А.Н. Сулейманова
(*Москва*)

ОБЗОР РАЗВИТИЯ АЛГОРИТМОВ ДЕРЕВЬЕВ РЕШЕНИЙ

Деревья решений – метод классификации и предсказания, распространенный в прикладных исследованиях в силу простоты применения и интерпретации. Ввиду большого количества самих алгоритмов, разрозненности литературы и программного обеспечения для работы с ними выбор одного из методов представляет собой непростую задачу. В результате исследователи предпочитают использовать хорошо знакомые и давно использующиеся алгоритмы, несмотря на их явные недостатки. Обзор ставит целью выделить и описать актуальные направления развития этого класса методов и систематизировать новации в его применении за 2014–2019 гг. Для выделения актуальных тематических направлений развития методов используется построение библиографической сети на ключевых словах. Обзор позволит упростить навигацию в растущем избытке алгоритмов и дополнить данные, представленные в предыдущих обзорах.

Ключевые слова: деревья решений, деревья классификации, деревья регрессии, библиографический анализ, сеть ключевых слов, большие данные.

Введение

Деревья решений – распространенный в прикладных социологических исследованиях метод классификации и прогнозиро-

Анна Наильевна Сулейманова – аспирант школы по социологическим наукам, преподаватель кафедры методов сбора и анализа социологической информации, факультет социальных наук, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: asuleymanova@hse.ru. ORCID: 0000-0002-4379-3835

вания[1]. К примеру, с помощью этого метода можно предсказать, какие категории респондентов склонны к различным типам неответов при заполнении опросника, чтобы заранее заложить в дизайн выборки поправку на смещения. С помощью деревьев решений можно производить логические правила, прогнозирующие тот или иной исход, например – выплатит потенциальный заемщик кредит или нет в зависимости от его социально-демографических признаков и отдельных событий кредитной истории. Кроме того, деревья решений используются для поиска взаимодействий признаков для последующего включения в регрессионные модели [1, с. 47].

Распространенность этого метода обусловлена простотой интерпретации получаемых решений, скоростью вычислений и доступностью программного обеспечения (в том числе бесплатного) [2; 3]. С момента первой публикации, посвященной деревьям решений, прошло более 50 лет [4], а последний исчерпывающий обзор этого класса методов датирован 2014 г. [2]. Поскольку более чем за пять лет с момента выхода указанного обзора взгляды на метод и его использование успели претерпеть существенные изменения, мы предприняли попытку систематически отразить новации в представлениях о свойствах, принципах работы и перспективах развития этих методов в социологических исследованиях и смежных науках. Для выделения актуальных тематик дискуссии на сегодняшний день мы использовали анализ библиографической сети, образованной публикациями за 2014–2019 гг., так или иначе затрагивающими деревья принятия решений.

В дискуссии к упомянутому выше обзору отмечается высокая фрагментированность исследований, касающихся деревьев решений [3]. Она обусловлена тремя факторами. Во-первых, деревья, в силу простоты использования и универсальности, применяются

[1] Термин «деревья решений» используется как обобщающий для деревьев классификации, деревьев регрессии и алгоритмов, совмещающих классификацию и регрессию.

почти во всех областях знания. Разнообразие проблем, решаемых с помощью деревьев решений, порождает и разнообразие самих алгоритмов, что само по себе может вызывать затруднения у исследователей. В результате для многих из них оказывается проще разработать новый, адекватный собственным запросам алгоритм, чем выбирать подходящий из имеющихся, что и становится вторым фактором фрагментированности литературы об этом классе методов. Третий фактор заключается в том, что модификации старых алгоритмов деревьев решений часто представляются как самостоятельные методы [3, p. 362]. Разнообразие как платных, так и бесплатных программных решений для применения этих методов только усугубляет методологическую раздробленность этих алгоритмов. Выходом из этой ситуации могут выступать разработка единого языка описания деревьев решений и единое программное обеспечение в формате Open Source [3, p. 366], однако до тех пор, пока этот консенсус не достигнут, в качестве опоры для исследователя могут выступать систематические обзоры.

Методология обзора

Предметом обзора выступают алгоритмы единичных деревьев решений. Цель обзора – систематизация развития алгоритмов деревьев решений за 2014–2019 гг. для упрощения навигации в избытке алгоритмов и дополнения информации, представленной в предыдущих обзорах.

Для подбора релевантной литературы и выстраивания структуры обзора применялись два подхода: экспертный и алгоритмический [5]. Первый заключался в поиске исследований, касающихся проблем, которые подняты в обзоре В. Ло и дискуссии к нему; экспертом в этом случае выступает автор обзора. Во второй части при помощи анализа библиографических сетей на основе ключевых слов были выделены актуальные направления развития рассматриваемого класса методов. Сочетание экспертного и алго-

ритмического подходов позволило добиться полноты отражения основной дискуссии вокруг деревьев решений, разворачивавшейся в 2014–2019 гг., за счет их комплементарности [5].

Алгоритмический подход был призван выявить методы и особенности их применения, находившиеся на повестке в научной литературе последние годы. В качестве способа определения актуальности того или иного алгоритма или специфики его применения мы воспользовались анализом библиографической сети по ключевым словам. Если при использовании экспертного подхода мы стремились проанализировать прибавление методологического знания, то алгоритмический подход призван, во-первых, продемонстрировать тенденции не только среди разработчиков алгоритмов, но и среди их пользователей без смещения на экспертизу автора, во-вторых – дополнить список обозреваемых публикаций на основе ключевых слов, которые не встретились в экспертной выборке, и, в-третьих, оценить адекватность экспертной выборки и структурировать обзор в соответствии с выявленными категориями ключевых слов.

Формирование выборки публикаций

Для сбора исходного массива данных были использованы данные баз научного цитирования РИНЦ и Web of Science за 2014–2019 гг.

В Web of Science поиск релевантной литературы производился по запросу «“Decision tree” or “Classification tree” or “Regression tree”» по названиям, аннотациям, и ключевым словам (поле «Topic») с 2014 по 2019 г. Дополнительно были уточнены категории Web of Science[1] для того, чтобы избавиться от случайных совпадений со

[1] Статьи были отобраны из следующих категорий Web of Science: “Computer Science Artificial Intelligence”, “Computer Science Theory Methods”, “Mathematics Applied”, “Computer Science Information Systems”, “Computer Science Interdisciplinary Applications”, “Statistics Probability”, “Computer Science Software

словосочетаниями, не относящимися к деревьям классификации как методу анализа данных. Полученная в результате база публикаций содержала 2939 записей, включающих как методологические, так и эмпирические статьи.

В РИНЦ поиск публикаций производился с помощью запроса «“деревья решений” или “деревья классификации” или “деревья регрессии”» с учетом морфологии по тематикам «социология» и «статистика» в названиях, аннотациях и ключевых словах, в результате чего был получен массив из 76 публикаций.

Для экспертного отбора релевантной литературы и в том, и в другом результирующем массиве отбирались публикации по социологии, статистике и смежным наукам, прибавлявшие методологическое знание о свойствах, применении и принципах работы деревьев решений, после чего выборка дополнялась при помощи экспертного поиска статей, соответствующих этим требованиям, но не попавших в исходный массив [5]. В результате из РИНЦ были отобраны 4 публикации[1] с методологическими новациями трех типов: 1) презентация нового авторского алгоритма или статистического инструмента; 2) презентация или предложение модификации алгоритма или способа подготовки данных для анализа; 3) результаты статистических экспериментов, демонстрирующих возможности или ограничения алгоритмов применительно к специфическим типам данных. В экспертную выборку не попадали статьи, нацеленные на демонстрацию применения известного алгоритма на результатах опросов, а также эмпирические статьи, использующие деревья в качестве метода анализа данных, чем и объясняется ее узость по сравнению с исходной. Из Web of Science при помощи тех же принципов были отобраны еще 40 статей.

Engineering”, “Multidisciplinary Sciences”, “Social Sciences Mathematical Methods”, “Medical Informatics”, “Social Sciences Interdisciplinary”.

[1] Из них 3 статьи не содержались в исходной выдаче из-за неполноты списков ключевых слов в статьях.

Алгоритмический подход

Из полученной базы публикаций Web of Science[1], содержавшей 2939 статей, при помощи программного обеспечения для работы с сетями WoS2Pajek и Pajek [6] была извлечена бимодальная (содержащая узлы двух типов) сеть совместной встречаемости ключевых слов в публикациях и самих публикаций. Бимодальная сеть была переведена в унимодальную, в которой в качестве узлов выступали только ключевые слова. Результатом стала сеть с 4350 ключевыми словами и 86 552 связями – совместными появлениями этих слов в публикациях. Далее из сети были удалены связи, наблюдавшиеся в выборке менее 20 раз[2]. Вместе с ними были усечены и одиночные узлы, которые оказались отрезаны от сети в результате этой операции. Затем были вручную удалены неинформативные узлы (такие как “algorithm”, “method”, “approach”, “analysis”, “technique”). Итоговым результатом обработки стала ненаправленная сеть, изображенная на *рис.* 1, включающая 46 узлов и 87 связей. Размер узла указывает на его степень – количество связей для данного узла; применительно к данной сети степень ключевого слова соответствует частоте его встречаемости с любым другим словом из сети.

При интерпретации полученной сети необходимо помнить, что для ее построения к ключевым словам применялся лемматизатор (инструмент, приводящий слова к начальной форме, что позво-

[1] Поскольку РИНЦ не позволяет производить выгрузку данных в формате, пригодном для анализа использованными в обзоре программными средствами, при алгоритмическом подходе была задействована только выборка, полученная из Web of Science.

[2] Имеется в виду, что совместная встречаемость двух ключевых слов составляла менее 20 случаев; данное число было выбрано в качестве порогового эмпирически, поскольку приводило к получению наиболее читаемой, но не слишком усеченной сети. Такой подход нетипичен для анализа библиографических сетей, однако иные опции (например, дифференциация узлов по степеням) не приводили к получению пригодной для интерпретации сети.

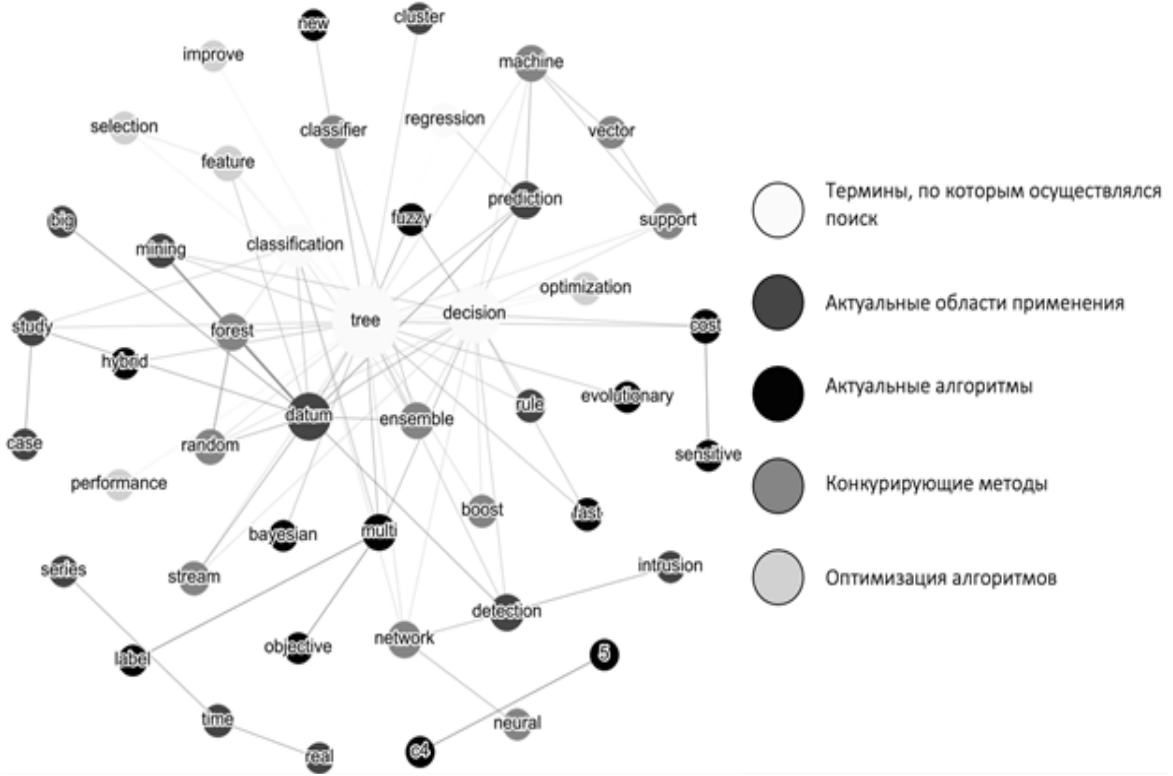

Рис. 1. Сеть ключевых слов из публикаций, касающихся деревьев решений, 2015–2019 гг.

ляет подсчитывать частоту возникновения слов вне зависимости от того, как они были использованы), а также о том, что ключевые слова не были преобразованы в словосочетания. Последний аспект обусловлен тем, что преобразование исходной базы потребовало бы не меньшей экспертизы, но значительно бóльших временных затрат, чем интерпретация сети по отдельным словам. Поиск же смысловых сочетаний в итоговой сети не представляет особых трудностей и достаточно однозначен. К примеру, связь узлов "big" и "datum" явным образом относится к большим данным ("big data"), а клика "support", "vector" и "machine" отсылает к методу опорных векторов. Анализ сети лег в основу структуры обзора, а выделенные алгоритмы и тематики позволили произвести дополнительную итерацию экспертного поиска литературы.

Обратимся непосредственно к сети. Ее узлы могут быть разделены на четыре смысловых кластера.

Актуальные алгоритмы: нечеткие деревья классификации ("fuzzy decision trees"); группа быстрых деревьев решений ("fast decision trees"); гибридные деревья ("hybrid trees"); многотемные деревья ("multi-label decision trees"), то есть деревья решений с пересекающимися классами целевой переменной; деревья решений, задействующие сразу несколько целевых переменных ("multi-objective decision trees"); чувствительные к издержкам деревья решений ("cost-sensitive decision trees") и алгоритм C4.5.

Оптимизация методов: преодоление смещений, связанных с выбором атрибута ("feature selection"); оптимизация ("optimization"); показатели качества ("performance").

Актуальные области применения: большие данные ("big data"); временные ряды ("time series"); кейс-стади ("case-study") и обнаружение вторжений ("intrusion detection") – поиск нехарактерных случаев, применяемый, к примеру, для отделения почтовых сообщений от спама. Первые две области имеют пересечение с социологическими исследованиями и будут рассмотрены ниже; кейс-стади и обнаружение вторжений не будут рассматриваться

подробнее, поскольку первое было и остается распространенным методом исследования для любых методов анализа данных, а второе хоть и может выступать подходом для решения определенных социологических проблем, но пока в социологии не применяется.

Конкурирующие методы интеллектуального анализа данных: метод опорных векторов (“support vector machine”); машинное обучение (“machine learning”), к которому, в том числе, относятся и сами деревья решений; ансамбли (“ensemble classifiers”), в частности – случайный лес (“random forest”) и бустинг (“boosting”).

Таким образом, в обзоре будут рассмотрены три темы повестки последних лет: актуальные алгоритмы, оптимизация методов и актуальные области применения. Четвертая тема – конкурирующие алгоритмы – не будет рассматриваться в рамках обзора. Для актуальных алгоритмов будут описаны направления их развития, новации относительно лучшей практики их применения и реализации в статистических пакетах; оптимизация методов будет рассмотрена с точки зрения проблем, выделенных в обзоре В. Ло [2]: проблемы смещений при выборе атрибута и разбиения узла, упрощения деревьев решений и работы с пропущенными данными; актуальные области применения будут описаны с точки зрения их пересечения со сферой социологических исследований и потенциала деревьев решений в этом контексте.

Актуальные алгоритмы деревьев решений

В этом разделе мы рассмотрим наиболее актуальные алгоритмы деревьев решений: C4.5, нечеткие и байесовские деревья, многотемные и многоцелевые деревья решений, а также деревья, учитывающие издержки.

Согласно полученной нами сети ключевых слов, одним из наиболее распространенных алгоритмов деревьев решений на сегодняшний день выступает C4.5 [7]. C4.5 демонстрирует стабильно приемлемые результаты, уступая различным ансамблям,

но превосходя другие распространенные алгоритмы единичных деревьев решений в простоте получаемого дерева, скорости и точности [8–11]. C4.5 реализован автором алгоритма в форме отдельного статистического инструмента в формате Open Source, а также в составе свободно распространяемого пакета *Weka* (с подробным описанием можно познакомиться в книге И. Виттена, Э. Фрэнка и М. Холла [12]), что делает его привлекательным за счет широкой доступности и простоты использования.

Благодаря возможностям современных компьютеров, распространение получили и более вычислительно затратные узкоспециализированные деревья, такие как *нечеткие деревья решений* (fuzzy decision trees, или FDT) [13–15], позволяющие работать с целевой переменной, элементы которой не принадлежат к одному из классов однозначно, но имеют определенную степень принадлежности к тому или иному классу, например – с лингвистическими категориями («холодный», «теплый» и «горячий» – не взаимоисключающие категории, с помощью которых можно описать ситуации «более или менее теплый», «скорее теплый, чем горячий» или «очень горячий»). Первое нечеткое дерево решений многотемной классификации было предложено в работе «Первый подход к использованию нечетких деревьев решений для многотемной классификации» [16]. Для социологических исследований инструменты, основанные на нечеткой логике, могут использоваться, например, для учета степени компетентности в экспертном опросе [17] или для представления социометрических данных [18]. За последние годы известные алгоритмы нечетких деревьев решений (к примеру, см.: [19–21]) были дополнены двумя новыми:

– дерево решений на основе нечеткого правила (Fuzzy Rule Based Decision Tree, или FRDT) [14], главной особенностью которого является использование сразу нескольких атрибутов для разбиения узла, что обеспечивает компактность дерева;

– GFID3 – нечеткий алгоритм ID3 (предшественник алгоритма C4.5) с использованием обобщенного энтропийного нечеткого

разбиения как критерия выбора атрибута [15], который способен делать поправку на возможную нелинейность функции принадлежности, что, по мнению авторов, лучше отражает реальный процесс принятия решений.

Говоря о практических рекомендациях применения нечетких деревьев, следует упомянуть исследование иранских ученых М. Зейналхани и М. Эфтехари, посвященное сравнению различных критериев остановки для этого типа алгоритмов [22]: согласно его результатам, предложенный авторами критерий NMGNI (нечеткий информационный выигрыш, помноженный на количество наблюдений в терминальном узле) превзошел критерии максимальной глубины, количества наблюдений в терминальном узле, нечеткого информационного выигрыша и нечеткой энтропии.

Главным ограничением для применения нечетких деревьев решений остается отсутствие имплементации в распространенных статистических пакетах. Нам не удалось найти опубликованных программных решений для применения подобных деревьев. Другие алгоритмы для работы с нечеткими множествами, основанные на деревьях решений, можно применить с помощью библиотек *frbs* [23] и *RoughSets* [24] для R. Соответственно, от исследователя требуются самостоятельное создание или адаптация нужного инструмента под конкретную задачу.

Байесовские деревья решений начали свое развитие еще в 1998 г. [25; 26]. Главное преимущество байесовских деревьев решений перед различными деревьями в русле частотной статистики – возможность работать с моделями, в которых число предикторов близко или превышает число наблюдений, недостаток же заключается в необходимости обосновывать применение методов пока еще мало распространенной байесовской статистики. Ряд проблем, характерных для этого типа деревьев, долгое время мешал их применению и распространениею. Источником этих проблем служит низкая сходимость цепей Маркова (распространенной техники байесовского вывода) применительно к байесовским деревьям,

а сами они заключаются в (а) недооценке неопределенности в апостериорных распределениях и (б) неизбежной переобученности получаемых деревьев [26, p. 886]. Сходимость – это способность цепи Маркова проходить через все возможные значения параметров и сходиться к стационарному распределению. Иными словами, в модели на основе цепи с низкой сходимостью некоторые значения признаков окажутся «неисследованными» цепью, а другие, напротив, будут пройдены чаще, чем того требует искомое распределение. В явном виде эти проблемы были выявлены еще в первых публикациях о методе, однако успешные попытки преодолеть их в отношении единичных деревьев были предприняты только в последние годы [28–30]. Два возможных решения этой проблемы предложены М. Пратолой [27]: процедура вращения дерева и моделирование «пространственной изменчивости» целевой переменной. Иными словами, это техники, проводящие цепь по тем вариантам структуры дерева, которые иначе были бы сочтены «невероятными». Оба решения демонстрируют улучшение качества цепей при относительно низких вычислительных затратах и при необходимости могут применяться вместе и в сочетании с другими известными способами получения приемлемых моделей байесовских регрессионных деревьев. Программное обеспечение для применения байесовских деревьев предоставляет исследователю достаточно широкий выбор: в среде для статистического анализа R реализованы пакеты *bayestree* [28], *BART* [28] и *bartMachine* [31] для алгоритма BART, а также *tgp* [32] для байесовского CART.

Отдельного внимания заслуживают многотемные[1] (*multi-label*) и многоцелевые (*multi-objective*) деревья решений, относящиеся к классу алгоритмов, решающих небинарные задачи классификации[2]

[1] Перевод заимствован у М.И. Петровского и В.В. Глазковой [68].

[2] К алгоритмам для решения небинарных задач классификации также относятся деревья для работы с порядковыми целевыми переменными или номинальными переменными, включающими больше двух классов, а также многовариантные

[32, p. 64]. Многотемные деревья решений предназначены для работы с целевыми переменными, классы которых не являются взаимоисключающими. Примером таких целевых переменных могут выступать анкетные вопросы, в которых респонденту предлагается отметить несколько или все подходящие варианты ответа. Стандартный подход к решению таких задач – один-против-всех, то есть классификация, производящаяся отдельно для каждого класса целевой переменной, позволяющая пользоваться готовыми классификаторами, однако совсем недавно начали возникать и специализированные инструменты для многотемной классификации [33–36]. Деревья решений для многотемной классификации реализованы в инструменте MULAN и пакете *utiml* для R [37].

Многоцелевые деревья решений [38] предназначены для работы с несколькими целевыми переменными одновременно. За последние годы были представлены деревья с частичным привлечением учителя (иными словами, использующие для классификации не только наблюдения с уже имеющимися значениями для зависимых переменных, но и наблюдения без этих значений) для многоцелевой регрессии [39]; сравнение многоцелевых алгоритмов с алгоритмами, объединяющими несколько одноцелевых решений, производилось в работе «Сравнение методов, основанных на деревьях решений, для многоцелевой регрессии на информационных потоках» [40], а возможное решение «близорукости» таких деревьев[1], отмеченное в этом исследовании, приводится в

деревья решений, работающие с «пакетами» наблюдений, для которых известен исход, но неизвестно, в каком из наблюдений пакета содержится предсказывающий исход атрибут.

[1] Под «близорукостью» алгоритмов машинного обучения подразумевается, что при выборе переменной или точки разбиения целесообразность этого выбора (в терминах соответствующего критерия, например – чистоты узла) оценивается только на данном ярусе дерева, а не с точки зрения оптимальности дерева целиком. Иными словами, дерево «видит» показатели качества только ближайших к текущему шагу возможных дочерних узлов, а не всех возможных терминальных.

работе «Предсказывающие деревья кластеризации для многоцелевой регрессии» [41]. Распространенное программное обеспечение для применения многоцелевых деревьев – система *CLUS* [42].

В 2014–2019 гг. внимание исследователей направлено также на развитие различных алгоритмов деревьев решений, учитывающих издержки (*cost-sensitive decision trees*). Под учетом издержек подразумевается возможность присуждения разных весов наблюдениям с разным значением целевой переменной. Например, банк понесет бóльшие издержки, если алгоритм, принимающий решение о выдаче кредита, классифицирует потенциально ненадежного заемщика как надежного и выдаст кредит, чем если потенциально надежный заемщик классифицируется как ненадежный и в кредите будет отказано. Если же решение будет принимать дерево, принимающее в расчет подобные издержки (например, неверная классификация потенциально ненадежного заемщика будет оцениваться им в три раза «дороже», чем неверная классификация потенциально надежного), то ненадежные заемщики будут классифицироваться точнее. Неправильно классифицировать одного ненадежного заемщика в этом случае для алгоритма будет настолько же критично, как ошибиться в трех надежных. Такие деревья применяются не только для непосредственного учета издержек мисклассификации, но и как способ преодоления смещений, связанных с несбалансированностью целевой переменной (об этом речь подробнее пойдет в следующем разделе; роль учета издержек здесь заключается в том, чтобы повысить «цену» неправильной классификации той категории, которая мало представлена в выборке). В частности, предложены: модификация алгоритма C4.5, учитывающая издержки [43]; деревья, позволяющие учитывать множество источников издержек одновременно [44]; нечеткие деревья решений, учитывающие издержки [45]; деревья, издержки для которых рассчитываются на основе обучающей выборки, а не задаются исследователем напрямую [46].

Оптимизация методов

Оптимизация алгоритмов деревьев решений происходит в двух основных направлениях: во-первых, это устранение смещений при выборе атрибута и расщеплении узла и, во-вторых, упрощение переобученных деревьев решений. Первое направление возникло в построенной нами сети ключевых слов (feature selection), а второе было добавлено на основе дискуссии к обзору В. Ло [2].

Проблема смещений при поиске атрибутов и расщеплении узла

Смещения при выборе атрибута (*attribute selection bias*) – это предпочтение алгоритмом для расщепления узла переменных с бóльшим количеством градаций. Нужно понимать, что устранение смещений в основном представляет собой итерационное исправление ошибок, обнаруженных в предыдущих версиях алгоритмов, и некоторые из них только предстоит обнаружить.

Существует несколько источников смещений: критерии выбора атрибута и расщепления как таковые; несоответствие между типом входных переменных и применяемых к ним критериев или преобразований; несбалансированность целевой переменной; наличие пропусков в данных.

Смещения, вызванные самим критерием выбора атрибута и расщепления узла, свойственны, в частности, распространенным «жадным» алгоритмам (*greedy algorithms*) – тем, которые принимают локально оптимальные решения при каждом расщеплении выборки [2]. Проблема смещений при выборе атрибута практически повсеместно поднимается при представлении нового алгоритма или его модификации, причем зачастую способы избегания смещений заимствуются у существующих алгоритмов, таких как GUIDE [47; 48] или деревьев условного вывода (*conditional inference trees*) [49]. Новый способ избежать таких смещений при помощи ограничения по классу предложен в работе «Выбор атрибута для

обучения деревья классификации с ограничением по классу» [50]. Получаемое дерево превосходит по интерпретируемости и показателям качества деревья, выращенные на тех же данных с помощью алгоритмов ID3, C4.5, REPTree и RandomTree [51, p. 22].

Смещения по причине несбалансированности целевой переменной свойственны всем распространенным алгоритмам, причем обычно в прикладных исследованиях эта проблема решается на уровне данных при помощи взвешивания или других методов ремонта выборки. Под несбалансированной переменной мы подразумеваем номинальную переменную, обладающую низкой вариацией: наблюдений, принадлежащих к одной категории этой переменной, значительно больше, чем принадлежащих к другим. Основной стратегией работы с несбалансированными данными выступают деревья, учитывающие издержки: мало представленной категории присуждается более высокая цена ошибки, чем категории, представленной адекватно, в результате чего дерево «вынуждено» принимать те решения, которые будут правильно классифицировать именно мало представленную категорию. Однако для случаев, когда отсутствует необходимая информация о предпосылках-издержках, были разработаны беззатратные деревья решений (*cost-free decision trees*), автоматически балансирующие мисклассификацию между классами целевой переменной [51]. Альтернативным решением здесь также могут выступить деревья решений, основанные на колеблющихся нечетких множествах, которые комбинируют сразу нескольких способов борьбы с несбалансированностью целевой переменной: на уровне входных данных и алгоритма, а также с помощью специального критерия выбора атрибута [52, p. 728].

Смещения при наличии пропусков проявляются в тех алгоритмах, которые работают с пропущенными значениями «как есть». Например, CART смещен относительно выбора атрибута в присутствии пропущенных значений: он склонен выбирать для расщепления узла переменные, содержащие больше пропусков, а

для суррогатного расщепления – содержащие меньше пропусков [53]. CHAID обнаруживает склонность к «порче» дерева (выбору не самого подходящего атрибута, неоптимальному расщеплению) в присутствии пропусков, в частности – если количество пропусков превышает 10% и атрибут с пропусками располагается близко к корню дерева [54]. На сегодняшний день открытыми остаются два аспекта, касающиеся работы с пропусками в контексте деревьев решений: 1) так же, как и при работе с любым другим методом анализа данных, перед деревьями стоит проблема определения степени случайности пропуска (пропуски бывают трех видов в зависимости от степени случайности: абсолютно случайные, случайные и систематические; подробнее о различиях между ними можно прочитать в статье Р. Литтла и Д. Рубина [55]); 2) неизученность соотношения между типом пропуска и наилучшим способом работы с ним. В частности, исследование влияния типа пропуска на случайные леса, использующие суррогатные переменные, показало, что, в отличие от других методов анализа данных, для деревьев решений и их ансамблей даже абсолютно случайные пропуски могут привести к систематическим смещениям в результатах [56]. С. Жучкова и А. Ротмистров оценили степень риска получения ложных выводов, рассмотрев работу с пропусками в предикторах «как есть», реализованную в алгоритме CHAID [54]. Авторы рассматривали деревья с разной степенью точности, разными долями пропусков и разным расположением предиктора с пропусками относительно корня дерева. В результате статистического эксперимента было установлено, что в целом алгоритм CHAID верно классифицирует наблюдения с пропусками, но в большинстве случаев пропуски вызывают изменения в структуре дерева [54, с. 104].

Упрощение деревьев решений

Как уже упоминалось, простота получаемого решения и его интерпретации – главное преимущество деревьев решений по

сравнению с другими методами классификации и предсказания, однако эти алгоритмы склонны к переобучению, то есть выращиванию слишком сложного и подходящего только для данной выборки дерева (иначе говоря, неприменимого к случаям, не покрытым выборкой). Существуют два способа борьбы с переобучением: останавливающие правила и прунинг – «обрезка» переобученных деревьев.

Алгоритмы, которые применяют основанные на проверке статистических гипотез правила остановки, могут пропускать важные эффекты взаимодействия при выращивании дерева [57]. К таким алгоритмам относятся, в частности, CTREE, QUEST и GUIDE. А. Альварес-Иглесиас и другие исследователи предложили простую модификацию алгоритма CTREE, позволяющую избегать ошибки такого рода: в предложенном ими алгоритме сначала выращивается насыщенное дерево (это производится с помощью увеличения уровня значимости для критерия остановки) и рассчитывается распределение p-значений с поправкой Бонферрони [57]. Затем С. Перейра и Р. Де Мелло предложили применять к насыщенному дереву прунинг, начиная с терминальных узлов дерева на основе поиска узлов, содержащих значимые различия [58, с. 94].

Таким образом, значительных изменений в области оптимизации деревьев классификации за последние годы не произошло. По-прежнему нет достаточно экспериментально и эмпирически обоснованных данных о поведении распространенных алгоритмов при наличии существенного количества пропусков и способов, позволяющих нивелировать этот эффект; проблемы смещений при расщеплении узла, в большинстве случаев, решаются заимствованными у сравнительно старых алгоритмов методами или не предлагают пользователю решения «из коробки», что существенно усложняет их применение на практике.

Актуальные области применения деревьев решений

В качестве актуальных областей применения деревьев решений мы выделили две сферы, имеющие отношение к социологическим исследованиям: большие данные и временные ряды, которые в значительной степени накладываются друг на друга, поскольку невозможность хранить постоянно возрастающие объемы данных приводит исследователей к необходимости интеллектуального анализа данных в реальном времени [58] (эта сфера представлена в сети узлами “real time”, “time series” и “data mining”).

Методы машинного обучения, в том числе и деревья решений, выступают одним из основных способов обработки и анализа больших данных [59–61]. Этому способствуют несколько факторов. Во-первых, методы машинного обучения лучше приспособлены к работе с массивами большого размера. Во-вторых, большие данные подлежат скорее разведывательному анализу, чем проверке гипотез: количество потенциальных предикторов, с одной стороны, ограничивает возможности привычных методов анализа данных, а с другой – открывает возможности для интеллектуального анализа данных (*data mining* и *data-driven research*). Деревья классификации как нельзя лучше справляются именно с такими задачами, осуществляя отбор релевантных предикторов. Наконец, в-третьих, базы данных большого объема позволяют искать различные нелинейные взаимодействия, на которые также направлены деревья решений [61; 62].

На сегодняшний день в социологических исследованиях деревья решений используются как для обработки, так и для анализа больших данных, в частности – в составе гибридных методов анализа. А. Кавеева и К. Гурин предлагают использовать алгоритмы деревьев решений для обработки больших данных, одним из характерных свойств которых является именно их неполнота. Идея авторов заключается в извлечении из имеющихся данных правил, которые предскажут значение пропущенной переменной [63].

Другим принципиальным свойством больших данных является их зашумленность [64, с. 63]. С учетом этой особенности была разработана модификация алгоритма C4.5 с заложенной в нее предпосылкой о частичной ненадежности данных – алгоритм Credal-C4.5, демонстрирующий более компактные и качественные деревья при обработке больших данных по сравнению с предшественником [65].

Основной вектор исследования временных рядов направлен на различные способы преобразования данных, чтобы к ним впоследствии можно было применить деревья решений и получить осмысленные результаты, например – вейвлет- и шейплет-преобразования [66; 67]. Вейвлет-преобразованием называют сложное преобразование временного ряда, основанное на вейвлетах – «маленьких волнах», коротких отрезках функции, подгоняемых под колебания ряда с помощью сжатия или растяжения этого отрезка. Такое преобразование позволяет очистить от шума и сократить массив информации, сохраняя при этом данные об изменениях характеристик ряда (среднее, дисперсия, период и т.д.) во времени. Шейплет-преобразования преследуют ту же цель, но вместо заранее заданных функций используют шейплеты – эталонные фрагменты самого временного ряда. Применение шейплет- или вейвлет-преобразований существенно повышает точность классификации временного ряда при помощи деревьев решений или других алгоритмов машинного обучения по сравнению с исходным.

Заключение

На основе анализа библиографической сети, построенной на ключевых словах, можно заключить, что наиболее актуальными направлениями развития алгоритмов деревьев решений за последние 5 лет стали следующие тематики:

– новые и распространенные алгоритмы деревьев решений: алгоритм C4.5, байесовские и нечеткие деревья решений, алгоритмы, чувствительные к издержкам;

– актуальные области применения деревьев решений, к которым можно отнести большие данные и работу с временными рядами;

– оптимизация существующих алгоритмов относительно возможностей работы с неполными данными, выращивания компактных деревьев или поиска новых способов их усечения, а также поиска и преодоления возможных источников смещения в результатах работы алгоритмов.

Не до конца изученными на сегодняшний день остаются следующие проблемы работы с деревьями классификации:

– применение алгоритмов к несбалансированным массивам на уровне алгоритма без операций над данными;

– соотношение степени случайности пропусков в данных и наилучшего способа работы с ними;

– проблема эффективного выбора того или иного метода построения дерева в зависимости от имеющихся данных и свойств самих алгоритмов.

Таким образом, в методологической литературе фокус смещается к более вычислительно затратным и узкоспециализированным алгоритмам деревьев решений и представлению программного обеспечения, поддерживающего эти методы, в формате Open Source.

Для широкой аудитории исследователей, безусловно, актуальными оказываются те алгоритмы, которые представлены в распространенных статистических пакетах или распространяются в виде простых в использовании отдельных инструментов с графическим интерфейсом. На сегодняшний день распространенные платные статистические пакеты предлагают пользователю только самые известные алгоритмы – CART, QUEST, CHAID. За более современными или специфичными алгоритмами приходится обращаться к библиотекам среды R, в подавляющем большинстве случаев не обладающим графическим интерфейсом и зачастую снабженным минимальной документацией. В этом отношении можно отметить тенденцию к повышению доступности алгоритмов

деревьев решений именно через последний вариант, позволяющий исследователю подобрать алгоритм, подходящий для конкретной исследовательской ситуации, без необходимости самостоятельно программировать дерево или обращаться к неоптимальным или неподходящим алгоритмам, реализованным в коммерческих статистических пакетах.

Выбор конкретного алгоритма опирается, в первую очередь, на задачу, стоящую перед исследователем: на сегодня деревья уже позволяют решать задачи, связанные с нечеткими массивами и временными рядами, а также адаптированы под байесовский подход к анализу данных; более того, большинство узкоспециализированных деревьев реализованы в виде сразу нескольких библиотек для R, перечисленных в этом обзоре, что делает их относительно доступными.

Если же данные не требуют специального подхода, следует проанализировать их качество:

– если целевая переменная несбалансирована, следует обращаться к беззатратным деревьям или деревьям, учитывающим издержки (в первом случае исследователю самостоятельно придется запрограммировать дерево – готовых решений на сегодняшний день не существует; последние же доступны в большинстве статистических пакетов);

– если предикторы сильно различаются по общему размаху (например, используются одновременно переменные возраста и пола), следует отдавать предпочтение несмещенным деревьям. Наиболее доступным из них является GUIDE, представленный в виде отдельного приложения с поддержкой большинства операционных систем и регулярно обновляемый создателем алгоритма[1]. Большинство более современных несмещенных деревьев осно-

[1] *Loh W.-Y.* GUIDE Classification and Regression Trees and Forests (version 36.2) // University of Wisconsin-Madison [site]. URL: http://pages.stat.wisc.edu/~loh/guide.html (date of access: 02.19.2021).

ваны на том же принципе, однако – в случае необходимости воспользоваться альтернативой – исследователь может воспользоваться деревьями с ограничением по классу, которые нужно запрограммировать самостоятельно, обратившись к алгоритму, описанному в статье «Выбор атрибута для деревьев решений с ограничением по классу» [50];

– если данные, с которыми имеет дело исследователь, содержат большое количество пропусков, мы рекомендуем отказаться от использования деревьев решений ввиду слабой изученности их поведения при наличии пропусков и обратиться к ансамблям деревьев решений.

В ситуации разрозненности литературы о деревьях решений и переизбытка самих алгоритмов задача выбора подходящего метода выращивания дерева становится достаточно нетривиальной. Помимо этого, многие распространенные статистические пакеты предлагают пользователю несколько устаревшие методы с известной склонностью к выращиванию смещенных деревьев, такие как CART, C4.5 и CHAID. Средствами выхода из этой ситуации в краткосрочном периоде нам видятся регулярные обзоры и сравнение новейших алгоритмов применительно к разным исследовательским ситуациям, а также изучение поведения алгоритмов в зависимости от различных несовершенств в данных и составление рекомендаций по их использованию, которые позволят снизить риск получения ложных результатов. В долгосрочной перспективе, безусловно, наиболее привлекательным выступает решение, предложенное в дополнении Т. Руша и А. Зилиса к обзору В. Ло [3]: унификация языка описания алгоритмов и создание консистентного программного обеспечения в формате Open Source.

ЛИТЕРАТУРА

1. *Жучкова С.В.* Поиск многомерной связи категориальных признаков: сравнение CHAID, логлинейного анализа и множественного анализа соответствий / С.В. Жучкова, А.Н. Ротмистров // Мониторинг общественного мнения: экономические и социальные перемены. 2019. № 2. С. 32–53.

2. *Loh W.* Fifty Years of Classification and Regression Trees // International Statistical Review. 2014. Vol. 82. No. 3. P. 329–348.

3. *Rusch T.* Discussion of Fifty Years of Classification and Regression Trees / T. Rusch, A. Zeileis // International Statistical Review. 2014. Vol. 82. No. 3. P. 361–367.

4. *Morgan J.* Problems in the Analysis of Survey Data, and a Proposal / J. Morgan, J. Sonquist // Journal of American Statistical Association. 1963. Vol. 58. No. 302. P. 415–434.

5. *Моисеев С.П.* Отбор источников для систематического обзора литературы: сравнение экспертного и алгоритмического подходов / С.П. Моисеев, Д.В. Мальцева // Социология: методология, методы, математическое моделирование. 2018. № 47. С. 7–43.

6. *Batagelj V.* Pajek: Program for Large Network Analysis / V. Batagelj, A. Mrvar // Connections. 1998. Vol. 21. No. 2. P. 47–57.

7. *Quinlan* J.R. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 1993.

8. An Up-to-date Comparison of State-of-the-art Classification Algorithms / C. Liu, X. Zhang, G. Almpanidis // Expert Systems with Applications. Vol. 82. P. 128–150.

9. *King R.* Statlog: Comparison of Classification Algorithms on Large Real-world Problems / R. King, C. Feng, A. Sutherland // Applied Artificial Intelligence. 1995. Vol. 9. No. 3. P. 289–333.

10. *Lim T.* Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms / T. Lim, W. Loh, Y. Shih // Machine Learning. 2000. Vol. 40. No. 3. P. 203–228.

11. *Бильгаева Л.П.* Исследование моделей деревьев решений в задаче классификации / Л.П. Бильгаева, В.В. Ларин, Д.А. Маслюк // Экспериментальные и теоретические исследования в XXI веке: проблемы и перспективы развития. Материалы XIII Всероссийской научно-практической конференции. Ростов н/Д.: ИУБиП, 2018. Р. 38–51.

12. *Witten I.* Data Mining: Practical Machine Learning Tools and Techniques / I. Witten, E. Frank, M. Hall. 3rd ed. San Francisco: Morgan Kaufmann Publishers Inc., 2011.

13. *Trabelsi A.* Decision Tree Classifiers for Evidential Attribute Values and Class Labels / A. Trabelsi, Z. Elouedi, E. Lefevre // Fuzzy Sets and Systems. 2019. Vol. 366. P. 46–62.

14. Fuzzy Rule Based Decision Trees / X. Wang, X. Liu, W. Pedrycz, I. Zhang // Pattern Recognition. 2015. Vol. 48. No. 1. P. 50–59.

15. *Jin C.* A Generalized Fuzzy ID3 Algorithm Using Generalized Information Entropy / C. Jin, F. Li, Y. Li // Knowledge-Based Systems. 2014. Vol. 64. P. 13–21.

16. *Prati R.* A First Approach towards a Fuzzy Decision Tree for Multilabel Classification / R. Prati, F. Charte, F. Herrera // 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Naples, 2017. P. 1–6.

17. *Баранов Л.Т.* Нечеткие множества в экспертном опросе / Л.Т. Баранов, А.И. Птушкин, А.В. Трудов // Социология: методология, методы, математическое моделирование. 2004. № 19. С. 142–157.

18. *Мухатдинова О.Р.* Построение и анализ социограмм на основе нечеткой логики // Социология: методология, методы, математическое моделирование. 2000. № 12. Р. 154–172.

19. *Chang R.* Fuzzy Decision Tree Algorithms / R. Chang, T. Pavlidis // IEEE Transactions on Systems, Man, and Cybernetics. 1977. Vol. 7. No. 1. P. 28–35.

20. *Olaru C.* A Complete Fuzzy Decision Tree Technique / C. Olaru, L. Wehenkel // Fuzzy Sets and Systems. 2003. Vol. 138. No. 2. P. 221–254.

21. *Cintra M.* A Fuzzy Decision Tree Algorithm Based on C4.5 / M. Cintra, M. Monard, H. Camargo // Mathware & Soft Computing. Vol.20. No. 1. 2013. P. 56–62.

22. *Zeinalkhani M.* Comparing Different Stopping Criteria for Fuzzy Decision Tree Induction through IDFID3 / M. Zeinalkhani, M. Eftekhari // Iranian Journal on Fuzzy Systems. 2014. Vol. 11. No. 1. P. 27–48.

23. *Bergmeir C.* frbs: Fuzzy Rule-based Systems for Classification / C. Bergmeir, M. Ben // Journal of Statistical Software. 2015. Vol. 65. No. 6. P. 1–30.

24. Implementing Algorithms of Rough Set Theory and Fuzzy Rough Set Theory in the R Package "roughSets" / L. Riza, A. Janusz, C. Bergmeir [et al.] // Information Sciences. 2014. Vol. 287. P. 68–89.

25. *Chipman H.* Bayesian CART Model Search / H. Chipman, E. George, R. McCulloch // Journal of American Statistical Association. 1998. Vol. 93. No. 443. P. 935–948.

26. *Denison D.* A Bayesian CART Algorithm / D. Denison, B. Mallick, A. Smith // Biometrika. 1998. Vol. 85. No. 2. P. 363–377.

27. *Pratola M.* Efficient Metropolis-Hastings Proposal Mechanisms for Bayesian Regression Tree Models // Bayesian Analysis. 2016. Vol. 11. No. 3. P. 885–911.

28. Parallel Bayesian Additive Regression Trees / M. Pratola, H. Chipman, J. Gattiker [et al.] // Journal of Computational and Graphical Statistics. 2014. Vol. 23. No. 3. P. 830–852.

29. *Linero A.* Bayesian Regression Trees for High-dimensional Prediction and Variable Selection // Journal of American Statistical Association. 2018. Vol. 112. No. 522. P. 626–636.

30. *Xu D.* A Bayesian Nonparametric Approach to Causal Inference on Quantiles / D. Xu, M. Daniels, A. Winterstein // Biometrics. 2018. Vol. 74. No. 3. P. 986–996.

31. *Kapelner A.* bartMachine: Machine Learning with Bayesian Additive Regression Trees / A. Kapelner, J. Bleich // Journal of Statistical Software. 2013. Vol. 70. No. 4. P. 1–40.

32. *Gramacy R.* tgp: An R Package for Bayesian Nonstationary, Semiparametric Nonlinear Regression and Design by Treed Gaussian Process Models // Journal of Statistical Software. 2007. Vol. 19. No. 9. P. 1–46.

33. *Gibaja E.* A Tutorial on Multilabel Learning / E. Gibaja, S. Ventura // ACM Computing Surveys. 2015. Vol. 47. No. 3. P. 1–38.

34. *Li P.* An Incremental Decision Tree for Mining Multilabel Data / Li P., X. Wu, X. Hu, H. Wang // Applied Artificial Intelligence. 2015. Vol. 29. No. 10. P. 992–1014.

35. *Bi W.* Bayes-optimal Hierarchical Multilabel Classification / W. Bi, J. Kwok // IEEE Transactions on Knowledge and Data Engineering. 2015. Vol. 27. No. 11. P. 2907–2918.

36. An Extensive Evaluation of Decision Tree-based Hierarchical Multilabel Classification Methods and Performance Measures / R. Cerri, G. Pappa, A. Carvalho, A. Freitas // Computational Intelligence. 2015. Vol. 31. No. 1. P. 1–46.

37. *Rivolli A.* The Utiml Package: Multi-label Classification in R / A. Rivolli, A. de Carvalho // R Journal. 2019. Vol. 10. No. 1. 2. P. 24-37.

38. *Blockeel H.* Top-down Induction of Clustering Trees / H. Blockeel, L. de Raedt, J. Ramon // ICML '98 Proceedings of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 1998. P. 55–63.

39. Semi-supervised Trees for Multi-target Regression / J. Levatić, D. Kocev, M. Ceci, S. Dzeroski // Information Sciences. 2018. Vol. 450. P. 109–127.

40. *Osojnik A.* Comparison of Tree-based Methods for Multi-target Regression on Data Streams / A. Osojnik, P. Panov, S. Dzeroski // Proceedings of the 4th International Conference on New Frontiers in Mining Complex Patterns (NFMCP'15). Berlin: Springer International Publishing. 2015. P. 17–31.

41. *Osojnik A.* Option Predictive Clustering Trees for Multi-target Regression / A. Osojnik, S. Dzeroski, D. Kocev // Computer Science and Information Systems. 2017. Vol. 17. No. 2. P. 118–133.

42. *Blockeel H.* Efficient Algorithms for Decision Tree Cross-validation / H. Blockeel, J. Struyf // Journal of Machine Learning Research. 2003. Vol. 3. No. 4–5. P. 621–650.

43. *Lee J.* AUC4.5: AUC-based C4.5 Decision Tree Algorithm for Imbalanced Data Classification // IEEE Access. 2019. Vol. 7. P. 106034–106042.

44. *Wu C.* Cost-sensitive Decision Tree with Multiple Resource Constraints / C. Wu, Y. Chen, K. Tang // Applied Intelligence. 2019. Vol. 49. No. 10. P. 3765–3782.

45. A Compact Evolutionary Interval-valued Fuzzy Rule-based Classification System for the Modeling and Prediction of Real-world Financial Applications with Imbalanced Data / J. Sanz, D. Bernardo, F. Herrera [et al.] // IEEE Transactions on Fuzzy Systems. 2015. Vol. 23. No. 4. P. 973–990.

46. *Bahnsen A.C.* Example-dependent Cost-sensitive Decision Trees / A.C. Bahnsen, D. Aouada, B. Ottersten // Expert Systems with Applications. 2015. Vol. 42. No. 19. P. 6609–6619.

47. *Fu W.* Unbiased Regression Trees for Longitudinal and Clustered Data / W. Fu, J. Simonoff // Computational Statistics and Data Analysis. 2015. Vol. 88. P. 53–74.

48. *Kim J.* Seemingly Unrelated Regression Tree / J. Kim, H. Cho // Journal of Applied Statistics. 2019. Vol. 46. No. 7. P. 1177–1195.

49. *Hothorn T.* Unbiased Recursive Partitioning: A Conditional Inference Framework / T. Hothorn, K. Hornik, A. Zeileis // Journal of Computational and Graphical Statistics. 2006. Vol. 15. No. 3. P. 651–674.

50. *Sun H.* Attribute Selection for Decision Tree Learning with Class Constraint / H. Sun, X. Hu // Chemometrics and Intelligent Laboratory Systems. 2017. Vol. 163. P. 16–23.

51. *Zhang X.* A New Strategy of Cost-free Learning in the Class Imbalance Problem / X. Zhang, B. Hu // IEEE Transactions on Knowledge and Data Engineering. 2014. Vol. 26. No. 12. P. 2872–2885.

52. *Sardari S.* Hesitant Fuzzy Decision Tree Approach for Highly Imbalanced Data Classification / S. Sardari, M. Eftekhari, F. Afsari // Applied Soft Computing Journal. 2017. Vol. 61. P. 727–741.

53. *Kim H.* Classification Trees with Unbiased Multiway Splits / H. Kim, W. Loh // Journal of American Statistical Association. 2009. Vol. 96. No. 454. P. 589–604.

54. Жучкова С.В. Возможность работы с пропущенными данными при использовании CHAID: результаты статистического эксперимента / С.В. Жучкова, А.Н. Ротмистров // Социология: методология, методы, математическое моделирование. 2018. № 46. P. 85–122.

55. *Little R.* The Analysis of Social Science Data with Missing Values / R. Little, D. Rubin // Sociological Methods and Research. 1989. Vol. 18. P. 292–326.

56. A New Variable Importance Measure for Random Forests with Missing Data / A. Hapfelmeier, T. Hothorn, K. Ulm, C. Strobl // Statistics and Computing. 2014. Vol. 24. No. 1. P. 21–34.

57. An Alternative Pruning Based Approach to Unbiased Recursive Partitioning / A. Alvarez-Iglesias, J. Hinde, J. Ferguson, J. Newell // Computational Statistics and Data Analysis. 2017. Vol. 106. P. 90–102.

58. *Pereira C.* TS-stream: Clustering Time Series on Data Streams / C. Pereira, R. de Mello // Journal of Intelligent Information Systems. 2014. No. 42. P. 531–566.

59. *Gomes C.* Presenting the Regression Tree Method and its Application in a Large-scale Educational dataset / C. Gomes, E. Jelihovschi // International Journal of Research & Method in Education. 2020. Vol. 43. No. 2. P. 201–221.

60. *Sorensen L.* "Big Data" in Educational Administration: An Application for Predicting School Dropout Risk // Educational Administration Quaterly. 2019. Vol. 55. No. 3. P. 404–446.

61. *Губа К.* Большие данные в социологии: новые данные, новая социология? // Социологическое обозрение. 2018. Т. 17. № 1. P. 213–236.

62. *Varian H.* Big Data: New Tricks for Econometrics // Journal of Economic Perspectives. 2014. Vol. 28. No. 2. P. 3–28.

63. *Prati R.* Emerging Topics and Challenges of Learning from Noisy Data in Nonstandard Classification: a Survey beyond Binary Class Noise / R. Prati, J. Luengo, F. Herrera // Knowledge and Information Systems. 2018. Vol. 60. No. 1. P. 1–35.

64. *Кавеева А.Д.* Локальные сети дружбы «ВКонтакте»: восстановление пропущенных данных о городе проживания пользователей / А.Д. Кавеева, К.Е. Гурин // Мониторинг общественного мнения: экономические и социальные перемены. 2018. Т. 3. № 145. Р. 78–90.

65. *Mantas C.* Credal-C4.5: Decision Tree Based on Imprecise Probabilities to Classify Noisy Data / C. Mantas, J. Abellán // Expert Systems with Applications. 2014. Vol. 41. No. 10. P. 4625–4637.

66. Classification Tree Methods for Panel Data Using Wavelet-transformed Time Series / X. Zhao, S. Barber, C. Taylor, Z. Milan // Computational Statistics and Data Analysis. 2018. Vol. 127. P. 204–216.

67. Classification of Time Series by Shapelet Transformation / J. Hills, J. Lines, E. Baranauskas [et al.] // Data Mining and Knowledge Discovery. 2014. Vol. 28. P. 851–881.

68. *Петровский М.И.* Метод многотемной (multi-label) классификации на основе попарных сравнений с отсечением наименее релевантных классов / М.И. Петровский, В.В. Глазкова // Математические методы распознавания образов: 13-я Всероссийская конференция. Т. 13. М.: МАКС Пресс, 2007. С. 197–200.

Suleymanova Anna,
National Research University Higher School of Economics (NRU HSE), Moscow, asuleymanova@hse.ru

An overview of the development of the decision tree algorithms

Classification trees are classification and prediction algorithms, common for empirical studies due to the simplicity of application and interpretation. Nevertheless, selection of the specific algorithm for the task is not a trivial objective, as the literature on the topic and software are quite scattered. As a result, researchers stick to familiar and long-used algorithms despite the seeming variety and specification. This review aims to reveal and structure development directions that emerged in the past five years. We apply both expert and algorithmic approaches to relevant literature selection. In particular, a network of keywords is applied to reveal the current topics of interest. The review allows one to navigate the growing overabundance of algorithms and elaborates previous reviews.
Keywords: decision tree, classification and regression tree, bibliographical analysis, keyword network analysis, big data.

References

1. Zhuchkova S.V., Rotmistrov A.N. In search of multivariate associations: comparison of CHAID, log-linear analysis, and multiple correspondence analysis (in Russian), *Monitoring Obshchestvennogo Mneniya: Ekonomicheskie i Sotsial'nye Peremeny (Monitoring of Public Opinion)*, 2019, 2, 32–53.
2. Loh W. Fifty years of classification and regression trees, *International Statistical Review*, 2014, 82 (3), 329–348.
3. Rusch T., Zeileis A. Discussion of Fifty years of classification and regression trees, *International Statistical Review*, 2014, 82 (3), 361–367.
4. Morgan J., Sonquist J. Problems in the Analysis of Survey Data, and a Proposal, *Journal of American Statistical Association*, 1963, 58 (302), 415–434.
5. Moiseev S.P., Maltseva D.V. Selection of sources for a systematic literature review: comparison of expert and algorithmic approaches (in Russian), *Sotsiologiya 4M / Sociology: methodology, methods, mathematical modeling*, 2018, 47, 7–43.

6. Batagelj V., Mvar A. Pajek: Program for large network analysis, *Connections*, 1998, 21 (2), 47–57.
7. Quinlan J.R. *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc., 1993.
8. Liu C., Zhang X., Almpanidis G. An up-to-date comparison of state-of-the-art classification algorithms, *Expert Systems with Applications*, 2017, 82, 128–150.
9. King R., Feng C., Sutherland A. Statlog: Comparison of classification algorithms on large real-world problems, *Applied Artificial Intelligence*, 1995, 9 (3), 289–333.
10. Lim T., Loh W., Shih Y. Comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, *Machine Learning*, 2000, 40 (3), 203–228.
11. Bilgaeva L.P., Larin V.V., Maslyuk D.A. A study of decision tree models for classification problems (in Russian), *Eksperimental'nie i teoreticheskie issledovaniya v XXI veke: problemy i perspektivy razvitiya (XIII All-Russian Scientific and Practical Conference proceedings)*, Rostov-on-Don: IMBL, 2018. P. 38–51.
12. Witten I., Frank E., Hall M. *Data Mining: practical machine learning tools and techniques.* 3rd ed. San Francisco: Morgan Kaufmann Publishers Inc., 2011.
13. Trabelsi A., Elouedi Z., Lefevre E. Decision tree classifiers for evidential attribute values and class labels, *Fuzzy Sets and Systems*, 2019, 366, 46–62.
14. Wang X., Liu X., Pedrycz W., Zhang I. Fuzzy rule based decision trees, *Pattern Recognition*, 2015, 48 (1), 50–59.
15. Jin C., Li F., Li Y. A generalized fuzzy ID3 algorithm using generalized information entropy, *Knowledge-Based Systems*, 2014, 64, 13–21.
16. Prati R., Charte F., Herrera F. A first approach towards a fuzzy decision tree for multilabel classification, *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Naples, 2017. P. 1–6.
17. Baranov L.T., Ptushkin A.I., Trudov A.V. Fuzzy sets in expert survey (in Russian), *Sotsiologiya 4M / Sociology: methodology, methods, mathematical modeling*, 2004, 19, 142–157.
18. Muhatdinova O.R. "Plotting and analysis of sociogram based on fuzzy logic" (in Russian), *Sotsiologiya 4M / Sociology: methodology, methods, mathematical modeling*, 2000 (12), 154–172.

19. Chang R., Pavlidis T. Fuzzy Decision Tree Algorithms, *IEEE Transactions on Systems, Man, and Cybernetics*, 1977, 7 (1), 28–35.
20. Olaru C., Wehenkel L. A complete fuzzy decision tree technique, *Fuzzy Sets and Systems*, 2003, 138 (2), 221–254.
21. Cintra M., Monard M., Camargo H. A fuzzy decision tree algorithm based on C4.5, *Mathware & Soft Computing,* 2013, 20 (1), 56–62.
22. Zeinalkhani M., Eftekhari M. Comparing different stopping criteria for fuzzy decision tree induction through IDFID3, *Iranian Journal on Fuzzy Systems*, 2014, 11 (1), 27–48.
23. Bergmeir C., Ben M. frbs: Fuzzy rule-based systems for classification, *Journal of Statistical Software*, 2015, 65 (6), 1–30.
24. Riza L., Janusz A., Bergmeir C. [et al.] Implementing algorithms of rough set theory and fuzzy rough set theory in the R package 'roughSets', *Information Sciences*, 2014, 287, 68–89.
25. Chipman H., George E., McCulloch R. Bayesian CART model search, *Journal of American Statistical Association*, 1998, 93 (443), 935–948.
26. Denison D., Mallick B., Smith A. A Bayesian CART algorithm, *Biometrika*, 1998, 85 (2), 363–377.
27. Pratola M. Efficient Metropolis-Hastings proposal mechanisms for Bayesian regression tree models, *Bayesian Analysis*, 2016, 11 (3), 885–911.
28. Pratola M., Chipman H., Gattiker J. et al. Parallel Bayesian additive regression trees, *Journal of Computational and Graphical Statistics*, 2014, 23 (3), 830–852.
29. Linero A. Bayesian regression trees for high-dimensional prediction and variable selection, *Journal of American Statistical Association*, 2018, 112 (522), 626–636.
30. Xu D., Daniels M., Winterstein A. A Bayesian nonparametric approach to causal inference on quantiles, *Biometrics*, 2018, 74 (3), 986–996.
31. Kapelner A., Bleich J. bartMachine: Machine learning with Bayesian additive regression trees, *Journal of Statistical Software*, 2013, 70 (4), 1–40.
32. Gramacy R. tgp: An R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models, *Journal of Statistical Software*, 2007, 19 (9), 1–46.
33. Gibaja E., Ventura S. A tutorial on multilabel learning, *ACM Computing Surveys*, 2015, 47 (3), 1–38.

34. Li P., Wu X., Hu X., Wang H. An incremental decision tree for mining multilabel data, *Applied Artificial Intelligence*, 2015, 29 (10), 992–1014.
35. Bi W., Kwok J. Bayes-optimal hierarchical multilabel classification, *IEEE Transactions on Knowledge and Data Engineering*, 2015, 27 (11), 2907–2918.
36. Cerri R., Pappa G., Carvalho A., Freitas A. An extensive evaluation of decision tree-based hierarchical multilabel classification methods and performance measures, *Computational Intelligence*, 2015, 31 (1), 1–46.
37. Rivolli A., de Carvalho A. The utiml package: multi-label classification in R, *R Journal*, 2019, 10 (2), 24–37.
38. Blockeel H., de Raedt L., Ramon J. Top-down induction of clustering trees, *ICML '98 Proceedings of the 15th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc., 1998. P. 55–63.
39. Levatić J., Kocev D., Ceci M., Dzeroski S. Semi-supervised trees for multi-target regression, *Information Sciences*, 2018, 450, 109–127.
40. Osojnik A., Panov P., Dzeroski S. Comparison of tree-based methods for multi-target regression on data streams, *Proceedings of the 4th International Conference on New Frontiers in Mining Complex Patterns (NFMCP'15)*. Berlin: Springer International Publishing, 2015. P. 17–31.
41. Osojnik A., Dzeroski S., Kocev D. Option predictive clustering trees for multi-target regression, *Computer Science and Information Systems*, 2017, 17 (2), 118–133.
42. Blockeel H., Struyf J. Efficient algorithms for decision tree cross-validation, *Journal of Machine Learning Research*, 2003, 3 (4–5), 621–650.
43. Lee J. AUC4.5: AUC-based C4.5 decision tree algorithm for imbalanced data classification, *IEEE Access*, 2019, 7, 106034–106042.
44. Wu C., Chen Y., Tang K. Cost-sensitive decision tree with multiple resource constraints, *Applied Intelligence*, 2019, 49 (10), 3765–3782.
45. Sanz J., Bernardo D., Herrera F., Bustince H., Hagras H. A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data, *IEEE Transactions on Fuzzy Systems*, 2015, 23 (4), 973–990.
46. Bahnsen A.C., Aouada D., Ottersten B. Example-dependent cost-sensitive decision trees, *Expert Systems with Applications*, 2015, 42 (19), 6609–6619.

47. Fu W., Simonoff J. Unbiased regression trees for longitudinal and clustered data, *Computational Statistics and Data Analysis*, 2015, 88, 53–74.
48. Kim J., Cho H. Seemingly unrelated regression tree, *Journal of Applied Statistics*, 2019, 46 (7), 1177–1195.
49. Hothorn T., Hornik K., Zeileis A. Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics*, 2006, 15 (3), 651–674.
50. Sun H., Hu X. Attribute selection for decision tree learning with class constraint, *Chemometrics and Intelligent Laboratory Systems*, 2017, 163, 16–23.
51. Zhang X., Hu B. A new strategy of cost-free learning in the class imbalance problem, *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26 (12), 2872–2885.
52. Sardari S., Eftekhari M., Afsari F. Hesitant fuzzy decision tree approach for highly imbalanced data classification, *Applied Soft Computing Journal*, 2017, 61, 727–741.
53. Kim H., Loh W. Classification trees with unbiased multiway splits, *Journal of American Statistical Association*, 2009, 96 (454), 589–604.
54. Zhuchkova S.V., Rotmistrov A.N. Handling missing data with CHAID: results of a statistical experiment (in Russian), Sotsiolo*giya 4M / Sociology: methodology, methods, mathematical modeling*, 2018 (46), 85–122.
55. Little R., Rubin D. The analysis of social science data with missing values, *Sociological Methods and Research*, 1989, 18, 292–326.
56. Hapfelmeier A., Hothorn T., Ulm K., Strobl C. A new variable importance measure for random forests with missing data, *Statistics and Computing*, 2014, 24 (1), 21–34.
57. Alvarez-Iglesias A., Hinde J., Ferguson J., Newell J. An alternative pruning based approach to unbiased recursive partitioning, *Computational Statistics and Data Analysis*, 2017, 106, 90–102.
58. Pereira C., de Mello R. TS-stream: clustering time series on data streams, *Journal of Intelligent Information Systems*, 2014 (42), 531–566.
59. Gomes C., Jelihovschi E. Presenting the regression tree method and its application in a large-scale educational dataset, *International Journal of Research & Method in Education*, 2020, 43 (2), 201–221.

60. Sorensen L. 'Big Data' in Educational Administration: An Application for Predicting School Dropout Risk, *Educational Administration Quaterly*, 2019, 55 (3), 404–446.
61. Guba K. Big Data in Sociology: New Data, New Sociology? (in Russian), *Sotsiologicheskoe Obozrenie / The Russian Sociological Review*, 2018, 17 (1), 213–236.
62. Varian H. Big data: New tricks for econometrics, *Journal of Economic Perspectives*, 2014, 28 (2), 3–28.
63. Prati R., Luengo J., Herrera F. Emerging topics and challenges of learning from noisy data in nonstandard classification: a survey beyond binary class noise, *Knowledge and Information Systems*, 2018, 60 (1), 1–35.
64. Kaveeva A.D., Gurin K.E. Local friendship networks in Vkontakte: Reconstruction of missing user's city information" (in Russian), *Monitoring Obshchestvennogo Mneniya: Ekonomicheskie i Sotsial'nye Peremeny / Monitoring of Public Opinion*, 2018, 3 (145), 78–90.
65. Mantas C., Abellán J. Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data, Expert *Systems with Applications*, 2014, 41 (10), 4625–4637.
66. Zhao X., Barber S., Taylor C., Milan Z. Classification tree methods for panel data using wavelet-transformed time series, *Computational Statistics and Data Analysis*, 2018, 127, 204–216.
67. Hills J, Lines J., Baranauskas E., Mapp J., Bagnall A. Classification of time series by shapelet transformation, *Data Mining and Knowledge Discovery*, 2014, 28, 851–881.
68. Petrovsky M., Glazkova V. Multi-label classification method based on pairwise comparisons with least relevant classes clipping (in Russian), *Mathematical Methods of Image Recognition*, XIII All-Russian Conference, 2007, 13, 197–200.