора данных и набор данных признакового пространства, включающий категориальные и числовые признаки для решаемой задачи классификации. На указанном этапе 110 система 200 получает входные данные для решения задачи классификации. Так, процесс получения данных может быть реализован посредством обращения системы 200 к хранилищам данных, например внутренним базам данных, которые содержат данные о целевом событии. В качестве хранилищ данных могут выступать как определенные клиенты (все данные хранятся в CSV-файлах), так и SQL-сервера и т.д. Обращение к указанным хранилищам может осуществляться посредством сети передачи данных, например сети Интернет, ЛВС и т.д. В одном частном варианте осуществления данные могут быть загружены в систему с помощью интерфейса ввода/вывода, например, пользователем системы 200.

Входным набором данных является набор исторических данных о целевом событии, соответствующий решаемой задачей классификации, достаточный для создания и обучения модели машинного обучения. Так, в одном частном варианте осуществления пользователь системы 200 выбирает тип задачи, которую необходимо решить с помощью заявленного способа. В соответствии с выбранным типом задачи система 200 получает соответствующий набор исторических данных, связанных с этой задачей. Например, задачами классификации, связанными с проблемами реального мира, могут являться такие задачи, как определение вероятности звонка клиента в центр обработки звонков (call center) в следующем меся-

це, вероятность обращения клиента в отделение банка в определенный период, вероятность обращения человека за медицинской помощью, прогнозирование загруженности транспортной системы города и т.д. Каждая из приведенных задач влияет на инфраструктуру, связанную с организацией, обслуживающей представленную область жизнедеятельности. Так, если речь идет о вероятности звонка клиента в центр обработки звонков, то проблематикой в данной сфере будет являться нагрузка на оборудование, т.е. отказоустойчивость инфраструктуры, выход из строя или неполадки в которой, как следствие, могут привести к снижению безопасности (невозможности предотвращения угроз). Схожая ситуация может возникать и в медицинских учреждениях, где прогнозирование человеко-потока напрямую влияет на скорость оказания медицинской помощи и необходимого количества медицинского персонала для своевременного оказания такой помощи. Как будет очевидно специалисту в данной области техники, задача классификации в данном решении является общим термином, отражающим процесс определения вероятности наступления целевого события, на основе которого совершается целевое действие, соответственно, под задачей классификации в данном решении понимается задача определения наступления целевого события на основе его исторических данных.

Стоит отметить, что в частных случаях реализации заявленного изобретения задача классификации сводится к бинарной задаче классификации, т.е. событие либо наступило, либо нет. Кроме того, важным критерием для решения задачи является наличие исторических данных. Например, нельзя решить задачу предсказания вероятности звонка клиента в контактный центр в следующем месяце, не имея исторических данных о его звонках.

Так, например, на этапе 110 входной набор данных может быть получен посредством подключения к базе данных, например SQL-серверу. Так, для осуществления данного процесса пользователь, в GUI системы, например системы 200, может задать следующий набор параметров: параметр конфигурации подключения к sql-серверу; параметр уникального идентификатора модели; параметр даты обучения модели; параметр даты валидации модели; параметр списка признаков для исключения из обучения модели. Указанные параметры, по сути, обеспечивают возможность создания и выделения вычислительным устройством памяти и мощности под реализацию дальнейших этапов способа 200.

Таким образом, на этапе 110 система 200 получает первый набор данных, содержащий период сбора данных для решаемой задачи классификации и набор данных признакового пространства, включающий категориальные и числовые признаки для решаемой задачи классификации. Период сбора данных может зависеть как непосредственно от необходимого горизонта определения целевого события, так и от количества исторических данных, хранящихся в хранилище данных. Так, период сбора данных может представлять собой 180 дней, 365 дней и т.д. Кроме того, в еще одном частном варианте осуществления период сбора данных может зависеть от количества исторических данных, например, минимально достаточного набора исторических данных (репрезентативного набора данных). Так, репрезентативным набором данных является такая выборка, в которой все основные признаки генеральной совокупности, из которой извлечена данная выборка, представлены приблизительно в той же пропорции или с той же частотой, с которой данный признак выступает в этой генеральной совокупности. Указанный критерий может быть проверен системой 200, например, на этапе выбора хранилища данных, например, по количеству хранящихся исторических данных (10000 записей и т.д.). Данные о количестве записей каждого раздела базы данных могут храниться и определяться из файла таблицы данных, например из метаданных.

Под признаковым пространством в данном решении понимается совокупность признаков, которые с достаточной полнотой отражают свойства объекта. Так, признаковое пространство в одном частном варианте осуществления включает в себя наборы категориальных признаков объекта и числовых признаков объекта. Категориальные признаки представляют собой по меньшей мере пол, город проживания, город рождения, перенесенные заболевания, вид транспортного средства и т.д. Числовые признаки представляют собой по меньшей мере возраст, количество обращений в организацию за выбранный период, количество транспортных средств и т.д. Т.е. признаковое пространство характеризует максимально полную и объективную информацию о свойствах исследуемого объекта. Очевидно, что в зависимости от решаемой задачи классификации, признаковое пространство объекта будет отличаться. Так, например, для прогнозирования вероятности звонка в контактный центр, первый набор данных для обучения будет содержать следующие категориальные и числовые признаки: количество звонков за последние три месяца, пол, возраст, место рождения, место проживание.

Таким образом, на этапе 110 получают первый набор данных для обучения в соответствии с решаемой задачей классификации, причем указанный набор данных содержит по меньшей мере период сбора данных и набор данных признакового пространства, включающий категориальные и числовые признаки для решаемой задачи классификации.

Далее, на этапе 120 получают второй набор данных для обучения, содержащий данные о наступлении целевого события решаемой задачи классификации. На указанном этапе 120 система 200 получает данные о наступлении или не наступлении целевого события. Так, второй набор данных для обучения формируется также на основе исторических данных и содержит в себе данные о том, произошло ли событие. Указанный набор данный может быть получен так же, как и первый набор данных, посредством

сети передачи данных, например посредством сети Интернет. Данные о наступлении целевого события могут храниться как в отдельной базе данных, так и столбце базы данных, содержащем данные для первого набора обучающих данных, сформированного на этапе 110. Так, для прогнозирования вероятности звонка в контактный центр данные о факте наступления или не наступления целевого события являются данными совершенного или не совершенного звонка определенным клиентом и/или пациентом.

Так, в одном частном варианте осуществления, на этапе 120, для формирования второго набора данных может использоваться, например, уникальный идентификатор модели; уникальный идентификатор целевого события; дата сбора данных; факт наступления целевого события;

Кроме того, в еще одном частном варианте осуществления дополнительно может быть создан третий набор данных, содержащий данные о признаковом пространстве целевого события решаемой задачи классификации. Указанный набор предназначен для ускорения процесса обучения модели машинного обучения. Так, указанный набор может быть сформирован с помощью параметра фильтра репрезентативности, т.е. фильтра, отражающего релевантные признаки для обучения модели. Так, в еще одном частном варианте осуществления третий набор данных может содержать набор данных признакового пространства целевого события, включающий категориальные и числовые признаки для решаемой задачи классификации.

В еще одном частном варианте осуществления данными о наступлении целевого события могут быть данные о выходе из строя оборудования при разном количестве звонков. Авторы настоящего изобретения отмечают, хотя и в качестве примера приводится проблема отказоустойчивости контактного центра, что представленное изобретение также может быть применено и в других сферах жизнедеятельности, например в системе управления климатом помещений. Так, данными о наступлении целевого события в таком случае будут данные о включении или не включении системы отопления при достижении определенной температуры.

На этапе 130 формируют репрезентативную выборку для обучения модели на основе первого и второго наборов данных.

На этапе 130 выполняется обработка первого и второго наборов данных для получения репрезентативной выборки для обучения модели машинного обучения.

Репрезентативной является выборка, чей Target Rate соответствует Target Rate Генеральной совокупности.

Так, из второго набора данных берутся данные о факте наступления целевого события. Указанные данные сопоставляются с данными из первого набора. Указанный процесс может быть реализован, например, посредством процедуры сопоставления и поиска по таблицам данных, например инструмент поиска в текстовых файлах. После нахождения релевантных (репрезентативных) значений указанные значения сопоставляются между собой и сохраняются в новый файл данных.

В еще одном частном варианте осуществления репрезентативная выборка для обучения модели может быть заранее сформирована и отправлена в систему 200, например, пользователем системы 200.

Таким образом, на указанном этапе 130 формируют репрезентативную выборку для обучения модели на основе первого и второго наборов данных.

Далее, способ 100 переходит к этапу 140.

На этапе 140 определяют архитектуру модели машинного обучения.

Определение архитектуры модели выполняется в несколько подэтапов, в ходе которых задают параметр качества модели; получают начальное количество деревьев модели; выполняют обучение модели на данных, полученных на этапе 110 и 120, при заданном начальном количестве деревьев и сравнивают параметр качества модели с заданным параметром качества модели; повторяют этап обучения модели, рекуррентно уменьшая количество деревьев, до критерия останова; формируют файл с архитектурой модели на основе количества деревьев, полученных после остановки алгоритма.

Рассмотрим более подробно указанный процесс определения архитектуры модели машинного обучения.

Так, под определением архитектуры модели машинного обучения понимается автоматическая настройка и подбор количества решающих деревьев модели машинного обучения, которые обеспечивают максимальную обобщающую способность прогнозирования целевого события. Для этого на указанном этапе 140 задается качество модели. Качество модели в одном частном варианте осуществления определяется параметром - коэффициентом Джини. Указанный коэффициент отражает долю уменьшения качества итоговой модели по сравнению с качеством лучшей модели, т.е. максимально возможным значением качества модели. Указанный параметр в одном частном варианте осуществления может быть задан вручную пользователем системы. В еще одном частном варианте осуществления коэффициент Джини может быть заранее установлен в системе 200 и может быть равен, например, 99.5%. Под качеством модели в данном решении следует понимать возможность модели правильно предсказывать заданный класс в решении задачи классификации.

Определение архитектуры модели машинного обучения выполняется автоматически системой 200, т.е. пользователь системы не участвует в процессе. Модель машинного обучения в одном частном варианте осуществления представляет собой модель градиентного бустинга, например модель LightGBM и

т.д. Более подробно, определение архитектуры модели машинного обучения состоит в настройке ее параметров. Так, большинство алгоритмов машинного обучения включают большое количество параметров, в том числе параметры регуляризации, параметры стохастического градиентного спуска, а также другие параметры, такие как максимальное количество деревьев. Параметры модели оказывают значительное влияние на точность прогнозирования результирующих моделей без четких значений, применимых к различным наборам данных. Традиционным способом поиска параметров является ручная настройка. Точность генерируемой модели машинного обучения зависит от минимальной ошибки прогнозирования. Так, в одном частном варианте осуществления параметром модели машинного обучения является параметр количества деревьев.

Для определения указанного параметра, а следовательно, для определения архитектуры модели, система 200 на следующем шаге устанавливает начальное количество деревьев, например, равное 5000 деревьям. Специалисту в данной области техники будет очевидно, что начальный параметр количества деревьев может быть установлен любым целым, положительным значением.

После этого выполняется обучение представленной конфигурации модели, в ходе которого данные, полученные на этапе 110, и данные, полученные на этапе 120, обрабатываются моделью для получения результатов прогнозирования модели. Так, в качестве модели может быть использована модель LIGhtGBm (более подробно модель описана по ссылке, например: https://lightgbm.readthedocs.io/en/latest/).

После обучения данной конфигурации модели, содержащей заданное количество деревьев, сравнивается качество указанной модели с пороговым параметром качества. Как указывалось выше, качество модели может сравниваться по коэффициенту Джини. Если параметр качества удовлетворяет заданному, то алгоритм завершается. Если параметр качества ниже заданного, то система 200 рекуррентно уменьшает начальное количество деревьев и повторяет подэтап обучения с новым количеством деревьев. Указанный процесс продолжается до критерия останова. В одном частном варианте осуществления критерий останова представляет собой по меньшей мере параметр, характеризующий минимальное количество деревьев модели машинного обучения, при сохранении параметра качества модели машинного обучения выше порогового параметра качества. Стоит отметить, что критерий останова на указанном этапе 140 упоминается в контексте критерия останова определения архитектуры модели машинного обучения и также может быть назван первым критерием останова.

После достижения критерия останова (первого критерия останова) файл конфигурации модели машинного обучения сохраняется в память системы 200. Итоговый файл конфигурации модели содержит настроенную архитектуру модели машинного обучения под решаемую задачу классификации, например под прогнозирование вероятности звонка в контактный центр. При этом, в соответствии с критерием останова, алгоритм подбирает наиболее оптимальное количество деревьев при сохранении качества модели. Как было указано выше, количество деревьев также влияет на скорость обучения модели.

Таким образом, на указанном этапе 140 осуществляется определение архитектуры модели машинного обучения, в ходе которой формируется файл конфигурации, содержащий количество решающих деревьев модели.

Далее, на этапе 150 определяют количество существенных признаков для модели машинного обучения.

Так, в обучающих наборах данных содержится множество признаков, однако не все признаки влияют на точность прогнозирования модели машинного обучения, при этом такие признаки могут быть достаточно объемными и, как следствие, приводить к уменьшению скорости обучения модели и повышению требуемых для обучения вычислительных ресурсов. Указанная особенность заявленного решения позволяет повысить скорость обучения за счет автоматического определения существенных признаков из общего набора признакового пространства. Для этого на указанном этапе 150 выполняются следующие шаги алгоритма: получают архитектуру модели машинного обучения, определенную на этапе 140 и параметр качества модели; получают набор данных признакового пространства; исключают по меньшей мере один признак из набора и выполняют обучение модели; определяют качество модели и сравнивают его с пороговым параметром качества модели; повторяют процесс до критерия останова; формируют набор существенных признаков на основе достижения критерия останова. Рассмотрим более подробно указанные шаги.

На первом шаге указанного этапа 150 архитектура модели из файла конфигурации с этапа 140 и пороговый критерий качества (коэффициент Джини) извлекаются системой 200. Так, например, файл конфигурации может быть извлечен из внутренней памяти системы 200. Соответственно, коэффициент Джини также может быть получен из предыдущего этапа 140.

На следующем шаге получают набор данных признакового пространства. Указанный набор, как было указано выше, был получен на этапе 110 и содержит все категориальные и числовые признаки, характеризующие решаемую задачу классификации. Так, например, категориальными признаками могут являться пол, семейное положение, тип занятости, наличие второго гражданства и т.д. Т.е. количество признаков может быть достаточно большим, при этом очевидно, что большинство из этих признаков не будут относится к определенной задаче классификации. Соответственно, числовые признаки также мо-

гут содержать большое количество нерелевантных данных. Для определения существенных признаков осуществляются следующие действия.

После получения набора данных признакового пространства исключают по меньшей мере один признак из набора и выполняют обучение модели.

Далее, определяют качество модели и сравнивают его с пороговым параметром качества модели. Если качество удовлетворяет заданному параметру, то повторяют предыдущий шаг и исключают еще один признак из признакового пространства. Так, при исключении некоторых признаков, качество модели может даже расти, в связи с чем необходимо добиться минимально возможного набора признаков, достаточного для обучения модели машинного обучения с высокой обобщающей способностью. Соответственно, указанная процедура исключения признаков и определения текущего качества модели повторяется до критерия останова. Критерий останова представляет собой по меньшей мере пороговый параметр качестве модели и является критерием останова определения существенных признаков, который также может упоминаться в данном техническом решении как второй критерий останова.

Так, в одном частном варианте осуществления система 200 проводит поочередное исключение признаков из модели, в ходе которого рекуррентно исключается каждый признак из репрезентативной выборки; создается модель с заданным значением первого параметра конечного количества деревьев; выполняется обучение модели; определяется качество модели. Указанные шаги повторяются для каждого признака. После этого исключенные признаки сортируются по убыванию качества модели. Далее, система 200 получает самый первый признак из отсортированного списка и сравнивает качество модели без этого признака с пороговым параметром качества модели. В случае невыполнения критерия останова отбрасывается данный признак. После этого процесс повторяется.

Таким образом, на указанном этапе 150 определяют набор существенных признаков для обучения модели машинного обучения из набора данных признакового пространства.

Далее, на этапе 160 генерируют модель машинного обучения на основе данных, полученных на этапах 140 и 150, и выполняют обучение указанной модели с помощью репрезентативной выборки, полученной на этапе 130.

Итоговая модель машинного обучения состоит из архитектуры, полученной на этапе 140, и существенных признаков, полученных на этапе 150.

Как указывалось выше, модель представляет собой LGBM модель. После формирования итоговой модели выполняется обучение указанной модели. Обучение модели выполняется на наборе данных, полученном на этапе 130.

Так, в одном частном варианте осуществления модель обучалась для определения вероятности звонка клиента в Банк в следующем месяце после скоринга для обеспечения необходимых вычислительных ресурсов, обеспечивающих бесперебойную работу клиентского центра обработки звонков.

Так, для обучения модели использовались активные клиенты (пользователи). Активные клиенты могут быть получены из базы данных звонков, которая хранится в организации. Активным является клиент, совершивший звонок в центр за предопределенный период, например месяц, неделя и т.д. Далее, из указанных клиентов выбирались клиенты (пользователи), которые в следующем периоде, после получения набора данных с активными клиентами, например, в следующем месяце, совершили звонок в Банк. Все остальные клиенты размечались как неактивные и не относились к таргету. Стоит отметить, что формирование обучающей выборки выполнялось в автоматическом режиме, т.е. после получения системой входного набора данных система, посредством, например, алгоритмов поиска и сопоставления данных в базах данных, осуществляла сопоставление активных клиентов и клиентов, совершивших звонок в следующем периоде. Признаковое пространство для обучения состояло из 500 признаков.

На первом этапе выполнялось обучение модели LGBMClassifier() с количеством деревьев, равным 105. Выше порога информативности оказалось 98 признаков.

На втором этапе выполнялось обучение модели LGBMClassifier() с количеством деревьев, равным 1. После итеративного увеличения количества деревьев критерий останова выполнился на 51 дереве, при этом метрика GINI стала равной 60,3%. Выше порога информативности оказалось 18 признаков.

На третьем этапе выполнялось обучение модели LGBMClassifier() с количеством деревьев, равным 51. После итеративного отбрасывания признака из модели метрика GINI стала равной 59,12%, при этом в модели осталось 6 признаков.

На четвертом этапе выполнялось обучение модели LGBMClassifier() с количеством деревьев, равным 1. После итеративного увеличения количества деревьев критерий останова выполнился на 31 дереве, при этом метрика GINI стала равной 59,04%.

В результате в итоговую модель вошли следующие признаки:

1. Количество входящих в Банк звонков за последний месяц.
2. Количество исходящих контактов с клиентом за последний месяц.
3. Количество входов в мобильное приложение за последний месяц.
4. Сумма поступлений на счета клиента за последний месяц.
5. Доход клиента.

6. Суммарный баланс на счетах клиента.

Таким образом, модель использует как информацию по коммуникации клиента с Банком, так и его денежные средства на счетах Банка.

Стоит отметить, что итоговое признаковое пространство, в зависимости от решаемой задачи классификации, будет содержать свой уникальный набор признаков.

Соответственно указанные признаки приведены исключительно для понимания принципа реализации заявленного технического решения и не призваны ограничивать варианты реализации.

На этапе 170, после обучения модели, также выполняют калибровку модели, полученной на этапе 160. Так, например, калибровка итоговой модели может быть осуществлена посредством модели Isotonic Regression.

Рассмотрим более подробно указанный процесс.

В заявленном решении качество модели измеряли с помощью метрики Gini, полученной на датасете для валидации. Датасет для валидации итоговой модели автоматически настраивался по следующему алгоритму: берутся данные из списка признакового пространства итоговой модели из следующего месяца и скорятся с помощью sql-скрипта итоговой модели. Полученный скор клиента джойнится с данными о факте наступления целевого события в следующем месяце после скоринга. Метрики классификации, полученные на данных из месяца обучения и данных, полученных из следующего месяца после обучения, приведены в таблице.

Распределение клиентов в Train и OOT выборке

	Данные для обучения	**Данные для валидации**
Месяц	2022-11-30	2022-12-31
Всего клиентов	9118	9215
Таргет клиенты	659	594
Нетаргет клиенты	8459	8620
Target Rate, доля	7,2%	6,5%
GINI	59,04%	58,76%

Результаты сравнения метрики Gini показывают, что заявленный способ несильно уменьшает метрику в течение времени.

На этапе 180 осуществляют обработку данных с помощью модели машинного обучения, полученной на этапе 160, в ходе которой получают вероятность наступления целевого события для решаемой задачи классификации. Так, в заявленном решении в одном частном варианте осуществления в обученную модель машинного обучения отправляют набор данных, которые необходимо обработать. Так, например, в качестве данных, которые обрабатываются с помощью указанной модели, в одном частном варианте осуществления берутся данные, например, 50000 активных клиентов. Как упоминалось выше, активными клиентами являются пользователи, которые совершали заданное количество обращений в контактный центр, например, хотя бы одно обращение в месяц. Стоит отметить, что тип данных, обрабатываемый сгенерированной моделью, всегда совпадает с обучающими данными. Т.е. если речь идет о прогнозировании нагрузки на контактный центр, то требуется определить количество пользователей. Аналогичные требования выполняются и для других задач, где требуется классификация (например, приемное отделение и т.д.). Далее, для полученных данных, модель выдает параметр вероятности звонка. На основе указанного параметра как раз и выполняются действия, связанные с указанным типом данных. Кроме того, значение вероятностей звонка клиентов также может быть использовано при формировании штата сотрудников, обслуживающих контактный центр.

Таким образом, на этапе 180 осуществляется обработка данных, в ходе которой получают вероятность наступления целевого события для решаемой задачи классификации.

На этапе 190 выполняют действие, соответствующее результатам, полученным на этапе 180.

На указанном этапе 190 результат определения вероятности наступления целевого события решаемой задачи классификации поступает в систему принятия решений. Система принятия решений может представлять собой, например, сервер, компьютер, планшет и т.д. и выполнена с возможностью совершения действий на основе полученного предсказания. Например, такими действиями могут являться упреждающие действия, связанные с недопущением перегрузки инфраструктуры организации. Так, продолжая пример вероятности звонка в контактный центр, система принятия решений выполнена с возможностью вычисления плотности потока звонков, на основе данных, полученных на этапе 180, и сравнения показателей плотности с показателями оборудования, обслуживающего указанные звонки. При превышении показателя плотности выше допустимого значения нагрузка на контактный центр может быть распределена между другими центрами.

Таким образом, в указанных материалах заявки раскрыт способ автоматического создания модели машинного обучения для решения задачи классификации.

На фиг. 2 представлен пример общего вида вычислительной системы 200, которая обеспечивает реализацию заявленного способа или является частью компьютерной системы, например сервером, персо-

нальным компьютером, частью вычислительного кластера, обрабатывающей необходимые данные для осуществления заявленного технического решения.

В общем случае система 200 содержит такие компоненты, как один или более процессоров 201, по меньшей мере одну память 202, средство хранения данных 203, интерфейсы ввода/вывода 204, средство В/В 205, средство сетевого взаимодействия 206, которые объединяются посредством универсальной шины.

Процессор 201 выполняет основные вычислительные операции, необходимые для обработки данных при выполнении способа 100. Процессор 201 исполняет необходимые машиночитаемые команды, содержащиеся в оперативной памяти 202. Память 202, как правило, выполнена в виде ОЗУ и содержит необходимую программную логику, обеспечивающую требуемый функционал.

Средство хранения данных 203 может выполняться в виде HDD, SSD дисков, рейд массива, флэш-памяти, оптических накопителей информации (CD, DVD, MD, Blue-Ray дисков) и т.п. Средства 203 позволяют выполнять долгосрочное хранение различного вида информации, например истории обработки транзакционных запросов (логов), идентификаторов пользователей и т.п.

Для организации работы компонентов системы 200 и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В 204. Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которые могут представлять собой, не ограничиваясь, PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п. Выбор интерфейсов 204 зависит от конкретного исполнения системы 200, которая может быть реализована на базе широкого класса устройств, например персональный компьютер, мейнфрейм, ноутбук, серверный кластер, тонкий клиент, смартфон, сервер и т.п.

В качестве средств В/В данных 205 может использоваться: клавиатура, джойстик, дисплей (сенсорный дисплей), монитор, сенсорный дисплей, тачпад, манипулятор мышь, световое перо, стилус, сенсорная панель, трекбол, динамики, микрофон, средства дополненной реальности, оптические сенсоры, планшет, световые индикаторы, проектор, камера, средства биометрической идентификации (сканер сетчатки глаза, сканер отпечатков пальцев, модуль распознавания голоса) и т.п.

Средства сетевого взаимодействия 206 выбираются из устройств, обеспечивающих сетевой прием и передачу данных, например Ethernet карту, WLAN/Wi-Fi модуль, Bluetooth модуль, BLE модуль, NFC модуль, IrDa, RFID модуль, GSM модем и т.п. С помощью средств 205 обеспечивается организация обмена данными между, например, системой 200, представленной в виде сервера, и хранилищем данных, содержащим первый и второй наборы данных для обучения, по проводному или беспроводному каналу передачи данных, например WAN, PAN, ЛВС (LAN), Интранет, Интернет, WLAN, WMAN или GSM.

Конкретный выбор элементов системы 200 для реализации различных программно-аппаратных архитектурных решений может варьироваться с сохранением обеспечиваемого требуемого функционала.

Представленные материалы раскрывают предпочтительные примеры реализации технического решения и не должны трактоваться как ограничивающие иные, частные примеры его воплощения, не выходящие за пределы испрашиваемой правовой охраны, которые являются очевидными для специалистов соответствующей области техники. Таким образом, объем настоящего технического решения ограничен только объемом прилагаемой формулы.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ автоматического создания модели машинного обучения для решения задачи классификации, выполняющийся по меньшей мере одним вычислительным устройством и содержащий этапы, на которых:

a) получают первый набор данных для обучения в соответствии с решаемой задачей классификации, причем указанный набор данных содержит по меньшей мере период сбора данных и набор данных признакового пространства, включающий категориальные и числовые признаки для решаемой задачи классификации;

b) получают второй набор данных для обучения, содержащий данные о наступлении целевого события решаемой задачи классификации;

c) формируют репрезентативную выборку для обучения модели на основе первого и второго наборов данных;

d) определяют архитектуру модели машинного обучения в ходе которой:

i) задают параметр качества модели;

ii) получают начальное количество деревьев модели;

iii) выполняют обучение модели на данных, полученных на этапе a) и b), при заданном начальном количестве деревьев и сравнивают параметр качества модели с пороговым параметром качества модели;

iv) повторяют шаг iii), рекуррентно уменьшая количество деревьев, до критерия останова;

v) формируют файл с архитектурой модели на основе количества деревьев, полученных на шаге iv);

e) определяют количество существенных признаков для модели машинного обучения, причем в хо-

де определения указанных признаков:

i) получают архитектуру модели машинного обучения, определенную на этапе d), и параметр качества модели;

ii) получают набор данных признакового пространства;

iii) исключают по меньшей мере один признак из набора и выполняют обучение модели;

iv) определяют качество модели и сравнивают его с пороговым параметром качества модели;

v) повторяют шаги ii), iii) до критерия останова;

vi) формируют набор существенных признаков на основе данных, полученных на шаге v);

f) генерируют модель машинного обучения на основе данных, полученных на этапах d) и e), и выполняют обучение указанной модели с помощью репрезентативной выборки, полученной на этапе c);

g) выполняют калибровку модели, полученной на этапе f);

h) осуществляют обработку данных с помощью модели машинного обучения, полученной на этапе f), в ходе которой получают вероятность наступления целевого события для решаемой задачи классификации;

i) выполняют действие, соответствующее результатам, полученным на этапе h).

2. Способ по п.1, характеризующийся тем, что первый набор данных является репрезентативным набором данных.

3. Способ по п.1, характеризующийся тем, что категориальные признаки представляют собой по меньшей мере следующие признаки: пол, семейное положение, город проживания.

4. Способ по п.1, характеризующийся тем, что числовые признаки представляют собой по меньшей мере следующие признаки: возраст, количество обращений пользователя.

5. Способ по п.1, характеризующийся тем, что критерий останова определения архитектуры модели машинного обучения представляет собой по меньшей мере параметр, характеризующий минимальное количество деревьев модели машинного обучения, при сохранении параметра качества модели машинного обучения выше порогового параметра качества.

6. Способ по п.1, характеризующийся тем, что критерий останова определения существенных признаков представляет собой по меньшей мере пороговый параметр качестве модели.

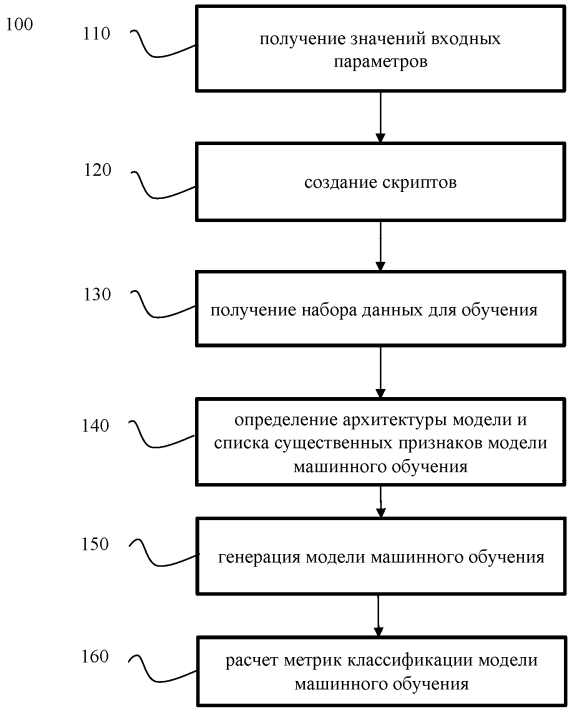
7. Способ по п.1, характеризующийся тем, что решаемая задача классификации представляет собой по меньшей мере вероятность звонка клиента в центр обработки звонков.

8. Способ по п.1, характеризующийся тем, что действие, соответствующее решаемой задаче классификации, представляет собой распределение нагрузки между оборудованием, обслуживающим центр обработки звонков.

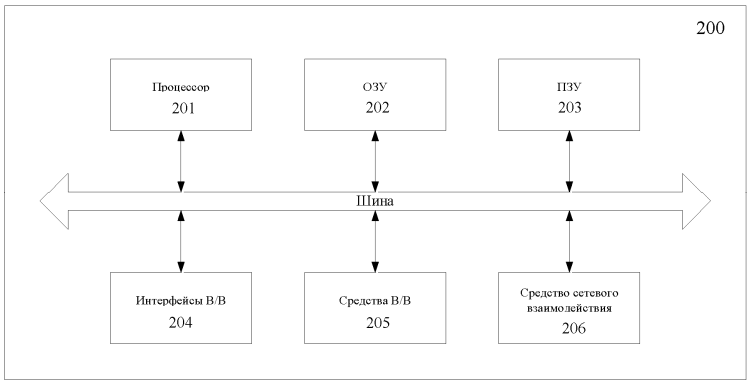
9. Система автоматического создания модели машинного обучения для решения задачи классификации, содержащая:

по меньшей мере один процессор;

по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа по любому из пп.1-8.


100
110
получение значений входных параметров
120
создание скриптов
130
получение набора данных для обучения
140
определение архитектуры модели и списка существенных признаков модели машинного обучения
150
генерация модели машинного обучения
160
расчет метрик классификации модели машинного обучения

Фиг. 1


200
Процессор
201
ОЗУ
202
ПЗУ
203
Шина
Интерфейсы В/В
204
Средства В/В
205
Средство сетевого взаимодействия
206

Фиг. 2