ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

На диссертационную работу Девяткина Дмитрия Алексеевича «Построение ансамблей деревьев решений с использованием линейных и нелинейных разделителей», представленную на соискание ученой степени кандидата физико-математических наук по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Актуальность темы исследований

Случайные композиции деревьев решений активно используются в системах интеллектуального анализа данных, как в России, так и за рубежом. Применение подобных композиций позволяет с приемлемой точностью учитывать негладкие зависимости между анализируемыми величинами, анализировать объекты дискретной природы, сохраняя при этом интерпретируемость полученных результатов. Однако композиции деревьев решений фиксированной глубины имеют ограниченную выразительную способность при анализе данных большой размерности. Использование многомерных разделителей в вершинах деревьев позволяет снять это ограничение, однако существующие алгоритмы их обучения имеют низкую вычислительную эффективность. Кроме того, деревья с многомерными разделителями имеют склонность к переобучению, поэтому требуется применение подходов к снижению их сложности.

Ограниченная практическая применимость существующих методов построения случайных композиций деревьев решений с многомерными разделителями привела к нехватке исследований в области создания программных архитектур систем обучения таких композиций, применимых для обработки больших массивов данных.

Таким образом, тема диссертационной работы, связанная с созданием вычислительно-эффективных методов и архитектуры распределенной системы построения случайных композиций деревьев решений с применением многомерных разделителей, является актуальной и востребованной.

Содержание диссертационного исследования

Диссертационная работа состоит из введения, трёх глав, заключения и списка литературы. Общий объем диссертации составляет 115 страниц с 26 рисунками и 8 таблицами. Список литературы состоит из 114 источников.

Введение содержит обоснование актуальности проблемы, связанной с созданием вычислительно-эффективных методов построения случайных ансамблей деревьев решений с применением линейных и нелинейных разделителей. Задается цель, и формулируются задачи исследования, аргументируется научная новизна, показывается теоретическая и практическая значимость полученных результатов.

В первой главе автор приводит результаты аналитического исследования современных методов оценки обобщающей способности алгоритмов обучения, построения деревьев решений с многомерными разделителями, обучения композиций алгоритмов, распределенных архитектур обучения композиций деревьев решений. В результате исследования выявлено, что необходима разработка новых распределенных архитектур систем, предназначенных для обучения композиций деревьев решений с многомерными разделителями. Отмечены недостатки существующих методов обучения деревьев решений и установлены критерии, которым должны соответствовать алгоритмы построения деревьев решений с многомерными разделителями. По итогам исследования сделан вывод об отсутствии подходов к оценке обобщающей способности случайных композиций деревьев решений, учитывающих основные параметры алгоритмов их обучения. Хочется отдельно положительно отметить первые разделы главы (разделы 1.1 и 1.2), в которых автор уделяет достаточно много внимания постановкам рассматриваемых задач и класическим подходам их решения – по этим обзорным материалам, в принципе, можно проводить занятия для студентов.

Вторая глава диссертации содержит описание алгоритма построения деревьев решений с многомерными разделителями, предложенного в диссертации, и оценку его вычислительной сложности. Показано, что при использовании линейный разделителей вычислительная сложность предложенного алгоритма соответствует сложности методов построения деревьев решений с одномерными разделителями. Представлена теоретическая оценка обобщающей способности, учитывающая влияние параметров алгоритмов обучения композиций деревьев решений. Указанная оценка предложена автором работы для подбора методов снижения сложности композиций деревьев решений. Эта теоретическая часть в значительной степени созвучна трём важнейшим понятиям «аппроксимация, устойчивость, сходимость», существенно связанным с задачами вычислительной математики. Отрадно, что автора волнует вопрос устойчивости и точности приведенных алгоритмов не только с практической точки зрения. Приведено также описание метода классификации объектов, характеризующихся наличием связей между признаками.

В третьей главе автор приводит описание архитектуры программного комплекса для обучения композиций деревьев решений с многомерными разделителями с помощью предложенных методов, а также результаты экспериментальных исследований этих методов на открытых размеченных наборах данных. Полученные результаты показывают, что предложенные методы позволяет значительно повысить качество решения задачи классификации с помощью лесов деревьев решений. Результаты исследования времени построения композиций деревьев решений показывают, что предложенная распределенная архитектура обеспечивает значительное увеличение быстродействия программного обеспечения обучения лесов деревьев решений на данных, характеризующихся признаковым пространством большой размерности.

Заключение содержит основные результаты диссертации.

Научная новизна

Научная новизна заключается в разработанном алгоритме обучения деревьев решений с применением многомерных разделителей, представленных оценке обобщающей способности случайных композиций деревьев решений и архитектуре системы глобально-распределенного построения композиций деревьев решений.

1. Предложен новый алгоритм обучения деревьев решений с применением многомерных разделителей. В этом алгоритме при построении вершин деревьев решений совместно оптимизируется отступ между разделяемыми объектами и произвольный критерий однородности данных. Этот метод применен для обучения деревьев решений в составе случайных лесов. Результаты экспериментальных исследований предложенного метода показали, что он позволяет добиться до 8% повышения точности и полноты классификации на открытых размеченных наборах данных.
2. Разработана оригинальная архитектура системы глобально-распределенного обучения случайных лесов деревьев решений с многомерными разделителями.
3. Предложен новый метод теоретической оценки обобщающей способности композиций деревьев решений. Показана практическая применимость метода для выбора подходов к снижению сложности композиций деревьев решений.

Степень обоснованности научных положений, выводов и рекомендаций

Все положения и выводы диссертации достоверны и научно обоснованы. Достоверность полученных результатов подтверждается экспериментальными данными, полученными на открытых размеченных наборах данных. Разработанные методы и алгоритмы основываются на корректном применении аппарата вычислительной математики, методов оптимизации, методов машинного обучения. Основные результаты апробировались на профильных конференциях и научных семинарах.

Теоретическая и практическая значимость работы

Теоретическая значимость работы обуславливается предложенными новыми методами обучения деревьев решений, определения обобщающей способности их композиций, а также архитектурой систем глобально-распределенного построения деревьев решений с многомерными разделителями.

Практическая значимость состоит разработке комплекса программ для обучения случайных лесов деревьев решений с линейными и нелинейными разделителями. Разработанное программное обеспечение может использоваться для решения прикладных задач интеллектуального анализа данных в различных областях науки и техники.

Замечания

1. В третьей главе диссертации отсутствует анализ влияния параметров регуляризации на скорость построения деревьев решений с помощью предложенного алгоритма.
2. В тексте диссертации, в особенности в первой главе, содержатся ошибки оформления и опечатки.
3. Во второй главе диссертационной работы приведен модифицированный метод обучения каскадных композиций лесов деревьев решений DeepForest (gcForest), однако отсутствует описание недостатков, на устранение которых направлена эта модификация.
4. В работе приведены графики времен работы разработанных алгоритмов, однако эти графики не поддерживаются табличными значениями времени вычислений, содержится мало информации о «железе» и отсутствуют некоторые важные детали о программной реализации разработанных алгоритмов (многопоточный режим, уровень утилизации ресурса GPU и проч.), с использованием которого проводились расчёты.
5. Разработанные алгоритмы и методы отчасти дополняют нейросетевой подход для задач классификации, а сравнение с ним в диссертации не слишком заметно.

Перечисленные замечания не снижают значимости полученных результатов и не влияют на общую положительную оценку диссертации.

Заключение

Диссертация Девяткина Д.А. представляет собой завершенное научно-квалификационное исследование, в котором решена актуальная научная задача повышения качества классификации методами на основе лесов деревьев решений. Диссертационная работа выполнена автором самостоятельно на высоком научном уровне. Представленные в работе результаты имеют теоретическое и эмпирическое обоснование. Они важны для дальнейшего развития методов построения деревьев решений и их композиций. Автореферат раскрывает основное содержание диссертации и соответствует требованиям ВАК. Результаты работы доказаны теоретически, либо подтверждены в ходе проведения эмпирических исследований.

Полагаю, что, несмотря на перечисленные недостатки, представленная к защите диссертация Девяткина Дмитрия Алексеевича «Построение ансамблей деревьев решений с использованием линейных и нелинейных разделителей» соответствует требованиям Положения о присуждении ученых степеней (Постановление Правительства Российской Федерации от 24 сентября 2013 года №842), предъявляемым к диссертациям на соискание ученой степени кандидата наук, а ее автор, Девяткин Д.А., заслуживает присуждения ученой степени кандидата кандидата физико-математических наук по специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Официальный оппонент:

Кандидат физико-математических наук Матвеев Сергей Александрович,

Доцент, ученый секретарь кафедры вычислительных технологий и моделирования факультета вычислительной математики и кибернетики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В.Ломоносова»

«28» ноября 2022

Кандидатская диссертация защищена по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ»

Адрес места основной работы 119991 ГСП-1 Москва, Ленинские горы, МГУ имени М.В. Ломоносова, д.1, стр. 52, 2-й учебный корпус, факультет ВМК

Собственноручную подпись Матвеева Сергея Александровича удостоверяю

«28» ноября 2022