ОТЗЫВ

официального оппонента Муравьёва Сергея Борисовича на диссертацию Ахматшина Фарида Галиулловича «Модели и алгоритмы автоматической группировки объектов для систем анализа и хранения данных на основе методов семейства k-средних», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.1. Системный анализ, управление и обработка информации, статистика

Актуальность темы исследования

Работа Ахматшина Ф.Г. посвящена исследованию методов автоматической группировки (кластеризации) данных и приближенного поиска ближайших соседей в контексте современных технологий анализа и хранения информации. В работе рассматривается применение этих методов в различных сферах, таких как анализ данных неразрушающих тестовых испытаний промышленной продукции, архивация данных в системах хранения данных и векторных базах данных. Подобные задачи сводятся к задачам минимизации значения некоторого критерия на основе выбранной меры подобия. Кроме того, эффективность решения подобных задач зависит от применяемых способов предварительной подготовки данных, таких как нормализация. Объемы накапливаемых и обрабатываемых данных постоянно возрастают, в связи с чем для их решения требуется разработка новых моделей и алгоритмов машинного обучения, в том числе автоматической группировки, в частности – алгоритмов машинного обучения для повышения эффективности управления в системах хранения данных.

Общая характеристика работы

Диссертация Ахматшина Ф.Г. представлена на 130 страницах, включая приложение, основной текст состоит из введения, пяти разделов и заключения.

Первый раздел посвящен вопросам нормализации данных в контексте автоматической группировки промышленной продукции, с акцентом на электрорадиоизделия (ЭРИ), для формирования однородных производственных партий на основе модели k-средних. Также рассматривается разработка нового подхода к нормализации данных промышленной продукции, который интегрирует нормализацию по допустимым значениям параметров оцениваемых характеристик продукции с оценкой Джеймса-Штейна.

Второй раздел посвящен разработке алгоритма кластеризации ЭРИ), который решает задачу разделения смешанной партии ЭРИ на однородные группы на основе результатов неразрушающих тестов. В данном исследовании рассматривается задача определения радиуса локальных концентраций (сгущений) в алгоритме кластеризации с заранее заданным числом кластеров, используя жадную эвристическую процедуру. Результаты были сравнены с различными методами нормализации данных для решения задачи кластеризации методом k-средних.

Третий раздел посвящен разработке нового алгоритма кластеризации, основанного на жадной агломеративной процедуре для построения индекса

векторной базы данных. Вычислительные эксперименты демонстрируют, что индекс, сгенерированный этим алгоритмом, достигает более высокой оценки полноты по сравнению с алгоритмами k-средних и k-means++.

Четвертый раздел фокусируется на создании инновационного алгоритма для автоматической группировки повторяющихся фрагментов блоков данных, интегрирующего метод k-средних и локально-чувствительное хэширование (LSH), предназначенного для применения в системах хранения данных.

Пятый раздел посвящен разработке новой процедуры инициализации центров кластеров для алгоритмов кластеризации, представляющей собой модификацию алгоритма k-means++. Полученная процедура инициализации применяется к большим данным. Учитывая меньшие вычислительные затраты, связанные с алгоритмом k-средних по сравнению с агломеративным алгоритмом, их применение в сочетании с модифицированным алгоритмом инициализации k-means++ к большим данным позволит уменьшить вычислительную сложность агломеративного алгоритма и улучшить его производительность.

В заключении сформулированы основные выводы и результаты, показано, что новые решения, повышающие эффективность (точность и стабильность результатов, т.е. улучшение достигаемых значений целевой функции за заданное время) алгоритмов автоматической группировки объектов при большом объеме входных данных привели к достижению цели диссертации.

Достоверность в обосновании научных положений, выводов и рекомендаций

В диссертационной работе Ахматшина Ф.Г. произведена разработка новых алгоритмов автоматической группировки, используемых в системах анализа и хранения данных, на повышение эффективности алгоритмов кластеризации при обработке больших данных в системах автоматической группировки объектов, в том числе в составе векторной СУБД и подсистем компрессии данных в составе систем хранения данных. Автор использует научные методы для обоснования полученных результатов. Достоверность подтверждается применением современных методов исследования, которые были использованы в большом наборе экспериментов.

Основные положения и результаты диссертационной работы докладывались и обсуждались на четырёх международных семинарах и конференциях, по теме диссертации опубликовано 14 научных работ в различных изданиях, индесируемых в ВАК, Scopus.

Теоретическая значимость

Теоретическая значимость состоит в дополнении эффективных алгоритмов решения задач автоматической группировки, а также алгоритмов предобработки данных для таких задач.

Практическая ценность

Исследования, приведенные в диссертационной работе, дополняют модельно-алгоритмический инструментарий, используемый в системах анализа данных результатов тестирования образцов промышленной продукции с повышенными требованиями качества, в частности – электронной компонентной базы космического применения, и могут использоваться в соответствующих

испытательных технических центрах. Кроме того,новая процедура инициализации центров кластеров для алгоритмов кластеризации, имеет универсальный характер и может применяться при обработке больших данных в любых системах автоматической группировки объектов. Новые алгоритмы кластеризации применяются для построения индекса для векторной базы данных и для разработки модели оптимального использования дискового пространства с учетом компрессии данных.

Оценка новизны результатов и выводов

Новыми научными результатами являются:

– подход к нормализации данных для предобработки входных данных, используемых в системах анализа данных результатов неразрушающих испытаний образцов промышленной продукции, комбинирующий нормализацию по допустимым значениям параметров оцениваемых значений продукции и оценке Джеймса-Штейна;

– алгоритм кластеризации системы анализа данных электрорадиоизделий на основе данных тестовых испытаний с использованием жадной эвристической процедуры выбора радиуса локальных концентраций по размеченным данным;

– алгоритм кластеризации для создания индекса векторной базы данных предназначенного для построения приближенного индекса поиска ближайших соседей, как компромисс между точностью и временем вычислений, значительно улучшающий метрику полноты в задачах приближенного поиска ближайших соседей;

– алгоритм автоматической группировки повторяющихся фрагментов блоков данных для использования в системах хранения данных на основе алгоритма k-средних совместно с применением локально чувствительного хэширования LSH;

– процедура инициализации центров кластеров для алгоритмов кластеризации, способная быстро находить приемлемое начальное решение при большом объеме данных.

Замечания по диссертационной работе

1. В работе недостаточно внимания уделено сравнению с другими современными алгоритмами кластеризации, такими как DBSCAN, это ограничивает понимание преимуществ и недостатков предложенных алгоритмов в контексте существующих решений.

2. Формальный анализ временной и пространственной сложности предложенных алгоритмов представлен фрагментарно. В работе отсутствует анализ временной сложности и вычислительных ресурсов, необходимых для выполнения предложенных подходов и алгоритмов.

3. В работе проведены вычислительные эксперименты, однако их объем и охват тестируемых сценариев можно расширить. Следовало бы протестировать алгоритмы на более разнообразных наборах данных, включая не только промышленные, но и стандартные датасеты из области машинного обучения.

4. В диссертации отсутствует детальное рассмотрение вопросов масштабируемости предложенных алгоритмов на больших объемах данных. Следует провести оценку их эффективности на сверхбольших выборках и сравнить с state-of-the-art методами.

5. В таблицах 2.1 – 2.8 неясен выбор временных промежутков, также не совсем понятно, почему не сократить запись до вида *среднее* $\pm$ *стандартное отклонение*, или привести в качестве результатов график типа «ящик с усами».

Заключение о соответствии диссертации требованиям и критериям, установленным Положением о порядке присуждения ученых степеней

Несмотря на приведенные замечания, диссертационная работа Ахматшина Ф.Г. является завершенным научным исследованием.

Представленная диссертационная работа отвечает требованиям п.9 «Положения о порядке присуждения ученых степеней» Постановления Правительства Российской Федерации от 24.09.2013 г. № 842, предъявляемым к кандидатским диссертациям, а ее автор Ахматшин Фарид Галиуллович заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.1 – Системный анализ, управление и обработка информации, статистика.

Официальный оппонент,
доцент института прикладных
компьютерных наук,
ФГАОУ ВО «Национальный
исследовательский
университет ИТМО»,
канд. техн. наук,
+7 (900) 644-22-06
smuravyov@itmo.ru

С.Б. Муравьёв

01 марта 2025 г.

Адрес организации:
194101, Россия, Санкт-Петербург, Кронверкский пр., д. 49, лит. А.

Подпись канд. техн. наук, доцента Муравьёва Сергея Борисовича заверяю