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

Деревья решений используются во многих приложениях, таких как распознавание изображений и извлечение информации. В узлах стандартных (одномерных) деревьев решений используются разделяющие гиперплоскости, параллельные осям. Построение таких одномерных разделяющих поверхностей тривиально, так как существует возможность перебора всех возможных порогов для каждого признака и, выбора наилучших параметров в соответствии с критерием разделения, однако обучение таких деревьев решений на наборах данных большой размерности может приводить к чрезмерному усложнению деревьев. В таких случаях деревья решений с многомерными разделяющими поверхностями в узлах позволяют упростить структуру деревьев и обеспечивают лучшую производительность. Однако обучение подобных деревьев решений связано с высокими вычислительными затратами. Важный вопрос, связанный с использованием подобных деревьев, касается потери разнообразия их структуры, происходящей при определении параметров многомерных разделяющих поверхностей с помощью методов оптимизации, и, следовательно, снижения обобщающей способности лесов таких деревьев. Поэтому необходимо развивать и применять методы редукции сложности, позволяющие найти баланс между сложностью получаемых лесов деревьев решений и их обобщающей способностью. Время обучения лесов 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 распределенных архитектур систем построения случайных лесов деревьев решений с многомерными разделителями.

Содержание диссертационного исследования

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.

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. По итогам анализа сформирован набор требований к алгоритмам обучения деревьев решений с линейными и нелинейными разделителями, отмечено, что существующие программные архитектуры предназначены для обучения деревьев с одномерными разделителями и не могут быть применены для создания систем построения 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, метод теоретической оценки обобщающей способности случайных ансамблей деревьев решений.

Третья глава содержит описание архитектуры и реализации программы для обучения случайных лесов деревьев решений с многомерными разделителями, а также результаты экспериментальных исследований предложенных методов с применением разработанного комплекса программ. Представленные результаты экспериментальных исследований показывают, что предложенные методы позволяет значительно улучшить результаты классификации на открытых размеченных наборах данных, а предложенная архитектура программного обеспечения обеспечивает значительный прирост скорости обучения при построении лесов деревьев решений большой глубины на данных большой размерности. Таким образом, применение программного обеспечения, основанного на этой архитектуре, для обработки данных большой размерности приводит к уменьшению общего машинного времени, необходимого для обучения и, как следствие, к снижению затрат на обучение.

В **заключении** приведены основные результаты работы.

Научная новизна

1. 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 данных. Этот метод применен для обучения деревьев решений в составе случайных лесов.

2. Предложена архитектура программного обеспечения распределенного обучения случайных лесов деревьев решений с многомерными разделителями.
3. Выполнено развитие теоретических подходов к снижению сложности случайных ансамблей деревьев решений, а именно:
   - Теоретически обоснована связь между равномерной устойчивостью алгоритмов обучения и формируемой структурой деревьев решений.
   - Предложена оценка обобщающей способности случайных ансамблей деревьев решений, учитывающая основные гиперпараметры алгоритмов их обучения.

Обоснованность и достоверность научных положений, выводов и рекомендаций

Достоверность и обоснованность результатов, полученных в диссертации, подтверждается корректным применением методов оптимизации, теории вероятностей и математической статистики, вычислительной математики и искусственного интеллекта. Достоверность полученных методов и оценок подтверждается результатам эмпирических исследований на открытых размеченных наборах данных. Полученные результаты не противоречат результатам, полученным другими авторами.

Теоретическая и практическая значимость работы

Теоретическая значимость работы обуславливается предложенным методом и архитектурой программных средств распределенного построения деревьев решений с многомерными разделителями. Предложенный в работе метод оценки обобщающей способности случайных ансамблей деревьев решений может использоваться в качестве теоретической основы при создании новых подходов к редукции сложности ансамблей деревьев решений.

Практическая значимость состоит в значительном повышении точности и полноты решения задач анализа данных с применением ансамблей деревьев решений (показано повышение точности на 8% на открытых наборах данных), а также в значительном повышении производительности программного обеспечения обучения ансамблей деревьев решений при обработке данных большой размерности.

Замечания

1. Текст рукописи следовало бы разбить на четыре главы, а именно вторую главу рукописи следовало бы разбить на две отдельные главы: вторую главу посвятить представлению метода построения деревьев решений с применением линейных и нелинейных разделителей, а третью главу посвятить оценке обобщающей способности композиций деревьев решений с линейными и нелинейными разделителями. Размеры глав в рукописи не пропорциональны по объему, так текст первой главы диссертации имеет избыточный объем.

2. В работе не рассмотрены на достаточном уровне вопросы выбора критериев останова при построении деревьев решений, а также используемые методы для обработки пропущенных значений, что не позволяет в полной мере сделать выводы о качестве работы предложенных алгоритмов.

3. Как известно выбор величины параметра регуляризации является основной проблемой при использовании регуляризирующих алгоритмов. Оценка качества и сложности метода оптимизации в предложенном автором алгоритме обучения дерева с линейными и нелинейными разделителями определяется параметром регуляризации C, который подбирается эмпирическим путем для каждого набора данных, непонятно почему автор не использовал численные или эвристические методы или способы оценки, которые позволили с приемлемой точностью определить оптимальное значение параметра регуляризации?

4. В рукописи приведены результаты экспериментальных исследований предложенных методов на размеченных наборах данных, но только для тех типов разделителей, с помощью которых получены максимальные оценки качества классификации. Не указаны оценки сложности обучения деревьев решений с помощью алгоритмов обучения, предложенных другими авторами. Отсутствие этих данных не позволяет сопоставить уровни прироста качества классификации и увеличения сложности алгоритмов обучения при использовании различных видов разделителей (линейных, нелинейных, одномерных) и методов построения деревьев решений.

5. В диссертации отсутствуют результаты оценки применения подобранного подхода к усилению и снижению сложности ансамблей к случайным лесам деревьев решений с одномерными разделителями.

6. Текст диссертации и автореферата содержит опечатки, грамматические ошибки и ошибки при оформлении формул, таблиц, рисунков.

Отмеченные недостатки не влияют на общую положительную оценку работы.

Заключение

Диссертационная работа Д.А. Девяткина «Построение ансамблей деревьев решений с использованием линейных и нелинейных разделителей» является завершенным научным исследованием и представляет научно-обоснованное решение задачи, имеющей значение для развития исследуемой отрасли знаний. Результаты, представленные в работе, имеют теоретическое обоснование, а также подтверждены практически в ходе экспериментальных исследований. Основные результаты и выводы диссертации получены автором лично, опубликованы автором в рецензируемых изданиях, входящих в Перечень ВАК, реализованы в виде программного обеспечения, апробированы на международной конференции и научных семинарах. Диссертационная работа соответствует паспорту специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей». Автореферат в достаточной мере отражает содержание работы.

Диссертация Девяткина Дмитрия Алексеевича выполнена на высоком научном уровне и соответствует требованиям п.9. Положения о присуждении ученых степеней, а ее автор заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

18.11.2022 г.

Официальный оппонент:

Доктор технических наук Вохминцев Александр Владиславович,

Заведующий научно-исследовательской лаборатории «Интеллектуальные информационные технологии и системы», профессор кафедры информационных технологий и экономической информатики Федерального государственного бюджетного учреждения высшего образования «Челябинский государственный университет»