УДК 004.8 **Информационные технологии**

Маркова Анастасия Владимировна, *магистрант, Московский государственный технический университет им. Н.Э. Баумана*
Афанасьев Арсений Геннадьевич, *ассистент, Московский государственный технический университет им. Н.Э. Баумана*
Афанасьев Геннадий Иванович, *кандидат технических наук, доцент, Московский государственный технический университет им. Н.Э. Баумана*

КЛАССИФИКАЦИЯ НА ОСНОВЕ АЛГОРИТМА ДЕРЕВА РЕШЕНИЙ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

Аннотация: Классификаторы дерева решений считаются выдающимися из наиболее известных методов представления классификации данных классификаторов. Разные исследователи из разных областей и с разным опытом рассматривали проблему расширения дерева решений на основе доступных данных, таких как машинное обучение, распознавание образов и статистика. В различных областях, таких как анализ медицинских заболеваний, классификация текстов, классификация пользовательских смартфонов, изображений и многих других, использование классификаторов дерева решений было предложено во многих отношениях. В этой статье подробно рассматривается подход к деревьям решений. Кроме того, всесторонне оцениваются и излагаются особенности работы, такие как используемые алгоритмы/подходы, наборы данных и достигнутые результаты. Кроме того, были обсуждены все проанализированные подходы, чтобы проиллюстрировать темы авторов и определить наиболее точные классификаторы. В результате обсуждается использование различных типов наборов, данных и анализируются их результаты.

Ключевые слова: машинное обучение, классификация, дерево решений, искусственный интеллект, алгоритм.

Abstract: Decision tree classifiers are considered to be the preeminent of the best known methods for representing the classification of data classifiers. Different researchers from different fields and backgrounds have addressed the problem of extending a decision tree based on available data such as machine learning, pattern recognition, and statistics. In various fields such as medical disease analysis, text classification, user smartphone classification, image classification, and many others, the use of decision tree classifiers has been proposed in many ways. This article takes a detailed look at the decision tree approach. In addition, specific features of the work, such as algorithms/approaches used, datasets and results achieved, are comprehensively assessed and presented. In addition, all analyzed approaches were discussed to illustrate the authors' topics and to determine the most accurate classifiers. As a result, the use of various types of datasets is discussed and their results are analyzed.

Keywords: machine learning, classification, decision tree, artificial intelligence, algorithm.

Введение

В настоящее время технологии сильно развились, особенно в области машинного обучения, которое полезно для уменьшения человеческого труда. В области искусственного интеллекта машинное обучение объединяет статистику и информатику для создания алгоритмов, которые становятся более эффективными, когда они зависят от соответствующих данных, а не от конкретных инструкций. Помимо распознавания речи, обнаружения изображений, локализации текста машинное обучение — это изучение вычислительных алгоритмов, которые автоматически совершенствуются на основе опыта. Он рассматривается как подмножество искусственного интеллекта [1; 2; 3]. Алгоритмы машинного обучения для того, чтобы производить предсказания или решения, которые не специально запрограммированы для этого, создают модельную популяцию на основе выборки, определяемой как «данные для обучения» [4]. В широкой области

приложений, таких как фильтрация электронной почты и компьютерное зрение, алгоритмы машинного обучения используются там, где сложно или нецелесообразно создавать традиционные алгоритмы для реализации требуемых функций [5]. У машинного обучения есть много применений, наиболее известным из которых является прогнозный анализ данных. Два основных механизма могут быть разбиты на выполнение классификации машинного обучения: разработка модели и оценка модели.

Используя один и тот же набор атрибутов, описывается любой экземпляр в каждом наборе данных, используемом алгоритмами машинного обучения. Атрибуты могут быть непрерывными, категориальными или бинарными [6]. Если кейсы распознаются с распознанными метками (правильными выходами), то обучение называется контролируемым [7]. Он также анализирует данные тестирования и создает производную задачу, которую можно использовать для новых примеров для сопоставления [8]. Однако каждый объект ввода данных имеет предварительно назначенную метку класса. Основная функция контролируемых алгоритмов состоит в том, чтобы изучить модель, которая создает одинаковую маркировку для предлагаемых данных и хорошо популяризирует невидимые данные. Это основная цель алгоритмов классификации.

Классификация пытается предсказать класс цели с наивысшей точностью. Алгоритм классификации выясняет связь между входным атрибутом и выходным атрибутом для построения модели, которая представляет собой процесс обучения [9; 10]. Объем данных, получаемых в средах интеллектуального анализа данных, огромен. Если набор данных правильно классифицирован и содержит минимальное количество узлов, то использование метода дерева решений является оптимальным [11].

Дерево решений — это древовидный метод, в котором любой путь, начинающийся от корня, описывается последовательностью, разделяющей данные, до тех пор, пока не будет достигнут логический результат в листовом узле [12]. Это иерархическая иллюстрация отношений знаний, которые

содержат узлы и соединения. Когда отношения используются для классификации, узлы представляют цели [13; 14].

В этой статье проводится всесторонний обзор последних и наиболее эффективных подходов, которые применялись исследователями за последние три года в отношении деревьев решений в различных областях машинного обучения. Кроме того, обобщаются детали этого метода, такие как использование алгоритмов/подходов, наборов данных и полученные результаты. Кроме того, в этом исследовании были выделены наиболее часто используемые подходы и методы с наивысшей точностью.

Алгоритм дерева решений

Одним из широко используемых методов интеллектуального анализа данных являются системы, создающие классификаторы [15]. В интеллектуальном анализе данных алгоритмы классификации способны обрабатывать огромный объем информации. Его можно использовать для создания предположений относительно имен классов категорий, для классификации знаний на основе обучающих наборов и меток классов, а также для классификации вновь полученных данных [16]. Алгоритмы классификации в машинном обучении содержат несколько алгоритмов, и в этой работе основное внимание уделяется алгоритму дерева решений в целом. Рисунок 1 иллюстрирует структуру дерева решений.

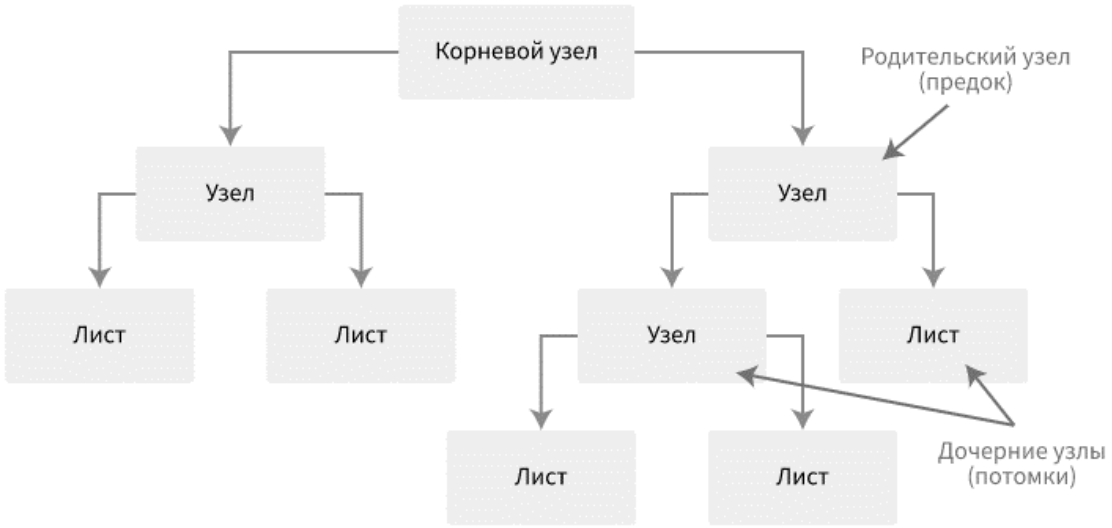

Рисунок 1. Структура дерева решений

Деревья решений являются одним из эффективных методов, обычно используемых в различных областях, таких как машинное обучение, обработка изображений и выявление закономерностей. Дерево решений представляет собой последовательную модель, которая эффективно и связно объединяет серию основных тестов, где числовая характеристика сравнивается с пороговым значением в каждом тесте. Концептуальные правила построить намного проще, чем числовые веса в нейронной сети связей между узлами. В основном для целей группировки используется дерево решений. Кроме того, этот алгоритм является обычно используемой моделью классификации в Data Mining. Узлы и ветви состоят из каждого дерева. Каждый узел представляет признаки в категории, подлежащей классификации, и каждое подмножество определяет значение, которое может быть принято узлом [17]. Из-за простоты анализа и точности в различных формах данных деревья решений нашли множество областей применения. На Рисунке 2 показан пример дерева решений.

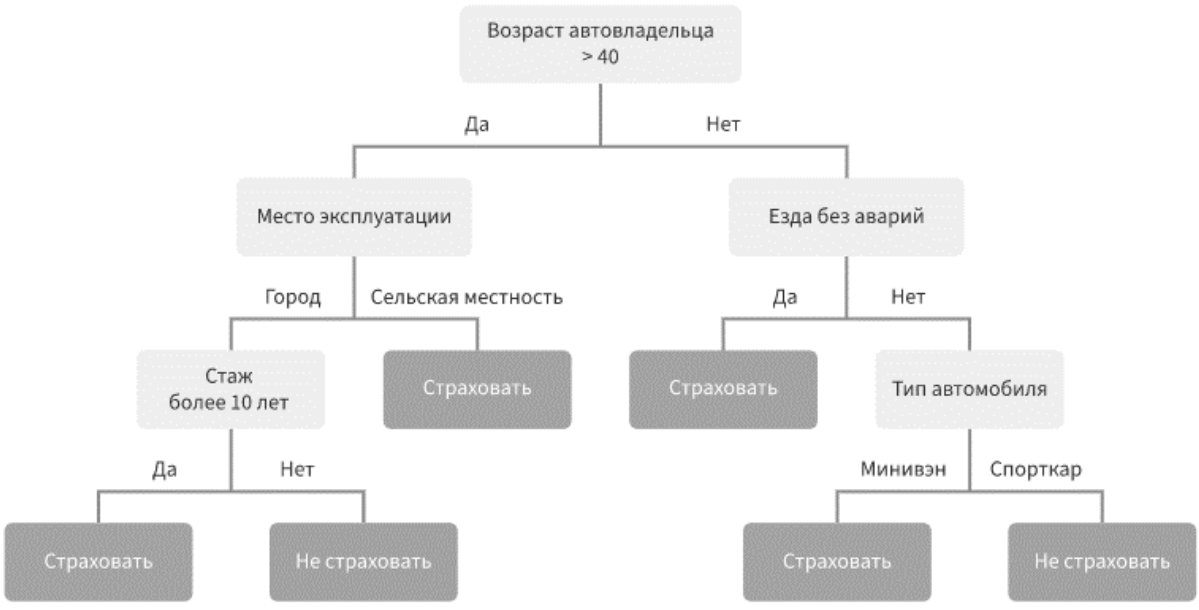

Рисунок 2. Пример дерева решений.

Типы алгоритмов дерева решений

Существует несколько типов алгоритмов дерева решений, таких как:

Итеративные дихотомии 3 (ID3), Преемник ID3 (C4.5), Дерево классификации и регрессии (CART) [18], Детектор автоматических взаимодействий CHi-squared (CHAID) [19], Сплайны многомерной адаптивной регрессии (MARS) [20], обобщенные, несмещенные, обнаружение и оценка взаимодействия (GUIDE), деревья условного вывода (CTREE) [21; 22], правило классификации с непредвзятым выбором и оценкой взаимодействия (CRUISE), Беспристрастное и эффективное статистическое дерево (QUEST) [23; 24]. В Таблице 1 показано сравнение часто используемых алгоритмов дерева решений [25].

Таблица 1. Сравнение популярных алгоритмов в методах дерева принятия решений.

Методы	CART	C4.5	CHAID	QUEST
Мера, используемая для сбора входных переменных	Коэффициент Джини	Энтропийный прирост информации	Критерий хи-квадрат	Хи-квадрат для категориальных переменных, дисперсионный анализ для непрерывных/порядковых переменных
Зависимые переменные	Категориальные, непрерывные	Категориальные, непрерывные	Категориальные	Категориальные
Входные переменные	Категориальные, непрерывные	Категориальные, непрерывные	Категориальные, непрерывные	Категориальные, непрерывные
Разделитель в каждом узле	Бинарный	N-арный	N-арный	Бинарный

Энтропия и прирост информации

Энтропия используется для измерения примеси или случайности набора данных. Значение энтропии всегда лежит между 0 и 1. Ее значение лучше, когда оно равно 0, и хуже, когда оно равно 1, т.е. чем ближе ее значение к 0, тем лучше, как показано на Рисунке 3. Если целью является G с разными значениями атрибутов, энтропия классификации множества S относительно состояний c. Как показано в Формуле 1.

$$\text{Entropy}(S) = \sum_{i=1}^{c} P_i \log_2 P_i \qquad (1)$$

Где P_i — отношение номера выборки подмножества к i-му значению признака.

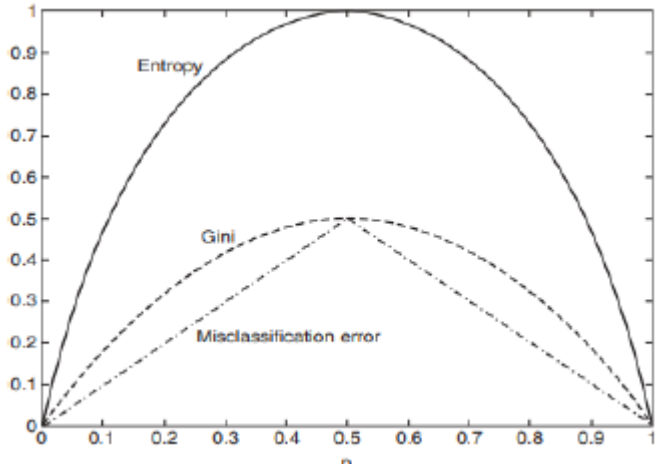

Рисунок 3. Значение энтропии.

Прирост информации является одним из показателей, используемых для сегментации, и его часто называют взаимной информацией. Это показывает, насколько хорошо известно значение случайной величины. Это противоположность энтропии, чем выше ее значение, тем лучше. Прирост данных $Gain(S,A)$ определяется следующим образом по определению энтропии, как показано в Формуле 2.

$$Gain(S,A) = \sum_{v \in V(A)} \frac{|S_v|}{|S|} \text{Entropy}(S) \qquad (2)$$

Где диапазон атрибута A равен $V(A)$, а S_v является подмножеством множества S, равным значению атрибута v.

Преимущества и недостатки дерева решений

Алгоритм дерева решений является частью семейства алгоритмов обучения с учителем, и его основная цель состоит в том, чтобы построить модель обучения, которую можно использовать для прогнозирования класса или значения целевых переменных с помощью правил принятия решений, выведенных из обучающих данных. Этот алгоритм можно использовать для решения задач регрессии и классификации, но он имеет преимущества и

недостатки, которые показаны в Таблице 2.

Таблица 2. Преимущества и недостатки.

Преимущества	Недостатки
1) Простота понимания 2) Формируют четкие и понятные правила классификации 3) Способны генерировать правила в областях, где специалисту трудно формализовать свои знания 4) Легко визуализируются, то есть могут «интерпретироваться» не только как модель в целом, но и как прогноз для отдельного тестового субъекта (путь в дереве) 5) Быстро обучаются и прогнозируют 6) Не требуется много параметров модели 7) Поддерживают как числовые, так и категориальные признаки	1) Чувствительны к шумам во входных данных. 2) Возможно переобучение дерева решений, из-за чего приходится прибегать к методу «отсечения ветвей», установке минимального числа элементов в листьях дерева или максимальной глубины дерева. 3) Сложный поиск оптимального дерева решений. 4) Дерево решений делает константный прогноз для объектов, находящихся в признаковом пространстве вне параллелепипеда, который охватывает не все объекты обучающей выборки.

Сравнение

Алгоритмы классификации деревьев решений состоят из нескольких типов, которые используются для генерации деревьев. Это осуществляется путем контроля как непрерывных, так и периодических атрибутов пропущенных значений. Дерево решений генерируется формой, которая обычно представляется в виде статистического классификатора и может использоваться для кластеризации. Узлы и ответвления включаются в дерево. Для каждого узла требуются задачи, основанные на одном или нескольких свойствах, т. е. сравнение значения атрибута с константой или использование других функций для сравнения более чем одного свойства. Для дерева решений сбор обучающих данных иногда называют деревом результатов. Чтобы включить классификации в машинное обучение и интеллектуальный анализ данных с использованием алгоритма. В следующей последовательности этот алгоритм применяется итеративно, и для классификации требуется трехэтапный процесс: создание модели (обучение), модель оценки (точность) и использование модели (классификация). Этап классификации дерева решений основан на проценте полученной информации, измеряемой энтропией. Метрика охвата используется для описания тестовых характеристик для узла в дереве и

называется шкалой выбора свойства (свойство). В качестве тестовой функции для текущего узла вычисляется свойство наилучшего знания. В некоторых исследованиях предлагались подходы для преодоления недостатков задач алгоритма, чтобы можно было рассчитать оптимальные деревья на основе обзора, который был выполнен ранее, без подробных деталей и примеров. Методы дерева решений показали, что описанных выше проблем можно избежать. Кроме того, он предоставит указанному набору данных соответствующее решение. Во многих исследованиях проводились разные наборы данных, и подход данного алгоритма использовался для устранения его недостатков и повышения производительности. В исследовании использовалось несколько методов оптимизации для укрепления дерева решений в хранимых наборах данных UCI ML; на основании результатов оценки было показано, что подход алгоритма имеет самую высокую точность, которая составляет 99,93% по сравнению с другими методами, такими как метод ближайших соседей, алгоритм синтаксического разбора и метод опорных векторов, которые менее эффективны, чем подход дерева решений. В задаче на сегментацию в исследовании использовался подход дерева решений для идентификации и извлечения кровеносной артерии для правильной сегментации диска зрительного нерва, что привело к более высоким результатам, равным 99,61%. В конечном счете, при использовании наборов данных библиотеки машинного обучения UCI и наборов данных CICIDS2017, алгоритм дерева решений оказался самым точным. Хотя в исследовании использовались XGBoost (алгоритм машинного обучения, основанный на дереве поиска решений и использующий фреймворк градиентного бустинга) и метод случайного леса на наборах данных курильщиков Китайского центра по контролю и профилактике заболеваний, было обнаружено, что подход дерева решений достиг наивысшей точности; что составляет 84,11%. Кроме того, на основе исследований с использованием метода ближайших соседей в наборах данных CICIDS2017, RNA-seq Malaria и Wisconsin Breast Cancer было обнаружено, что подход DT имел самую высокую точность во всех трех

исследованиях. Кроме того, его точность была выше при использовании наборов данных CICIDS2017, точность которых достигла 99,91%.

Заключение

Классификаторы дерева решений известны своим улучшенным представлением результатов производительности. Из-за их высокой точности оптимизированные параметры разделения и улучшенные методы обрезки деревьев (ID3, C4.5, CART, CHAID и QUEST) обычно используются всеми признанными классификаторами данных. Отдельные наборы данных используются для обучения выборок из огромного набора данных, что, в свою очередь, влияет на точность тестового набора. У деревьев решений есть несколько возможных проблем с надежностью, адаптацией масштабируемости и оптимизацией высоты. Но, в отличие от других методов классификации данных, деревья решений создают эффективный набор правил, который прост для понимания. В этой статье рассматриваются самые последние исследования, которые проводятся во многих областях. Кроме того, детали, используемые в методах/алгоритмах, наборы данных использовались авторы и достигнутые результаты, связанные с точностью, суммируются для деревьев решений. Наконец, наилучшая точность, достигнутая для алгоритма дерева решений, составляет 99,93%, когда он использует репозиторий машинного обучения в качестве набора данных.

Библиографический список:

1. Афанасьев Г.И., Дин Но. Применение глубокого обучения в визуализации диагностики туберкулеза // Искусственный интеллект в автоматизированных системах управления и обработки данных (ИИАСУ 22). Сборник статей Всероссийской научной конференции. В 2 т. (Москва. МГТУ им. Н.Э. Баумана. 27-28 апреля 2022 г.). - М.: МГТУ им. Н.Э. Баумана, Т.2. С. 237-241.

2. Вей Пхьоу Ту, Афанасьев Г.И. Интеллектуальная система идентификации человека по отпечаткам пальцев // Искусственный интеллект в

автоматизированных системах управления и обработки данных (ИИАСУ 22). Сборник статей Всероссийской научной конференции. В 2 т. (Москва. МГТУ им. Н.Э. Баумана. 27-28 апреля 2022 г.). - М.: МГТУ им. Н.Э. Баумана, Т.2. С. 346-351.

3. Чжан Чжибо, Афанасьев Г.И. Применение моделей глубокого обучения в области рекомендательных систем, основанных на содержании //Искусственный интеллект в автоматизированных системах управления и обработки данных (ИИАСУ 22). Сборник статей Всероссийской научной конференции. В 2 т. (Москва. МГТУ им. Н.Э. Баумана. 27-28 апреля 2022 г.). - М.: МГТУ им. Н.Э. Баумана, Т.1. С. 278-284.

4. D. Maulud, A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” Journal of Applied Science and Technology Trends, vol. 1, no. 4, pp. 140–147, 2020.

5. G. Carleo et al., “Machine learning and the physical sciences,” Reviews of Modern Physics, vol. 91, no. 4, p. 045002, 2019.

6. S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: A review of classification techniques,” Emerging artificial intelligence applications in computer engineering, vol. 160, no. 1, pp. 3–24, 2007.

7. S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “Machine learning: a review of classification and combining techniques,” Artif Intell Rev, vol. 26, no. 3, pp. 159–190, Nov. 2006.

8. D. Sharma and N. Kumar, “A Review on Machine Learning Algorithms, Tasks and Applications,” vol. 6, pp. 2278–1323, Oct. 2017.

9. S. Patil and U. Kulkarni, “Accuracy Prediction for Distributed Decision Tree using Machine Learning approach,” in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Apr. 2019, pp. 1365–1371.

10. N. S. Ahmed and M. H. Sadiq, “Clarify of the random forest algorithm in an educational field,” in 2018 International Conference on Advanced Science and Engineering (ICOASE), 2018, pp. 179–184.

11. D. V. Patil and R. S. Bichkar, “A Hybrid Evolutionary Approach To

Construct Optimal Decision Trees With Large Data Sets," in 2006 IEEE International Conference on Industrial Technology, Dec. 2006, pp. 429–433.

12. F. Yang, "An Extended Idea about Decision Trees," in 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Dec. 2019, pp. 349–354.

13. A. Shamim, H. Hussain, and Maqbool Uddin Shaikh, "A framework for generation of rules from decision tree and decision table," in 2010 International Conference on Information and Emerging Technologies, Jun. 2010, pp. 1–6.

14. Priyanka and D. Kumar, "Decision tree classifier: a detailed survey," International Journal of Information and Decision Sciences, vol. 12, no. 3, pp. 246–269, 2020.

15. R. Kumar and R. Verma, "Classification algorithms for data mining: A survey," International Journal of Innovations in Engineering and Technology (IJIET), vol. 1, no. 2, pp. 7–14, 2012.

16. S. S. Nikam, "A comparative study of classification techniques in data mining algorithms," Oriental journal of computer science & technology, vol. 8, no. 1, pp. 13–19, 2015.

17. A. Dey, "Machine learning algorithms: a review," International Journal of Computer Science and Information Technologies, vol. 7, no. 3, pp. 1174–1179, 2016.

18. A. E. Brodley and P. E. Utgoff, "Multivariate decision trees," Machine learning, vol. 19, no. 1, pp. 45–77, 1995.

19. G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," Energy, vol. 32, no. 9, pp. 1761–1768, Sep. 2007.

20. S. Singh and P. Gupta, "Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey," International Journal of Advanced Information Science and Technology (IJAIST), vol. 27, no. 27, pp. 97–103, 2014.

21. L. Rokach and O. Maimon, "Top-Down Induction of Decision Trees Classifiers—A Survey," Systems, Man, and Cybernetics, Part C: Applications and

Reviews, IEEE Transactions on, vol. 35, pp. 476–487, Dec. 2005.

22. T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, “A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms,” Machine learning, vol. 40, no. 3, pp. 203–228, 2000.

23. W.-Y. Loh, “Fifty Years of Classification and Regression Trees,” International Statistical Review, vol. 82, Jun. 2014.

24. S. R. Jiao, J. Song, and B. Liu, “A Review of Decision Tree Classification Algorithms for Continuous Variables,” in Journal of Physics: Conference Series, 2020, vol. 1651, no. 1, p. 012083.

25. Y.-Y. Song and Y. Lu, “Decision tree methods: applications for classification and prediction,” Shanghai archives of psychiatry, vol. 27, pp. 130–5, Apr. 2015.