На правах рукописи

Девяткин Дмитрий Алексеевич

Построение ансамблей деревьев решений с использованием линейных и нелинейных разделителей

Специальность 2.3.5 (05.13.11) —
«Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва, 2022

Работа выполнена в федеральном государственном бюджетном учреждении науки «Федеральный исследовательский центр “Информатика и управление” Российской академии наук»

Научный руководитель:	**Соченков Илья Владимирович,** кандидат физико-математических наук
Официальные оппоненты:	**Вохминцев Александр Владиславович,** доктор технических наук, Федеральное государственное бюджетное образовательное учреждение высшего образования "Челябинский государственный университет", институт информационных технологий, профессор кафедры информационных технологий и экономической информатики **Матвеев Сергей Александрович,** кандидат физико-математических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования Московский государственный университет им. М.В. Ломоносова, факультет вычислительной математики и кибернетики, ученый секретарь кафедры вычислительных технологий и моделирования
Ведущая организация:	Федеральное государственное автономное образовательное учреждение высшего образования "Российский университет дружбы народов"

Защита состоится 15 декабря 2022 г. в 16 часов на заседании диссертационного совета 24.1.120.01 при Федеральном государственном бюджетном учреждении науки Институт системного программирования им. В.П. Иванникова РАН по адресу: 109004, г. Москва, ул. А. Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Институте системного программирования им. В. П. Иванникова РАН.

Автореферат разослан «____» _____ 2022 года.

Ученый секретарь
диссертационного совета
24.1.120.01,
кандидат физико-математических наук Зеленов С.В.

Общая характеристика работы

Актуальность темы.

Методы на основе дифференцируемых функций, такие как нейронные сети, используются для решения задач анализа структурированных данных, текстов и изображений. В некоторых случаях применимость этих методов ограничена: например, в исследовании A. Ek эмпирически показано, что незначительные изменения в тексте могут приводить к существенным изменениям его тональности и смысла, которые не всегда выявляются многослойными нейронными сетями, в том числе с архитектурой «Трансформер». В работах M. Tanchik и др., Z. Khan и др. на примерах восстановления изображений МРТ и оценки вегетационных индексов по данным аэрофотосъёмки показано, что многослойные нейронные сети не позволяют решать задачи регрессионного анализа с приемлемой точностью, если моделируемая зависимость имеет негладкий характер. В подобных задачах существенное значение имеет дискретная природа составляющих частей анализируемых объектов в контексте применения методов машинного обучения. Между тем, решение этих задач имеет большую практическую значимость в различных отраслях экономики, позволяя снизить стоимость, либо автоматизировать отдельные этапы технологических процессов.

Большей точности решения в таких случаях можно было бы достигнуть при применении моделей машинного обучения с дискретными зависимыми переменными. В основе таких моделей могут лежать деревья решений и их композиции (ансамбли). Так, в исследованиях Z. Zhou, J. Ren, Y. Chen, Л. Уткина, предложены каскадные композиции случайных лесов деревьев решений, на некоторых задачах превосходящие по качеству анализа методы на основе нейронных сетей. Однако деревья решений имеют ограниченную выразительную способность при фиксированной высоте, так как количество учитываемых признаков определяется числом узлов дерева. Они также характеризуются низкой вычислительной эффективностью при анализе данных большой размерности. Одним из подходов к решению этой проблемы является обучение деревьев решений с многомерными разделителями в узлах, например, с наклонными гиперплоскостями (oblique trees). Исследования S. Murphy, B. Menze и других показали, что подобные деревья решений имеют низкую обобщающую способность (то есть способность формировать корректные результаты для объектов, не использовавшихся при обучении), поэтому они применимы только в составе рандомизированных композиций, построенных методами бэггинга, со-ббэггинга или случайного леса. Кроме того, большинство подходов к построению таких деревьев имеют низкую вычислительную эффективность и большое количество гиперпараметров. При использовании таких композиций также могут наблюдаться эффекты,

связанные с переобучением, поэтому необходимо использовать методы регуляризации, позволяющие найти баланс между сложностью получаемых алгоритмов и их обобщающей способностью.

Время обучения композиций деревьев решений с линейными или нелинейными разделителями на данных большого объема существенно превосходит время построения ансамблей деревьев решений с одномерными разделителями. Кроме того, для построения различных видов разделителей требуются разные типы вычислительных ресурсов. Поэтому актуальной становится задача разработки специализированных распределенных архитектур систем построения композиций деревьев решений с многомерными разделителями.

Целью работы является повышение качества (полноты и точности) решения задачи классификации на основе лесов деревьев решений путем создания вычислительно-эффективного метода построения случайных ансамблей деревьев решений с применением линейных и нелинейных разделителей.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Разработать и реализовать метод построения деревьев решений применением линейных и нелинейных разделителей.
2. Развить теоретические основы для регуляризации случайных ансамблей деревьев решений: разработать методы оценки обобщающей способности случайных ансамблей деревьев решений.
3. Разработать методы классификации объектов сложной структуры (изображения, тексты) на основе исследованных деревьев решений с линейными и нелинейными разделителями.
4. Провести экспериментальное исследование разработанного метода, сравнить его с другими методами, в том числе на задачах обработки изображений.
5. Разработать модель (архитектуру) для организации глобально распределенного обучения случайных лесов деревьев решений с линейными и нелинейными разделителями.
6. Разработать комплекс программ для обучения случайных лесов деревьев решений с линейными и нелинейными разделителями.

Основные положения, выносимые на защиту:

1. Вычислительно-эффективный (линейная сложность) метод построения узлов деревьев решений с применением линейных и нелинейных разделителей, при обучении которых совместно оптимизируется отступ между разделяемыми объектами и произвольный критерий однородности данных.

2. Оценка обобщающей способности случайных ансамблей деревьев решений, учитывающая основные гиперпараметры алгоритмов их построения.
3. Метод классификации объектов, характеризующихся наличием связей между признаками.
4. Архитектура программного обеспечения глобально распределенного обучения случайных лесов деревьев решений с линейными и нелинейными разделителями.
5. Комплекс программ для обучения случайных лесов деревьев решений с линейными и нелинейными разделителями.

Перечисленные положения относятся к направлениям исследований 4, 7, 8, и 9 паспорта специальности 2.3.5.

Научная новизна.

Разработан оригинальный вычислительно-эффективный метод построения узлов деревьев решений с применением линейных и нелинейных разделителей, при обучении которых совместно оптимизируется отступ между разделяемыми объектами и произвольный критерий однородности. Этот метод применен для обучения деревьев решений в составе случайных лесов. Предложена архитектура программного обеспечения глобально распределенного обучения случайных лесов деревьев решений с линейными и нелинейными разделителями. Выполнено развитие теоретических подходов к подбору методов регуляризации случайных ансамблей деревьев решений:

1. Теоретически обоснована связь между равномерной стабильностью (устойчивостью) алгоритмов обучения и формируемой структурой деревьев решений.
2. Предложена оценка обобщающей способности случайных ансамблей деревьев решений, учитывающая основные гиперпараметры алгоритмов их построения.

Теоретическая и практическая значимость.

Теоретическая значимость состоит в развитии формальных подходов к исследованию обобщающей способности случайных ансамблей деревьев решений. Эти подходы могут быть применены для разработки или подбора методов регуляризации случайных ансамблей деревьев решений. Теоретические результаты могут быть востребованы в ИПУ РАН, ИПС РАН, ИСП РАН, ФИЦ ИУ РАН при создании новых методов и систем интеллектуального анализа данных.

Практическая значимость состоит в повышении точности и полноты решения задач анализа данных с применением ансамблей деревьев решений (показано повышение точности на 8% на наборе данных Cifar-10, и на 2% на наборах Letter и USPS), а также в повышении (более чем в четыре раза)

производительности распределенного программного обеспечения обучения ансамблей деревьев решений при обработке данных большой размерности. Результаты работы использованы при реализации проектов № 075-15-2020-799, № 075-15-2020-805, поддержанных Министерством образования и науки Российской Федерации, проекта 19-29-07163 мк, поддержанного Российским фондом фундаментальных исследований. Предложенные методы использовались при решении задач анализа психолингвистических характеристик сообщений в социальных сетях, выявления когнитивных операций в научных текстах, анализа нормативно-правовых документов, а также задач оценки вегетационных индексов по цветным изображениям. Результаты работы могут применяться при создании прикладных систем интеллектуального анализа данных в сельском хозяйстве, промышленном производстве, в энергетике, на транспорте.

Апробация работы. Основные результаты работы были представлены на следующих конференциях и семинарах:

- XI Международная научно-практическая конференция «Интегрированные модели и мягкие вычисления в искусственном интеллекте» (ИММВ-2022, Коломна, 16-19 мая 2022 г.
- Научный семинар "Математические модели информационных технологий" департамента анализа данных и искусственного интеллекта "Интеллектуальные системы и структурный анализ" НИУ ВШЭ 25.03.2021, Москва.
- Научный семинар по системному программированию под руководством академика РАН А.И. Аветисяна по теме «Построение рандомизированных ансамблей деревьев решений с использованием ядерных разделителей» , 16.06.2022, Москва.
- Научный семинар кафедры информационных технологий РУДН «Построение рандомизированных ансамблей деревьев решений с использованием ядерных разделителей», 17.06.2022, Москва.

Личный вклад. Все выносимые на защиту результаты получены лично автором.

Публикации.

Основные результаты по теме диссертации изложены в шести работах, опубликованных в изданиях, рекомендованных ВАК или приравненных к ним (Scopus/Web Of Science), а также представлены в форме зарегистрированной программы для ЭВМ. В статьях [1, 3] вместе с соавторами была поставлена задача и проводилась редакторская правка, остальная часть выполнена соискателем. Работы [4-7] полностью выполнены автором.

Объем и структура работы. Диссертация состоит из введения, трёх

глав и заключения. Полный объем диссертации составляет 115 страниц с 26 рисунками и 8 таблицами. Список литературы содержит 114 наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках диссертационной работы, формулируется цель и ставятся задачи, перечисляются основные положения, выносимые на защиту, излагается научная новизна, теоретическая и практическая значимость представляемой работы.

Первая глава содержит обзор существующих подходов к обучению деревьев решений с линейными и нелинейными разделителями, построению случайных ансамблей подобных деревьев, оценки их обобщающей способности. В начале главы даются основные определения и понятия, связанные с обучением деревьев решений и построением их композиций. Под *алгоритмом классификации* понимается алгоритм, возвращающий для заданного произвольного объекта соответствующую ему метку. *Алгоритмом обучения* называется процедура построения алгоритма классификации на основе заданной обучающей выборки. В работе рассматриваются алгоритмы классификации – бинарные *деревья решений,* в каждом *узле* которых с применением *разделителя* происходит распределение анализируемых объектов по двум подмножествам, соответствующим левому и правому поддеревьям. *Разделитель* может представлять собой простое решающее правило, в котором значение определенного признака сравнивается с заданным порогом, а может являться более сложным линейным, либо нелинейным алгоритмом классификации объектов по поддеревьям. При обучении деревьев решений, как правило, оптимизируется некоторый *критерий неоднородности данных*, такой как неоднородность Джини (Gini impurity), информационная энтропия и другие. Под *ансамблем* понимается линейная *композиция* деревьев решений, под с*лучайным ансамблем* понимается композиция, отдельные алгоритмы которой обучены на случайных подмножествах обучающих данных (*бэггинг*) или случайных подмножествах данных и признаков (*случайный лес*).

Раздел 1.1 посвящен анализу оценок обобщающей способности методов классификации, в частности деревьев решений и их ансамблей. Анализ показывает, что критерии неоднородности данных, учитываемые при обучении деревьев решений, влияют на их структуру, тогда как структура в свою очередь влияет на обобщающую способность. По итогам анализа сформирован набор требований к алгоритмам обучения деревьев решений с линейными и нелинейными разделителями:

- Алгоритмы формирования структуры деревьев должны быть квази-оптимальными, чтобы повысить случайность формируемых деревьев.

– При построении узлов деревьев решений должен оптимизироваться отступ между данными в поддеревьях и заданные критерии неоднородности данных в поддеревьях.

В современных исследованиях деревьев решений используются оценки, основанные на сложности алгоритмов классификации (VC-размерность, Радемахеровская сложность). Это делает невозможным обобщение полученных результатов на случайные ансамбли деревьев решений, так как в этом случае важным фактором, влияющим на обобщающую способность, является случайность процесса выборки обучающих данных, то есть на обобщающую способность оказывают влияние особенности алгоритма обучения ансамбля. В таком случае для оценки обобщающей способности алгоритмов классификации могут быть использованы подходы, основанные на равномерной случайной стабильности (устойчивости) алгоритмов обучения. Основным инструментом для вывода таких оценок являются неравенства концентрации меры МакДиармида для вещественных функций от нескольких случайных величин. В этих неравенствах отклонение функции от математического ожидания зависит от верхней границы изменения значения этой функции при изменении одного из её параметров. В случае исследования алгоритмов обучения, равномерная стабильность определяет верхнюю границу изменения функции потерь при замене одного из объектов обучающего набора данных. Такие оценки являются граничными и завышенными, тем не менее, они позволяют качественно оценить влияние изменений алгоритмов построения деревьев и гиперпараметров этих алгоритмов на обобщающую способность. Подобные оценки предложены для случайных ансамблей произвольных алгоритмов классификации, они позволяют исследовать влияние некоторых гиперпараметров ансамбля на обучающую способность. Однако, большую практическую значимость представляет изучение эффекта, достигаемого при комбинации таких подходов как деревья решений и случайные ансамбли.

По итогам обзора выявлено, что методы оценки обобщающей способности случайных ансамблей деревьев решений на основе равномерной стабильности алгоритмов обучения отсутствуют. Создание таких методов позволило бы сформировать дополнительные требования к алгоритмам обучения и регуляризации случайных ансамблей деревьев решений, подбору гиперпараметров, повысить их обобщающую способность.

Раздел 1.2. содержит обзор методов классификации, в том числе обучения деревьев решений с линейными и нелинейными разделителями. Эти методы можно разделить на группы с учетом их двух основных особенностей: применяемого подхода к формированию структуры деревьев и используемого алгоритма оптимизации функции потерь при обучении узлов деревьев. В

результате группирования методов по подходу к формированию структуры деревьев решений выделяются методы обучения деревьев решений с фиксированной структурой. Для решения этой задачи может применяться, в том числе, метод обратного распространения ошибки. Методы усиления обученных деревьев решений состоят в построении деревьев решений с применением стандартных алгоритов (CART, ID3, C4.5) и дальнейшей модификации условий в узлах и эмпирических оценок вероятностей классов в листьях этих деревьев. К третьей группе относятся методы обучения деревьев решений с применением жадных алгоритмов. В этих методах дерево решений строится рекурсивно от корня, в ходе обучения каждого узла оптимизируется некоторая локальная функцимя потерь, связанная с этим узлом. Подобный поход позволяет повысить вариативность структур порождаемых деревьев при условии обучения на случайных подмножествах объектов и признаков.

В результате группирования методов обучения деревьев решений по применяемой функции потерь и подходам к ее оптимизации при построении узлов деревьев можно выделить методы, в которых оптимизируется целочисленная функция потерь; методы, в которых применяются произвольные гладкие функции потерь, где для оптимизации используется стохастический градиентный спуск, при этом, достижение глобального оптимума не гарантируется; методы, в которых используются выпукло-вогнутые (сумма выпуклых и вогнутых функций), либо выпуклые функции потерь. Наибольшей производительности при обучении деревьев решений можно добиться при использовании выпуклой функции потерь, так как в этом случае могут быть использованы алгоритмы оптимизации, имеющие линейную вычислительную сложность относительно размера обучающего набора данных.

По итогам аналитического исследования сформулированы дополнительные требования к алгоритмам обучения деревьев решений с линейными и нелинейными разделителями:

- Обучение узлов деревьев решений должно сводиться к решению задачи оптимизации гладкой выпуклой функции потерь с ограничениями-неравенствами.
- Узлы деревьев решений должны выполнять классификацию объектов в спрямляющих признаковых пространствах.

По итогам анализа выявлено, что методы обучения деревьев решений с линейными или нелинейными разделителями, удовлетворяющие всем заданным в разделах 1.1 и 1.2 требованиям, отсутствуют.

В разделе 1.3 приведен обзор методов классификации объектов сложной структуры. Отмечены недостатки существующих подходов к решению этой задачи на основе случайных лесов деревьев решений. Раздел

1.4 содержит обзор архитектур и систем распределенного обучения деревьев решений и их композиций. По итогам обзора отмечено, что рассмотренные архитектуры предназначены для обучения деревьев с одномерными разделителями и не могут быть обобщены для деревьев решений с линейными или нелинейными разделителями.

Во **второй главе** представлен метод построения деревьев решений с применением линейных и нелинейных разделителей, метод оценки обобщающей способности случайных ансамблей деревьев решений, метод классификации объектов, характеризующихся наличием связей между признаками.

Раздел 2.1 содержит описание метода построения деревьев решений применением линейных и нелинейных разделителей. Метод включает в себя рекурсивный Алгоритм 1. На каждом шаге алгоритма строится узел дерева, который разделяет обучающие данные, затем эта процедура рекурсивно повторяется для «левого» и «правого» подмножеств обучающих данных, пока не будет достигнута заданная высота дерева.

Алгоритм 1. Обучение дерева с линейными и нелинейными разделителями (Девяткин, 2021)

Вход: набор данных D_m, параметр регуляризации C. J - порог выбра способа распределения классов по поддеревьям (точный/жадный), K – количество запусков жадной процедуры распределения классов по поддеревьям.

1: **вызвать** *BuildTree*(D_m)
2: BuildTree(D):
3: **Если** D содержит объекты одного класса:
4: **Выход**
5: **Иначе:**
6: **Если** $|U| < J$:
7: s:=*all_distributions(Y,H={+1,-1})*
8: s_{best}:=*sort(Impurity(s))[rnd(1..N)]* \\$U \to H$
9: **Иначе**
10: s_{best}:=*greedy_find_best_impurity(D)*\\ $U \to H$
11: $L^*_1, \dots, L^*_m := L(h_i, -h_i), h_i \in H_{best}$
12: $w^*, \varepsilon^*_1, \dots, \varepsilon^*_m := optimize\_node(C, X, L^*_1, \dots, L^*_m$ \\Решить задачу обучения SVM с масштабированными переменными невязки $\frac{\varepsilon^*_1}{L^*_1}, \dots, \frac{\varepsilon^*_m}{L^*_m}$
13: $D_l := D[classify(D, w^* \geq 0)]$
14: $D_l := D[classify(D, w^* < 0)]$
15: **вызвать** *BuildTree(D_l)*

16: **вызвать** *BuildTree(D_r)*

Рассмотрим построение разделителя как задачу бинарной классификации, в которой при обучении примеры $D_m = \{< x_i, y_i > | i = 1, .., m; x_i \in X_m, y_i \in Y_m\}, X_m \sim P_X, Y_m \sim P_Y, D_m \sim P_{XY}$ с метками классов y_i, выбирающихся из множества U, распределяются по поддеревьям и $H = \{-1,+1\}$– метки этих поддеревьев. На первых шагах алгоритма (шаги 5-10) устанавливается целевое распределение классов в поддеревьях. Некоторые критерии могут разделять объекты одного класса по разным поддеревьям, но в предложенном методе эта особенность игнорируется в угоду скорости обучения. Если количество классов $|U|$ превышает пороговое значение J, являющееся гиперпараметром алгоритма, перебираются все возможные распределения классов по поддеревьям и вычисляются значения соответствующего критерия неоднородности (шаги 6-7). Затем полученный список вариантов распределений сортируется по критерию неоднородности, и целевое распределение c_s выбирается случайным образом среди первых N элементов отсортированного списка. Если количество классов $|U|$ больше, чем J, для выявления целевого распределения классов по поддеревьям применяется жадная процедура (шаг 9). Процедура начинается с генерации случайного распределения классов по поддеревьям. Затем сгенерированное распределение итеративно изменяется и сохраняются модификации, улучшающие критерий неоднородности данных. Эта процедура повторяется K раз, затем выбирается наилучшее распределение $c_S : U \to H$. В результате в процесс обучения узла дерева решений добавляется рандомизация, которая позволяет уменьшить корреляцию между отступами деревьев. Затем алгоритм использует полученное распределение c_s для определения поддерева для каждого объекта из обучающего набора данных D_m: $H_{best} = c_s(Y_m)$.

Следующие шаги алгоритма (шаги 11-12) состоят в обучении разделителя в узле дерева решений. В процессе обучения одновременно оптимизируется отступ между объектами в поддеревьях и критерий неоднородности. Для вычислительно-эффективного построения разделителей, аналогично методам обучения SVM классификации структур, задается непрерывная гладкая функция потерь, которая отражает зависимость между параметрами разделителя и значением критерия неоднородности данных (Рис. 1).

Обучение разделителя на наборе данных из m обучающих примеров производится путем решения следующей задачи оптимизации:

$$w^*, \xi^* = \arg\min_{w,\xi}\left(\frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m}\xi_i\right) \quad (1)$$

При ограничениях:

$$\forall i, w^T x h_i \geq 1 - \frac{\xi_i}{L(h_i, -h_i)}$$

где w^* — параметры разделяющей гиперплоскости, ξ^* — значения переменных невязки, C – параметр регуляризации, а $L(h_i, h)$ отражает прирост критерия неоднородности данных в случае отнесения объекта i к некорректному поддереву – h_i вместо h_i (Рис. 1).

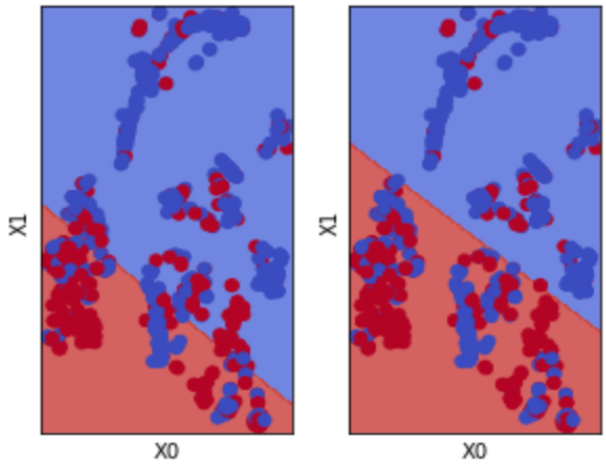

Рисунок 1. Влияние масштабирования переменных невязки $\xi_1,..,\xi_m$ на результаты построения разделяющей гиперплоскости. Без масштабирования (слева), с масшитабированием – справа. Набор данных Titanic, критерий – неоднородность Джини

Если применить к этой задаче условия Каруша-Куна-Таккера, то получится двойственная задача оптимизации:

$$a^* = \arg\max_a\left(-\frac{1}{2}\sum_{i=1..m}\sum_{j=1..m} a_i a_j K(x_i, x_j) + \sum_{i=1..m} a_i\right) \quad (2)$$

При ограничениях:

$$\sum_{i=1..m}\frac{a_i}{L(h_i, -h_i)} \leq \frac{C}{m}$$

где a_i — вес примера i *из* обучающей выборки (отличный от нуля для опорных векторов), а $K(x_i, x_j)$ — положительно определенное ядро. В

отличие от классификации структур, эта задача эффективно решается в явном виде, поскольку классов всего два (два поддерева). Гиперпараметр регуляризации C должен быть подобран эмпирически для каждого набора данных.

Шаги 13-16 алгоритма направлены на классификацию всех обучающих данных по поддеревьям с помощью построенного разделителя и рекурсивного обучения новых узлов дерева разделению обучающих данных поддеревьев: D_l, D_r.

В диссертации показано, что решение задач (1) или (2) позволяет оптимизировать критерий неоднородности данных.

На основе дереьев, обученных с помощью представленного алгоритма, с использованием подходов, предложенных Лео Брейманом, формируются случайные леса.

Раздел также содержит оценки вычислительной сложности представленного метода. Для деревьев решений с одномерными разделителями сложность обучения составляет $O\big(mflog(m)\big)$, так как сбалансированные деревья решений содержат не более $log\,(m)$ узлов, а сложность обучения каждого узла определяется необходимостью перебора f признаков и m обучаюших примеров, чтобы подобрать признак и порог, оптимизирующий снижение критерия неоднородности данных.

При обучении деревьев решений с линейными и нелинейными разделителями на верхнем уровне применяется тот же рекурсивный алгоритм, что и для стандартных деревьев. Поэтому сложность обучения узла в этом случае зависит от типа разделителя (линейный или нелинейный) и от метода оптимизации, используемого для обучения разделителя. В случае линейного разделителя можно использовать метод покоординатного спуска. Этот метод имеет линейную сложность относительно количества обучающих примеров. Следовательно, оценка сложности такая же, как и для деревьев, с одномерными разделителями $O(mflog(m))$. В случае нелинейных ядер одним из применимых подходов к обучению является метод структурной минимальной оптимизации (SMO). Вычислительная сложность этого метода составляет $O(m^2f)$. Поэтому итоговая оценка сложности является полиномиальной, что может являться проблемой при обучении на больших наборах данных $O(m^2flog(m))$.

Раздел 2.2 содержит метод теоретической оценки обобщающей способности случайных ансамблей деревьев решений. Предварительные экспериментальные исследования разработанного метода показали, что при использовании нелинейных разделителей наблюдается переобучение формируемых лесов. Для снижения переобучения необходимо создать или подобрать подход к регуляризации лесов деревьев решений. Для подбора

этого подхода необходимо разработать теоретическую оценку, учитывающую влияние особенностей алгоритма обучения ансамбля и его гиперпараметров (высота деревьев, количество деревьев, количество листьев, параметры обучения узлов деревьев) на обобщающую способность. Для упрощения задачи в рамках работы рассматривалось только влияние случайного выбора объектов при обучении деревьев ансамбля, то есть формально получены оценки обобщающей способноости для бэггинга на деревьях решений. Влияние случайного выбора подмножества признаков, используемых при обучении, учитывалось только в том смысле, что этот выбор способствует сокращению размерности признакового пространства.

Для вывода оценки обобщающей способности необходимо задать формальное описание деревьев решений и их ансамблей.

Пусть $H = \{h_1, h_2, .., h_\Gamma\}$– множество длины Γ листовых функций $s = h_i(x, y)$, которые отображают объекты и их метки на отрезок $\mathbb{R}^f \times U \to [0..1]$, где Γ – количество листьев в дереве. Все функции из множества H возвращают эмпирическую оценку вероятности того, что объект x с меткой y принадлежит листу с индексом i.

Зададим $Q = \{q_1, q_2, .., q_n\}$ – множество алгоритмов выбора узла $s = q_{j_{D_{m_j}}}(x)$, обученных на подмножествах $D_{m_j} \subseteq D_m$. Алгоритм $q_{D_j}(x)$ возвращает 1 если x принадлежит к узлу j, и 0 в противном случае.

Цепочкой решений для листа с индексом i будем считать произведение результатов всех алгоритмов выбора узла, соответствующих пути от корня до этого листа i, и значения функции выбора листа i.

Потребуем, чтобы цепочка решений удовлетворяла следующим свойствам:

1. Существует строгий порядок обучения алгоритмов выбора узла в цепочке.
2. Каждый алгоритм выбора узла в цепочке определяет какое подмножество обучающих данных будет использовано при обучении следующего узла в цепочке.
3. Каждый алгоритм выбора узла в цепочке определяет какое подмножество обучающих данных не будет учитываться при обучении узла в цепочке.

Вывод оценки обобщающей способности случайных ансамблей деревьев решений состоит из следующих шагов:

1. Вывод оценки равномерной стабильности алгоритма обучения цепочек решений.
2. Вывод оценки равномерной стабильности алгоритма обучения случайного ансамбля деревьев решений (как линейной композиции зависимых цепочек решений).
3. Вывод оценки обобщающей способности случайного ансамбля деревьев

решений, зависящей от равномерной стабильности алгоритма обучения этого ансамбля.

Для оценки равномерной стабильности алгоритма обучения цепочек решений получим сначала выражение для более общего случая – произведения результатов конечного числа произвольных алгоритмов классификации, обученных с использованием равномерно стабильных алгоритмов.

Лемма 1. *(Девяткин). Пусть* $Pr(x) = \prod_{q \in G} q(x)$ *- произведение результатов всех алгоритмов классификации q с областью значений [0..1] из множества G мощности* $n \in \mathbb{N}^+$ *. Эти функции обучены с помощью алгоритмов, оптимизирующих некоторые B-Липшицевы функции потерь. Положим, что эти алгоритмы* γ_q*-равномерно стабильны. Тогда равномерная стабильность алгоритма обучения этого произведения Pr(x) ограничена сверху следующим образом:*

$$\gamma_{Pr} \leq \frac{1}{n}\left(\sum_{i=1}^{n-2}\left(\frac{B}{2}\right)^i + 2\left(\frac{B}{2}\right)^{n-1}\right)\sum_{q \in G}\gamma_q = \varphi(B,n)\sum_{q \in G}\gamma_q \quad (8)$$

Следствие 1. *(Девяткин) Для указанного множества G равномерная стабильность алгоритма обучения произведения* $Pr(x)$ *асимптотически ограничена сверху равномерной стабильностью алгоритмов обучения элементов G:* $\gamma_{Pr} = O(\gamma_q)$.

Применяя свойства 1-3 цепочек решений и полагая алгоритмы обучения узлов в цепочке решений равномерно стабильными получим оценку равномерной стабильности для цепочек решений.

Следствие 2. *(Девяткин). Пусть соблюдаются свойства 1-3 цепочек решений. Алгоритмы обучения узлов в цепочке решений равномерно стабильны:* $\gamma = O(\frac{1}{m})$. $P_D(j)$*:* $\mathbb{N} \to [0..1]$ *- вероятность того, что элементы из набора* D_{j-1} *представлены также в наборе* D_j*, тогда равномерная стабильность алгоритма обучения цепочки решений длины* $n \in \mathbb{N}^+$ *ограничена сверху:*

$$\gamma_{Pr} \leq \varphi(B,n)\sum_{q \in G}\frac{\gamma_q}{\prod_{j=1}^{i} P_D(j)} \quad (9)$$

Следствие 3. *(Девяткин). Пусть соблюдаются свойства 1-3 цепочек решений, а алгоритмы обучения узлов в цепочке решений равномерно стабильны, тогда алгоритм обучения цепочки решений является равномерно стабильным*.

В существующих подходах к оценке равномерной стабильности случайных ансамблей деревьев решений предполагается независимость случайных параметров выбора обучающих данных для отдельных алгоритмов

в ансамбле. Однако, случайные параметры выбора обучающих данных для цепочек решений, составляющих дерево, зависимы (цепочки должны иметь совпадающие подмножества узлов). Поэтому в ходе исследования был создан подход, не накладывающий ограничений на зависимость этих параметров. Зададим условия, которым должен удовлетворять алгоритм построения случайного ансамбля деревьев решений:

1. Пусть $\mathcal{R} = \{0,1\}^m$. Алгоритм обучения ансамбля использует случайные векторы $\boldsymbol{r} = \{\boldsymbol{r_1}, \boldsymbol{r_2}, .., \boldsymbol{r_T} | \boldsymbol{r_t} \in \mathcal{R}\}$ для выбора объектов из D_m с целью обучения алгоритмов f_t: $\boldsymbol{r_{ti}} = 1$ если соответствующий объект i из D_m используется, 0 – иначе. $D(\boldsymbol{r_t})$ – набор данных, сгенерированный для обучения алгоритма f.
2. Несколько копий одного и того же объекта не влияют на результат обучения.

Теорема 1. *(Девяткин). Пусть ансамбль из T деревьев решений обучается методом бэггинга на наборе данных D_m размера m, каждое дерево с индексом i из этого ансамбля содержит Γ_i цепочек решений длины n_j. Пусть для обучения узлов деревьев используются алгоритмы с равномерной стабильностью γ, которые оптимизируют B-Липшицеву функцию потерь. Пусть выполняются условия 1 и 2. Тогда равномерная стабильность алгоритма бэггинга β_m будет ограничена сверху:*

$$\beta_m \leq \frac{B}{T} \sum_{k=1}^{m} \sum_{i=1}^{T} \sum_{j=1}^{\Gamma_i} \gamma_{Pr}(i,j,k) \frac{k}{m} P_{\boldsymbol{r_{ij}}}\left(d\left(\boldsymbol{r_{ij}}\right) = k\right), \tag{10}$$

где: $\gamma_{\mathrm{P}r}(i,j,k) \leq \varphi(B, n_j) \sum_{l=1}^{n_j} \frac{\gamma_k}{\prod_{t=1}^{l} P_{D(\boldsymbol{r_{ij}})}(t)}$, $d\left(\boldsymbol{r_{ij}}\right)$ *- кол-во различных элементов в* $D(\boldsymbol{r_{ij}})$, $P_{\boldsymbol{r_{ij}}}(d\left(\boldsymbol{r_{ij}}\right) = k)$ *- вероятность того, что в наборе данных* $D(\boldsymbol{r_{ij}})$ *ровно k элементов.*

Если алгоритмы обучения узлов в цепочке решений равномерно стабильны $\gamma_k = O\left(\frac{1}{k}\right)$, тогда значение выражения (10) будет снижаться при уменьшении числа листьев Γ_i и стремлении распределения вероятностей переходов от корня к узлам деревьев решений, задаваемого величинами $\prod_{t=1}^{l} P_{D(\boldsymbol{r_{ij}})}(t)$ к равномерному. В противном случае в составе деревьев решений будет расти число нестабильных цепочек решений, содержащих узлы с низкой вероятностью перехода $\prod_{t=1}^{l} P_{D(\boldsymbol{r_{ij}})}(t)$. Анализ влияния размера ансамбля на усредненное распределение вероятностей переходов от корня к узлам деревьев решений показывает (Рис. 2), что с увеличением количества деревьев это распределение «сглаживается». Автором доказана экспоненциальная сходимость равномерной стабильности бэггинга на

деревьях решений к ее математическому ожиданию.

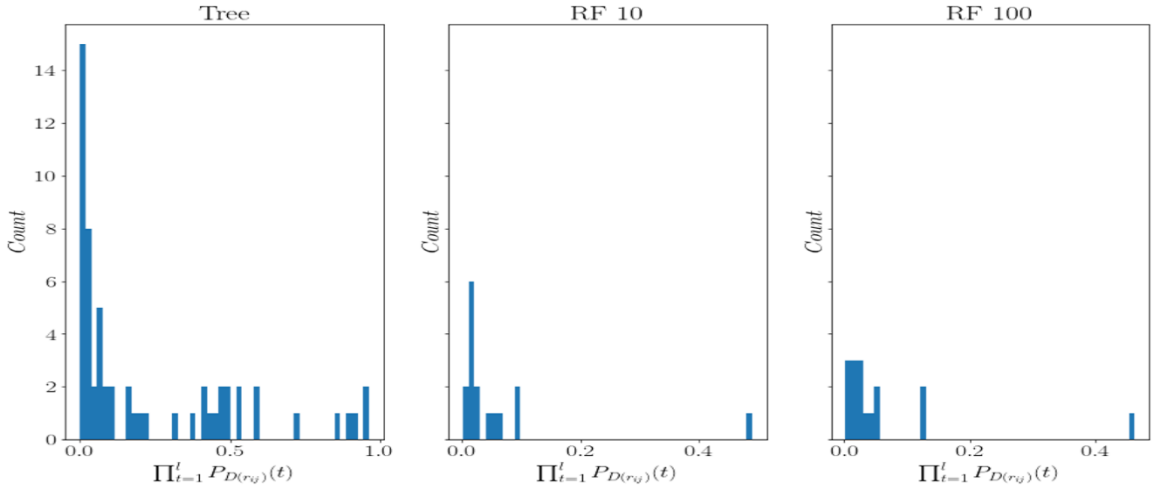

Рисунок 2. Распределение вероятностей переходов от корня к узлам дерева решений (слева), усредненное распределение вероятностей переходов от корня к узлам деревьев решений леса из 10 деревьев (середина) и из 100 деревьев (справа)

Следствие 4. *(Девяткин). Равномерная стабильность алгоритма бэггинга сходится с увеличением количества деревьев к своему математическому ожиданию μ с экспоненциальной скоростью. То есть:* $\forall T > 0: P(\beta_m(T) - \mu) \leq O(e^{-T})$.

Полученный результат доказывает многократно полученные эмпирические результаты, указавающие на избыточность ансамблей размером 1000 и более деревьев.

Аналогично оценке равномерной стабильности выведена оценка обобщающей способности случайного ансамбля деревьев решений.

Теорема 2. *(Девяткин). Пусть выполняются условия 1 и 2,* $F_{D(\boldsymbol{r})}(x, y) = \frac{1}{T}\sum_{t=1}^{TT} f_{D(\boldsymbol{r_t})}(x, y)$ *- ансамбль деревьев решений, построенный алгоритмом бэггинга на н.о.р. наборе данных D_m размером m. Каждое дерево ансамбля имеет Γ цепочек решений и равномерная стабильность алгоритма бэггинга ограничена сверху величиной β_m. Тогда с вероятностью не менее $1 - \delta$ верно следующее:*

$$R(F_{D(r)}) \leq \hat{R}(F_{D(r)}) + 2\beta_m + \frac{1}{2}B(\ln\left(\frac{1}{\delta}\right)\frac{2M}{3} + \sqrt{\left(\ln(\frac{1}{\delta})\frac{2M}{3}\right)^2 + 8ln\frac{1}{\delta}(B^2m(\beta_m)^2 + \frac{(2M)^2\Gamma^3}{T}))} \qquad (12)$$

В полученном выражении величина $B^4m(\beta_m)^2$является верхней гранью дисперсии ошибки, связанной с данными. Эта величина зависит от равномерной стабильности алгоритма обучения случайного ансамбля, то есть «сглаживание» распределения вероятностей переходов от корня к узлам

деревьев решений леса приводит к снижению дисперсии ошибки. Величина $\frac{(2BM)^2\Gamma^3}{T}$ является верхней гранью дисперсии ошибки, связанной с параметрами выборки данных $\boldsymbol{r}$, эта величина зависит от размера ансамбля T. Доказательства полученных оценок приведены в диссертации, результаты согласуются с существующими исследованиями в области стабильности бэггинга (Elisseeff, Friedman, Hall, Barlett и др.).

На основе полученных оценок можно сформулировать следующие неформальные критерии подбора метода регуляризации случайного ансамбля деревьев решений:

- Сокращение общего числа цепочек (листьев) в деревьях.
- Сокращение длины цепочек решений.
- Сокращение цепочек решений с низкими вероятностями перехода.

Проведено исследование методов регуляризации случайных ансамблей деревьев решений на соответствие сформулированным критериям подбора. Большая часть подходов к регуляризации случайных ансамблей состоит в искусственном внесении случайных различий между деревьями композиции. Вместе с тем, подобные подходы могут отрицательно повлиять на смещение ошибки. S. Ren и др. предложили отбросить эмпирические оценки вероятностей классов, хранящиеся в листьях деревьев предварительно обученного случайного леса, и переопределить их. Для получения новых значений оптимизируется глобальная функция потерь, сходная с применяемой в методе опорных векторов. Оптимизация глобальной функции потерь может привести к переобучению ансамбля. Для преотвращения этого эффекта используется прунинг, который состоит в соединении соседних листьев деревьев, в случае, когда норма разности векторов этих листьев (векторов переопределенных эмпирических вероятностей классов) близка к нулю. Эта операция приводит к уменьшению количества цепочек решений, уменьшению их длины, повышению вероятности переходов от корня к узлам деревьев, то есть выполняются все предложенные критерии подбора метода регуляризации, при этом за счет глобальной оптимизации функции потерь не происходит увеличения эмпирическонго риска.

В разделе 2.3 представлен метод классификации объектов, характеризующихся наличием связей между признаками. Метод представляет собой модификацию подхода DeepForest, в котором обработка данных происходит последовательно на нескольких слоях. Каждый слой имеет следующую структуру (Рис. 4). На основе каждого объекта из обучающего набора методом «скользящего окна» генерируется набор объектов, помеченных классом исходного объекта. Эти объекты используются для обучения случайного леса деревьев решений с линейными и нелинейными

разделителями, далее выполняется усиление и прунинг обученного леса. После процедуры прунига лес используется для формирования векторных представлений (эмбеддингов) обрабатываемых объектов для следующего слоя. Эти представления включают исходные признаки объектов, а также векторы эмпирических вероятностей, возвращаемые деревьями леса.

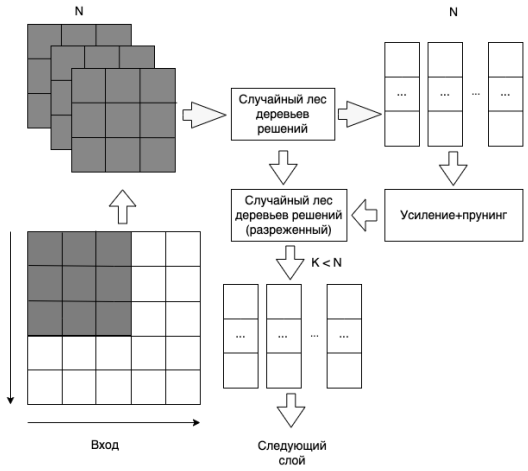

Рисунок 3. Схема метода классификации объектов, характеризующихся наличием связей между признаками (один слой)

Основными отличиями модифицированного метода от оригинального DeepForest являются:

- Использование случайных лесов деревьев решений с нелинейными разделителями в качестве базового алгоритма классификации, что позволяет учитывать связи между признаками анализируемых объектов, уменьшить количество слоев обработки данных и, следовательно, повысить производительность.
- Отказ от применения экстремально случайных лесов деревьев решений (Extremely Random Forests), вместо них для повышения обобщающей способности используются усиление и прунинг из раздела 2.4, что позволяет повысить стабильность работы метода.

Третья глава посвящена программной реализации и экспериментальным исследованиям представленных методов. Раздел 3.1. содержит описание архитектуры и реализации комплекса программ для обучения случайных лесов деревьев решений с линейными и нелинейными разделителями.

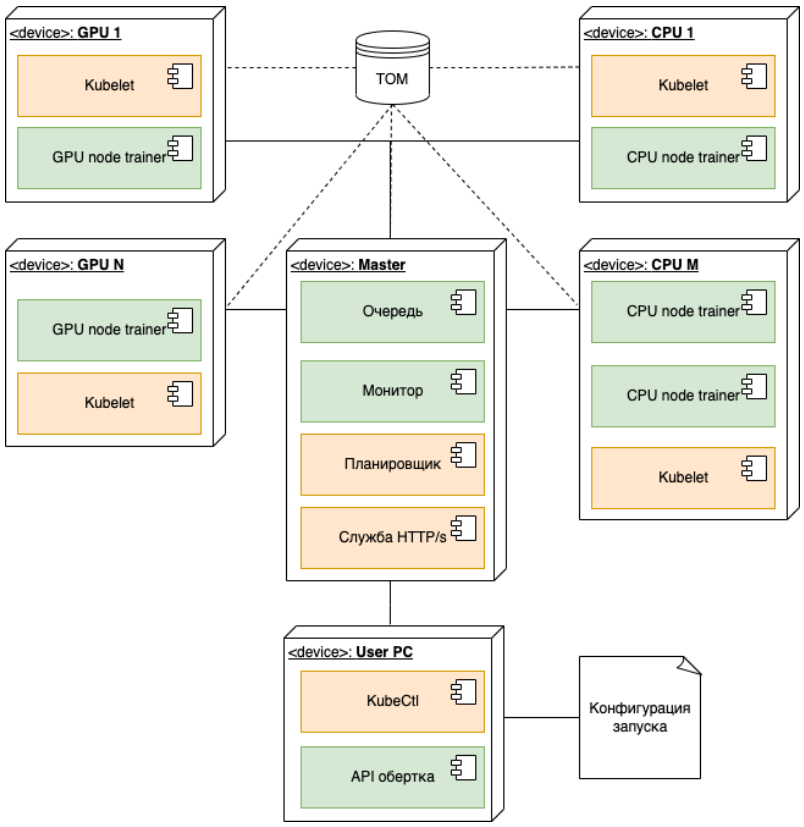

Рисунок 4. Архитектура комплекса программ для обучения случайных лесов деревьев решений с линейными и нелинейными разделителями

Для проведения экспериментальных исследований предложенных методов была разработана архитектура для организации глобально распределенного обучения случайных лесов деревьев решений с линейными и нелинейными разделителями и на ее основе разработан комплекс программ. При программной реализации методов обучения случайных лесов подобных деревьев необходимо учитывать, что их построение имеет следующие особенности:

1. Время обучения отдельных деревьев ансамбля может существенно отличаться ввиду отличий в структуре и высокой вычислительной сложности построения многомерных разделителей, поэтому распараллеливание обучения на уровне отдельных деревьев может привести к простою оборудования.

2. Существует необходимость подбора аппаратных средств и библиотек для оптимизации в зависимости от типа разделителя: если линейный разделитель может быть построен за приемлемое время с использованием ресурсов центральных процессоров, то для обучения нелинейных разделителей необходимо задействование графических или тензорных процессоров.

Для сокращения времени обучения лесов деревьев решений может применяться облачная вычислительная инфраструктура, однако она также накладывает определенные ограничения:

- теоретически неограниченное количество доступных виртуальных вычислительных узлов при ограниченном времени доступа к ним,
- относительно высокие задержки при передаче данных между узлами.

При использовании облачной инфраструктуры важную роль играет экономический фактор – стоимость обучения ансамбля, пропорциональная затраченному процессорному времени.

Для того чтобы соответствовать перечисленным ограничениям программное обеспечение обучения случайных лесов деревьев решений должно соответствовать следующим требованиям:

- параллелизм на уровне отдельных разделителей, что позволит снизить простой аппаратных ресурсов и назначать задания на обучение различным видам вычислительных узлов в зависимости от типа разделителя.
- возможность динамического масштабирования вычислительных ресурсов.
- минимизация передачи данных между узлами вычислительной системы.

Комплекс программ для обучения случайных лесов деревьев решений с линейными и нелинейными разделителями состоит из следующих основных компонентов (Рис. 5):

– Сервис обучения узлов деревьев решений с применением графических ускорителей (GPU node trainer).

– Сервис обучения узлов деревьев решений с использованием ресурсов центрального процессора (CPU node trainer).

– Сервис «Очередь задач на обучение узлов деревьев решений (Очередь)» реализует конкурентный доступ к очереди для чтения и записи задач на обучение узлов деревьев решений.

– Монитор выполняет контроль загруженности имеющихся аппаратных средств и состояние очереди. В случае роста размера очереди и наличия свободных аппаратных средств Монитор инициирует запуск дополнительных сервисов обучения узлов. В случае выявления простаивающих сервисов обучения узлов инициируется их остановка и освобождение аппаратных ресурсов.

Время обучения разделителя в системе труднопрогнозируемо, так как оно зависит от количества обучающих примеров, их конфигурации в признаковом пространстве, типа разделителя, производительности рабочего узла, что усложняет централизованную балансировку нагрузки и

распределение задач. Поэтому в предложенной архитектуре централизованный механизм распределения заданий на обучение разделителей отсутствует: каждый компонент обучения самостоятельно выполняет мониторинг Очереди на предмет наличия подходящих заданий, что позволяет упростить балансировку нагрузки и снизить количество пересылаемых сообщений между рабочими узлами.

На рис. 6 показана последовательность действий, выполняемых в системе при обучении разделителя дерева решений. Обучение разделителя выполняется сервисом NodeTrainer. В системе может параллельно функционировать множество подобных сервисов. Перед запуском обучения леса NodeTrainer с помощью компонента загрузки данных DataLoader получает и сохраняет локальную копию обучающих данных. Далее NodeTrainer запрашивает у Очереди задач битовую маску, определяющую примеры из обучающей выборки, которые будут использоваться для построения разделителя. Очередь помечает загруженную задачу флагом «на исполнении». В случае, если время обучения превысит заранее заданный порог, Очередь сбросит этот флаг, что приведет к перезапуску задачи на другом NodeTrainer. По итогу обучения полученная модель разделителя загружается в Очередь, которая инициирует сохранение результатов обучения на диск. Приведенная схема запуска заданий позволяет минимизировать передачу обучающих данных между вычислительными узлами в процессе обучения, так как для указания обучающих примеров при запуске заданий необходима лишь битовая маска.

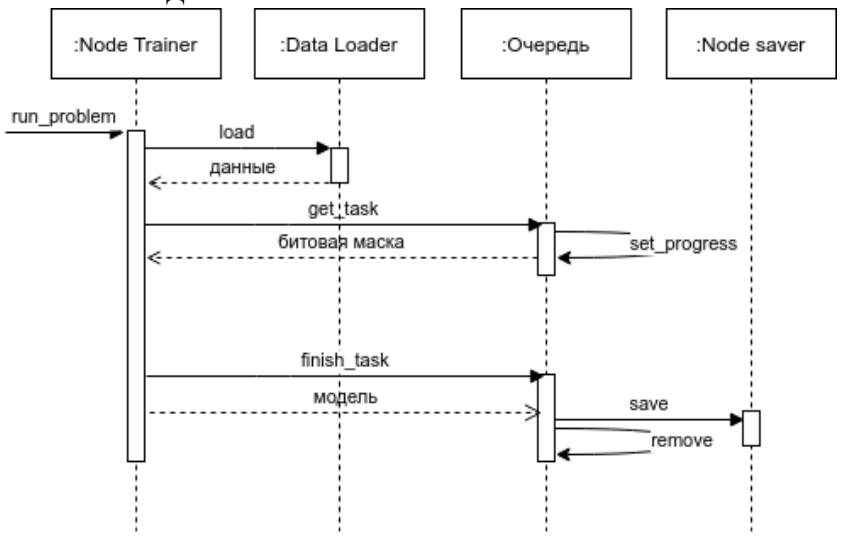

Рисунок 6. Диаграмма обучения узла дерева случайного леса

Для разработки комплекса программ использовались языки программирования C++/Cython/Python, программные библиотеки LibLinear,

LibSVM, ThunderSVM, Numpy, Scipy, утиллиты Kubernetes и Docker.

Раздел 3.2. содержит результаты экспериментальных исследований предложенных методов с применением разработанного комплекса программ.

В работе применялся стандартный подход к оценке качества методов классификации на размеченных выборках, используемый в аналогичных работах, посвященных разработке деревьев решений с многомерными разделителями: G. DeSalvo, M. Mohri, M. Naruzi и др. Для подбора гиперпараметров исследуемых методов использовалась статистическая процедура перекрестного скользящего контроля. Для оценки качества классификации использовалась отложенная выборка. Оценивались показатели точности (*accuracy, precision*) и полноты (*recall*) классификации с макроусреднением. Для лесов деревьев решений с ядерными разделителями выполнялся подбор следующих гиперпараметров: количество деревьев *T={30,100,300}*, параметр регуляризации при построении разделителей *C={100, 1000, 3000, 5000}*, максимальная глубина дерева *n={3,4,5,6,7}*, доля признаков, которые необходимо учитывать в каждом узле *f={0.08,0.1,0.2,0.3,0.4,0.5},* коэффициенты регуляризации (до 0,9), а также параметры ядра (*gamma={10,100}* для Гауссовского ядра, *degree=3* для полиномиального ядра). Интервалы подбора гиперпараметров найдены эмпирически в ходе предварительных исследований разработанных методов.

Экспериментальные исследования разработанного метода проводились на открытых размеченных наборах данных из коллекции UCI: SatImage, USPS, Letter, MNIST, наборе CIFAR-10, наборах YoutubeChannels и Titanic. Исследовались следующие методы: случайный лес деревьев решений (Random Forest), случайный лес деревьев, построенных методом CO2 (CO2 Forest), случайный лес деревьев WODT, случайный лес деревьев решений с линейными и нелинейными разделителями, метод построения которых предложен в настоящей работе (Kernel Forest).

В таблице 1 представлены полученные результаты (*accuracy* на отложенной выборке) экспериментального исследования методов построения случайных лесов деревьев решений различных видов. Наилучшие результаты для использованных наборов данных распознавания были получены с применением разделителей с Гауссовским ядром. Оценки *precision* и *recall* представлены в основном тексте диссертации.

Таблица 1. Результаты экспериментальных исследований (*accuracy*) методов построения случайных лесов с деревьями различных видов

Набор данных/ Метод	MNIST	USPS	Letter	SatImage	Cifar-10	Yutube Channels
Random Forest	0.972	0.936	0.963	0.911	0.501	0.938

CO2 Forest	0.981	0.945	**0.982**	0.911	-	0.821
WODT	0.943	0.905	0.879	0.876	-	0.914
Kernel Forest	0.991	0.946	0.975	0.918	0.581	0.944
Kernel Forest + усиление	**0.992**	0.952	0.981	0.920	0.590	0.955
Kernel Forest + усиление + базовый прунинг	**0.992**	0.953	**0.982**	0.920	0.591	0.955
Kernel Forest + усиление + l^2 прунинг	**0.992**	**0.954**	**0.982**	**0.921**	**0.591**	**0.957**

В таблице 1 также показаны результаты экспериментальных исследований регуляризации случайных лесов деревьев решений с ядерными разделителями. Следует отметить, что усиление из раздела 2.4 позволяет значительно улучшить результаты классификации на этих наборах данных (до 0,6%), повышая применимость деревьев со сложными ядрами. В то же время применение прунинга в дополнение к усилению практически не привело к улучшению результатов классификации.

В таблице 2 представлены полученные результаты (*accuracy* на отложенной выборке) экспериментальных исследований методов анализа объектов со сложной структурой.

Таблица 2. Результаты экспериментальных исследований (*accuracy*) методов анализа объектов со сложной структурой

Набор данных	Kernel Forest	Deep Kernel Forest	Deep Forest
MNIST	0.991	**0.994**	0.993
USPS	0.946	**0.979**	0.959
Letter	0.975	**0.985**	0.974
SatImage	0.918	**0.931**	0.930
Cifar-10	0.581	**0.632**	0.618

Точность предложенной модификации для анализа объектов со сложной структурой сопоставима с Deep Forest.

Раздел также содержит результаты оценки времени обучения ансамблей деревьев решений с линейными и нелинейными разделителями. Исследование проводилась на случайным образом выбранных подмножествах наборов данных MNIST и Cifar-10. Оценивалось время проводилось обучение лесов деревьев решений с линейными и нелинейными

(с Гауссовским ядром) разделителями высотой 10. На рис. 7 представлена зависимость времени обучения леса деревьев решений из 100 деревьев высотой 10 от размера обучающей выборки и алгоритма (для набора данных CIFAR-10). Согласно полученным результатам, предложенная архитектура обеспечивает значительный прирост скорости обучения при построении лесов деревьев решений большой глубины (10 и более) на данных большой размерности (более 3 тыс. признаков), таким образом, применение программного обеспечения, основанного на этой архитектуре, для обработки таких данных приведет к снижению общего машинного времени, необходимого для обучения и, как следствие, к снижению затрат на обучение.

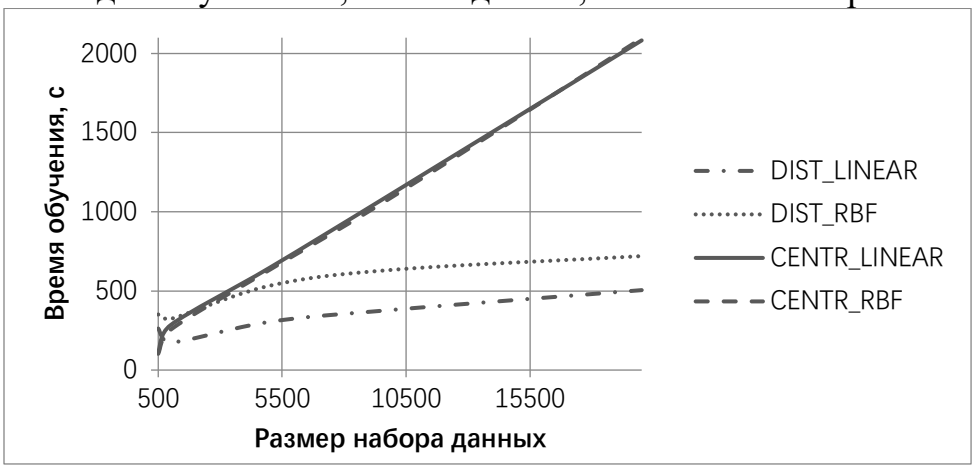

Рис. 7. Время обучения леса деревьев решений с линейными разделителями на наборе данных CIFAR-10 (100 деревьев высотой 10)

Полученные результаты согласуются с теоретическими оценками из раздела 2.1, а предложенная архитектура для организации глобально распределенного обучения случайных лесов деревьев решений с линейными и нелинейными разделителями позволяет многократно (более 4 раз по сравнению со стандартными подходами к распараллеливанию построения ансамблей на уровне отдельных деревьев) ускорить обучение ансамблей за счет рационального использования вычислительных ресурсов.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Разработан и реализован метод построения деревьев решений с применением линейных и нелинейных разделителей, при построении которых оптимизируется отступ между разделяемыми объектами и произвольный критерий однородности.
2. Проведены экспериментальное исследование разработанного метода, выполнено его сравнение с другими методами.

3. Разработаны методы оценки обобщающей способности случайных ансамблей деревьев решений. Оценки применены для подбора подход к регуляризации случайных лесов деревьев решений с линейными и нелинейными разделителями. Доказаны соответствующие теоремы и следствия.
4. Разработан метод классификации объектов, характеризующихся наличием связей между признаками.
5. Разработана архитектура программных средств глобально распределенного построения случайных лесов деревьев решений с линейными и нелинейными разделителями.
6. Реализован комплекс программ для обучения лесов деревьев решений с линейными и нелинейными разделителями.

Публикации автора по теме диссертации

1. *Девяткин Д. А*., *Григорьев О.Г.* Метод обучения деревьев решений с нелинейными разделителями // Искусственный интеллект и принятие решений, №3, 2022. С. 95-104.
2. *Dmitry Devyatkin, Oleg Grogoriev* Random Kernel Forests // in IEEE Access, vol. 10, pp. 77962-77979, 2022, doi: 10.1109/ACCESS.2022.3193385.
3. *В. В. Жебель, Д. А. Девяткин, Д. В. Зубарев, И. В. Соченков* Методы кросс-языкового поиска тематически похожих нормативно-правовых документов на основе машинного обучения // Искусственный интеллект и принятие решений. – 2022. – № 2. – С. 27-35. – DOI 10.14357/20718594220203.
4. *Devyatkin Dmitry*. Estimation of vegetation indices with Random Kernel Forests // Sensors, 2022.
5. *Devyatkin D.* Extraction of Cognitive Operations from Scientific Texts //Russian Conference on Artificial Intelligence. – LNCS, Springer, Cham, 2019. – С. 189-200.
6. *Девяткин Д. А*. Система распределенного построения случайных лесов деревьев решений с линейными и нелинейными разделителями // Системы высокой доступности, №3, 2022. – С. 59-67.
7. Свидетельство о регистрации программы для ЭВМ "Программный комплекс для обучения случайных лесов деревьев решений с нелинейными разделителями". Девяткин Д.А., 2022.

Девяткин Дмитрий Алексеевич
Построение ансамблей деревьев решений с использованием линейных и нелинейных разделителей
Автореф. дис. на соискание ученой степени канд. физ.-мат. наук
Подписано в печать ___.___.____. Заказ №_____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография_____