

ОТЗЫВ

официального оппонента доктора технических наук,
профессора КРАВЦА Олега Яковлевича

на автореферат диссертации Ахматшина Фариды Галиуллиновича на тему
«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

Диссертационная работа направлена на разработку новых алгоритмов автоматической группировки, используемых в системах анализа данных, в том числе выполняющих функции управления системами хранения данных, а также на повышение эффективности алгоритмов автоматической группировки при обработке больших объемов данных, в том числе в составе подсистем компрессии данных в составе систем хранения данных.

Актуальность диссертационной работы

Развитие систем анализа данных требует разработки адекватных методов обработки и предобработки данных (в том числе, нормализации), новых моделей и алгоритмов автоматической группировки для той или иной предметной области. Вследствие увеличения объемов обрабатываемых и анализируемых данных усложняется внутренняя структура систем анализа данных, включая подсистемы хранения данных, требуя применения алгоритмов машинного обучения, в том числе алгоритмов кластерного анализа, для повышения эффективности управления в системах хранения данных (дисковые массивы, векторные базы данных). При этом увеличение объема обрабатываемых данных выявляет низкую вычислительную эффективность существующих алгоритмов. Решению задач в этих областях и посвящена данная работа.

Основными научными результатами диссертационного исследования Ахматшина Ф.Г. являются:

1. Предложен новый 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 характеристик продукции и оценки Джеймса–Штейна.

2. Предложен новый алгоритм кластеризации для системы анализа данных электрорадиоизделий на основе данных тестовых испытаний с использованием жадной эвристической процедуры выбора радиуса локальных концентраций по размеченным данным.

3. Предложен новый алгоритм кластеризации для построения индекса векторной базы данных для приближенного поиска ближайших соседей, обеспечивающий компромисс между точностью и временем вычислений, существенно улучшает метрику полноты в задачах приближенного поиска ближайших соседей.

4. Предложен новый алгоритм автоматической группировки повторяющихся фрагментов блоков данных на основе алгоритма k -средних совместно с локально-чувствительным хэшированием (LSH), обеспечивающий увеличение эффективности сжатия данных в системах хранения данных.

5. Предложена новая процедура инициализации центров кластеров для алгоритмов автоматической группировки больших объемов данных, использующая вспомогательную структуру данных – массив слагаемых для вычисления суммы квадратов расстояния.

Практическая ценность состоит в дополнении модельно-алгоритмического инструментария, используемого 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.

Обоснованность и достоверность результатов

Достоверность теоретических результатов и выводов проведенного автором исследования определяется тщательной проработкой проблем в ходе решения всех решенных автором задач исследования, корректным применением методов системного анализа, теории тепловых процессов, методов оптимизации, теории вероятностей и математической статистики, теории автоматического управления, а также методологии создания прикладных программных систем. Обоснованность практических выводов и эффективность разработанных алгоритмов подтверждена внедрением в реальном производстве. Оригинальность и новизна полученных результатов подтверждается анализом научно-технической литературы по направлению диссертационного исследования, а также значительным количеством научных проектов и грантов соискателя, поддержанных как на государственном уровне, так и различными грантовыми фондами.

Автором опубликовано 14 научных работ по теме диссертации, в том числе 5 статей в научных изданиях, входящих в Перечень ВАК (категории K2), 5 статей в изданиях, индексируемых в Web of Science и Scopus, 4 статьи в других изданиях. Получено свидетельство о государственной регистрации программ для ЭВМ.

Замечания

1. В работе отсутствует описание методов семейства k-средних, не описаны границы и состав этого семейства.

2. В диссертации (на стр. 22, раздел «Результаты вычислительных экспериментов») статистическую значимость различий в значениях индекса Рэнда (RI) в табл. 1.2 следовало бы доказать с применением методов доказательства статистических гипотез.

3. В диссертации большие скачки стандартного отклонения RI нивелируют рост значений RI . Поэтому выводы по главе 1 следовало бы обосновывать более формально, с применением методов доказательства статистических гипотез.

4. В формуле (1.2) не объясняется, почему в знаменателе дроби используется среднее значение μ вместо координаты центра кластера c_j .

5. В таблицах 3.1 и 3.2 исследуется построение индекса с различным числом центров (от 16 до 4096), а на рис. 3.1 и 3.2 график построен по результатам построения индекса с 1024-мя центрами. Неясен выбор данных именно для указанного количества центров.

Приведенные замечания не снижают теоретической значимости и практической ценности диссертационного исследования, не противоречат основным защищаемым понижениям и не снижают общей высокой оценки проделанной работы.

Выводы

Диссертация Ахматшина Ф.Г. посвящена важной и актуальной задаче автоматической группировки (кластеризации) в быстроразвивающейся сфере деятельности по анализу данных неразрушающих тестовых испытаний промышленной продукции, в системах архивации данных в системах хранения данных, в векторных базах данных. Диссертация Ахматшина Ф.Г. – самостоятельная, выполненная лично автором, законченная научно-квалификационная работа в которой на основании выполненных автором исследований изложены новые подходы, алгоритмы и процедуры, внедрение которых вносит значительный вклад в развитие страны за счет повышения эффективности (вычислительной производительности, а также повышения качества результата по внешним и внутренним критериям качества) алгоритмов автоматической группировки объектов для систем анализа и хранения данных.

Автореферат и опубликованные статьи достоверно отражают основное содержание диссертации и полностью ей соответствуют.

Диссертация Ахматшина Ф.Г. «Модели и алгоритмы автоматической группировки объектов для систем анализа и хранения данных на основе методов семейства k -средних» полностью

удовлетворяет требованиям Министерства науки и высшего образования Российской Федерации, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, в частности п.п. 9-14 действующего «Положения о присуждении ученых степеней», и соответствует паспорту научной специальности 2.3.1 – Системный анализ, управление и обработка информации, статистика, а её автор Ахматшин Фарид Галиуллович заслуживает присуждения ему ученой степени кандидата технических наук.

Официальный оппонент

Профессор кафедры автоматизированных и вычислительных систем
ФГБОУ ВО «Воронежский государственный технический университет»,
доктор технических наук, профессор

_____ Кравец Олег Яковлевич
«06» ~~февраля~~ 2025 г.

Адрес: 394006, г. Воронеж, ул. 20-летия Октября, 84
ФГБОУ ВО «Воронежский государственный технический университет»
(ФГБОУ ВО «ВГТУ»)
Тел.: +7(473)271-59-05; +7(473)271-52-68
E-mail: rector@cchgeu.ru, csit@bk.ru

Я, Кравец Олег Яковлевич, даю согласие на включение моих персональных данных в документы, связанные с работой диссертационного совета, и их дальнейшую обработку.

_____ Кравец Олег Яковлевич

Подпись Кравца Олега Яковлевича заверяю:

Проректор по науке и инновациям,
д-р техн. наук, доцент



_____ Башкиров Алексей Викторович