SN Компьютерные науки (2021) 2:160
https://doi.org/10.1007/s42979-021-00592-x

ОБЗОР СТАТЬИ

Машинное обучение: алгоритмы, реальные приложения и исследования Направления

Икбал Х. Саркер1,2

Получено: 27 января 2021 г. / Принято: 12 марта 2021 г. / Опубликовано онлайн: 22 марта 2021 г.

Абстрактный

В нынешнюю эпоху Четвертой промышленной революции (4IR или Industry 4.0) цифровой мир обладает огромным количеством данных, таких как данные Интернета вещей (IoT), данные кибербезопасности, мобильные данные, бизнес-данные, данные социальных сетей, данные о здоровье и т. д. Для разумного анализа этих данных и разработки соответствующих интеллектуальных и автоматизированных приложений ключевым является знание искусственного интеллекта (ИИ), в частности, машинного обучения (МО) . В этой области существуют различные типы алгоритмов машинного обучения, такие как контролируемое, неконтролируемое, полуконтролируемое и обучение с подкреплением. Кроме того, глубокое обучение, которое является частью более широкого семейства методов машинного обучения, может разумно анализировать данные в больших масштабах. В этой статье мы представляем всесторонний взгляд на эти алгоритмы машинного обучения , которые можно применять для повышения интеллекта и возможностей приложения. Таким образом, ключевой вклад этого исследования заключается в объяснении принципов различных методов машинного обучения и их применимости в различных областях реального мира , таких как системы кибербезопасности, умные города, здравоохранение, электронная коммерция, сельское хозяйство и многое другое. Мы также выделяем проблемы и потенциальные направления исследований

на основе нашего исследования. В целом, эта статья призвана служить ориентиром как для академических кругов, так и для профессионалов отрасли, а также для лиц, принимающих решения в различных реальных ситуациях и областях применения, особенно с технической точки зрения.

Ключевые слова Машинное обучение · Глубокое обучение · Искусственный интеллект · Наука о данных · Принятие решений на основе данных · Предиктивная аналитика · Интеллектуальные приложения

Введение

Мы живем в эпоху данных, когда все вокруг нас подключено к источнику данных, и все в нашей жизни записывается в цифровом виде [21, 103]. Например, в современном электронном мире есть множество различных видов данных, таких как данные Интернета вещей (IoT), данные кибербезопасности, данные умных городов, бизнес-данные, данные смартфонов, данные социальных сетей, данные о здоровье, данные COVID-19 и многое другое. Эти данные могут быть структурированными, полуструктурированными или неструктурированными, кратко обсуждается в разделе «Типы реальных данных и методы машинного обучения» , число которых растет с каждым днем.

Извлечение информации из этих данных может быть использовано для создания различных интеллектуальных приложений в соответствующих областях. Например, для создания автоматизированной и интеллектуальной системы кибербезопасности, управляемой данными, могут быть использованы соответствующие данные кибербезопасности [105]; для создания персонализированных контекстно-зависимых интеллектуальных мобильных приложений могут быть использованы соответствующие мобильные данные [103] и т. д. Таким образом, срочно необходимы инструменты и методы управления данными, имеющие возможность извлекать информацию или полезные знания из данных своевременно и разумно, на которых основаны реальные приложения.

Искусственный интеллект (ИИ), в частности, машинное обучение (МО) быстро развивались в последние годы в контексте анализа данных и вычислений, что обычно позволяет приложениям функционировать интеллектуальным образом [95]. МО обычно предоставляет системам возможность учиться и совершенствоваться на основе опыта автоматически, без специального программирования, и обычно упоминается как самый популярный

Данная статья является частью тематического сборника «Достижения в области вычислительных подходов для искусственного интеллекта, обработки изображений, Интернета вещей и облачных приложений» под гостевыми редакторами Бхану Пракаш КН и М. Шивакумар.

* Икбал Х. Саркер
msarker@swin.edu.au

1 Технологический университет Суинберна, Мельбурн, Виктория 3122, Австралия

2 Факультет компьютерных наук и техники, Читтагонгский университет инженерии и технологий, 4349 Чаттограм, Бангладеш

новейшие технологии четвертой промышленной революции (4ПР) или Industry 4.0) [103, 105]. «Industry 4.0» [114] — это, как правило, постоянная автоматизация традиционных производственных и промышленных практик, включая разведочную обработку данных, с использованием новых интеллектуальных технологий, таких как автоматизация машинного обучения. Таким образом, для разумного анализа этих данных и разработки соответствующих реальных приложений алгоритмы машинного обучения являются ключом. Алгоритмы обучения можно разделить на четыре основных типа, такие как контролируемое, неконтролируемое, полуконтролируемое и обучение с подкреплением в области [75], кратко обсуждаемой в разделе «Типы реальных данных и методы машинного обучения» . Популярность этих подходов к обучению растет с каждым днем, что показано на рис. 1, на основе данных, собранных из Google Trends [4] за последние пять лет. Ось x рисунка указывает конкретные даты и соответствующую оценку популярности в диапазоне от 0 (минимум)

до 100 (максимум) показано на оси Y. Согласно рис. 1, значения показателей популярности для этих типов обучения в 2015 году были низкими и растут с каждым днем. Эти статистические данные мотивируют нас к изучению машинного обучения в этой статье, которое может играть важную роль в реальном мире посредством автоматизации Industry 4.0.

В целом, эффективность и результативность решения машинного обучения зависят от природы и характеристик данных и производительности алгоритмов обучения. В области алгоритмов машинного обучения существуют методы классификационного анализа, регрессии, кластеризации данных, проектирования признаков и снижения размерности, обучения ассоциативным правилам или обучения с подкреплением для эффективного построения систем, управляемых данными [41, 125].
Кроме того, глубокое обучение возникло из искусственной нейронной сети, которая может быть использована для интеллектуального анализа данных, что известно как часть более широкого семейства подходов машинного обучения [96]. Таким образом, выбор правильного алгоритма обучения, который подходит для целевого приложения в конкретная область является сложной. Причина в том, что цель различных алгоритмов обучения различна, даже результат различных алгоритмов обучения в схожей категории может различаться в зависимости от характеристик данных [106]. Таким образом, важно понимать принципы различных алгоритмов машинного обучения и их применимость в различных областях реального мира, таких как системы IoT, службы кибербезопасности, бизнес-системы и системы рекомендаций, умные города, здравоохранение и COVID-19, контекстно-зависимые системы, устойчивое сельское хозяйство и многое другое, что кратко объясняется в разделе.

«Применение машинного обучения» .

Основываясь на важности и потенциале «машинного обучения» для анализа данных, упомянутых выше, в этой статье мы даем всесторонний обзор различных типов алгоритмов машинного обучения , которые могут применяться для повышения интеллекта и возможностей приложения. Таким образом, ключевым вкладом этого исследования является объяснение принципов и потенциала различных методов машинного обучения, а также их применимости в различных областях реального мира, упомянутых ранее. Таким образом, цель этой статьи состоит в том, чтобы предоставить базовое руководство для тех представителей академических кругов и промышленности , которые хотят изучать, исследовать и разрабатывать автоматизированные и интеллектуальные системы, управляемые данными, в соответствующих областях на основе методов машинного обучения.

Основные выводы данной статьи перечислены ниже:

– Определить область нашего исследования, принимая во внимание природу и характеристики различных типов реальных данных и возможности различных методов обучения.

– Предоставить комплексное представление об алгоритмах машинного обучения, которые можно применять для повышения интеллектуальности и возможностей приложений, управляемых данными.

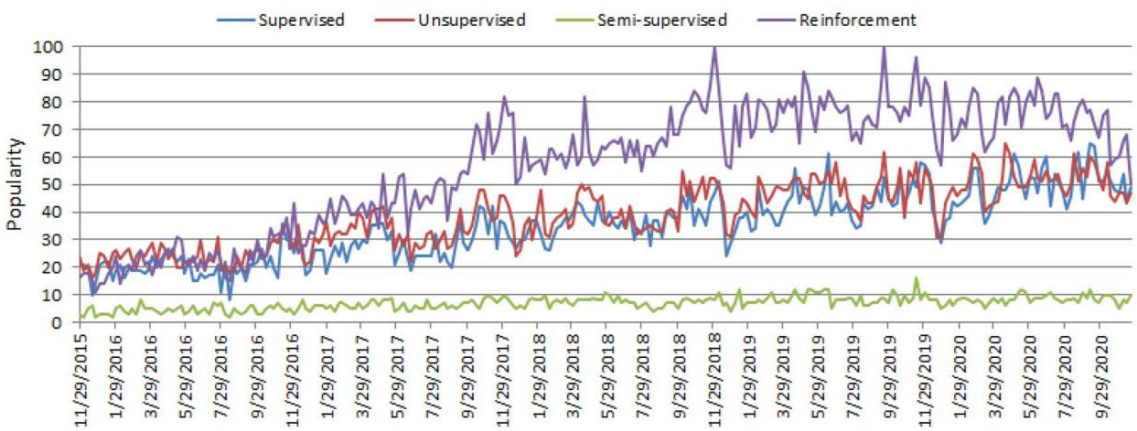

Рис. 1. Оценка мировой популярности различных типов алгоритмов МО (контролируемых, неконтролируемых, полуконтролируемых и с подкреплением) в диапазоне от 0 (мин.) до 100 (макс.) с течением времени, где ось x представляет информацию о временной метке, а ось y представляет соответствующую оценку.

– Обсудить применимость решений на основе машинного обучения в различных реальных прикладных областях.
– Выделить и обобщить потенциальные направления исследований в рамках нашего исследования интеллектуального анализа данных и услуг.

Остальная часть статьи организована следующим образом. В следующем разделе представлены типы данных и алгоритмы машинного обучения в более широком смысле и определены рамки нашего исследования. Мы кратко обсуждаем и объясняем различные алгоритмы машинного обучения в следующем разделе, за которым следуют различные области применения в реальном мире, основанные на алгоритмах машинного обучения. В предпоследнем разделе мы выделяем несколько исследовательских вопросов и потенциальных будущих направлений, а в заключительном разделе завершается эта статья.

Типы реальных данных и машин Методы обучения

Алгоритмы машинного обучения обычно потребляют и обрабатывают данные, чтобы изучить соответствующие закономерности относительно людей, бизнес-процессов, транзакций, событий и т. д. Далее мы обсудим различные типы данных реального мира, а также категории алгоритмов машинного обучения.

Типы реальных данных

Обычно доступность данных рассматривается как ключ к построению модели машинного обучения или систем реального мира, управляемых данными [103, 105]. Данные могут быть разных форм, таких как структурированные, полуструктурированные или неструктурированные [41, 72].
Кроме того, «метаданные» — это еще один тип, который обычно представляет данные о данных. Далее мы кратко обсудим эти типы данных.

– Структурированный: имеет четко определенную структуру, соответствует модели данных, следующей стандартному порядку, которая высокоорганизована и легкодоступна, и используется сущностью или компьютерной программой. В четко определенных схемах, таких как реляционные базы данных, структурированные данные обычно хранятся, т. е. в табличном формате. Например, имена, даты, адреса, номера кредитных карт, информация о запасах, геолокация и т. д. являются примерами структурированных данных.
– Неструктурированные: С другой стороны, не существует заранее определенного формата или организации для неструктурированных данных, что значительно усложняет их сбор, обработку и анализ, в основном они содержат текст и мультимедийные материалы. Например, данные датчиков, электронные письма, записи блогов, вики и документы текстовых процессоров, файлы PDF, аудиофайлы, видео, изображения, презентации, веб-страницы и многое другое. другие типы деловых документов можно рассматривать как неструктурированные данные.
– Полуструктурированные: Полуструктурированные данные не хранятся в реляционная база данных, подобная структурированным данным, упомянутым выше, но обладающая определенными организационными свойствами, которые облегчают ее анализ. HTML, XML, документы JSON, базы данных NoSQL и т. д. являются некоторыми примерами полуструктурированных данных.
– Метаданные: это не обычная форма данных, а «данные о данных» . Основное различие между «данными» и «метаданными» заключается в том, что данные – это просто материал, который может классифицировать, измерять или даже документировать что-либо относительно свойств данных организации. С другой стороны, метаданные описывают соответствующую информацию о данных, придавая ей большую значимость для пользователей данных. Базовым примером метаданных документа может быть автор, размер файла, дата создания документа, ключевые слова для определения документа и т. д.

В области машинного обучения и науки о данных исследователи используют различные широко используемые наборы данных для разных целей.
Это, например, наборы данных кибербезопасности, такие как NSL-KDD [119], UNSW-NB15 [76], ISCX'12 [1], CIC-DDoS2019 [2], Bot-IoT [59] и т. д., наборы данных смартфонов, такие как журналы телефонных звонков [84, 101], журналы SMS [29], журналы использования мобильных приложений [137] [117], журналы уведомлений мобильных телефонов [73] и т. д., данные IoT [16, 57, 62], данные сельского хозяйства и электронной коммерции [120, 138], данные о здоровье, такие как болезни сердца [92], сахарный диабет [83, 134], COVID-19 [43, 74] и т. д., и многие другие в различных областях применения. Данные могут быть разных типов, обсуждавшихся выше, которые могут различаться в зависимости от приложения в реальном мире. Для анализа таких данных в определенной проблемной области и извлечения из данных полезных идей или знаний для создания реальных интеллектуальных приложений можно использовать различные типы методов машинного обучения в соответствии с их возможностями обучения, что обсуждается далее.

Типы методов машинного обучения

Алгоритмы машинного обучения в основном делятся на четыре категории: контролируемое обучение, неконтролируемое обучение, полуконтролируемое обучение и обучение с подкреплением [75], как показано на рис. 2. Далее мы кратко обсудим каждый тип метода обучения с точки зрения его применимости для решения реальных задач.

– Контролируемое: Контролируемое обучение обычно является задачей машинного обучения по изучению функции, которая отображает входные данные в выходные данные на основе выборочных пар входных данных и выходных данных [41]. Оно использует помеченные обучающие данные и набор обучающих примеров для вывода функции. Контролируемое обучение выполняется, когда определенные цели идентифицированы для достижения