ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ»

На правах рукописи

АТЛАСОВ Денис Игоревич

УПРАВЛЕНИЕ ПРОЦЕССАМИ ОБРАБОТКИ ГЕТЕРОГЕННЫХ ДАННЫХ В РАМКАХ ИНФОРМАЦИОННЫХ СИСТЕМ С МНОГОМЕРНЫМИ АТРИБУТАМИ НА ОСНОВЕ МЯГКОЙ МАКСИМИННОЙ ОЦЕНКИ

Специальность: 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей

Диссертация
на соискание учёной степени кандидата технических наук

Научный руководитель:
д.т.н., профессор Кравец Олег Яковлевич

Воронеж – 2025

Оглавление

Введение

Актуальность темы. Сложные и разнообразные программные системы требуют постоянного анализа циркулирующих в них данных, так как именно данные определяют полезность и значимость результатов работы программных систем. Рекомендательные системы на основе гетерогенных информационных сетей обеспечивают единый подход к объединению различной вспомогательной информации, которую можно комбинировать с основными алгоритмами рекомендаций для эффективного повышения производительности. Актуальна и проблема извлечения общего сигнала из разнородных данных. Поскольку гетерогенность преобладает в крупномасштабных системах, цель - эффективный в вычислительном отношении оценщик с хорошими статистическими свойствами при различной степени неоднородности данных. Большой вклад в разработку методов и средств управления гетерогенными данными больших вычислительных систем и сетей внесли Бостром Н., Лекун Я., Маркус Г., Bishop C., Goodfellow I.

Одной из актуальных предметных областей задач управления данными гетерогенных систем является управление знаниями. Размер зерна знаний в приближенном пространстве напрямую влияет на неопределенность приблизительного множества. С точки зрения интуитивного понимания, чем больше объем знаний, тем меньше информации, тем больше будет неопределенность; чем меньше объем знаний, тем больше информации, тем меньше будет неопределенность. Это центр исследований в области искусственного интеллекта. Важным является измерение неопределенности для гетерогенной информационной системы, которое отражает способность этой системы к классификации и повышению точности классификации данных.

Таким образом, актуальность темы диссертационного исследования продиктована необходимостью разработки моделей и методов управления

гетерогенными данными информационных систем с многомерными атрибутами на основе мягкой максиминной оценки и активного обучения.

Тематика диссертационной работы соответствует научному направлению ФГБОУ ВО «Воронежский государственный технический университет» «Вычислительные комплексы и проблемно-ориентированные системы управления».

Целью работы является разработка моделей и методов управления гетерогенными данными информационных систем с многомерными атрибутами на основе мягкой максиминной оценки и активного обучения.

Задачи исследования. Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ проблем управления гетерогенными данными информационных систем с многомерными атрибутами на основе мягкой максиминной оценки и активного обучения.

2. Разработать мягкую максиминную оценку для гетерогенных данных, содержащих уникальные вариационные компоненты, обеспечивающую сохранение статистических свойств и лучшую вычислительную эффективность

3. Предложить архитектуру системы управления гетерогенной информационной системой, обеспечивающую эффективную редукцию многомерных анализируемых атрибутов.

4. Создать алгоритм обнаружения наилучшей модели машинного обучения для анализа гетерогенных данных, обеспечивающий сокращение объемов данных и повышение точности анализа совокупности данных.

5. Разработать алгоритм идентификации нулевых значений в гетерогенных базах данных, обеспечивающий более эффективную и точную оценку нулевых значений.

Объект исследования: процессы обработки гетерогенных данных в рамках информационных систем с многомерными атрибутами.

Предмет исследования: средства математического и программного обеспечения процессов обработки гетерогенных данных в рамках информационных систем с многомерными атрибутами.

Методы исследования. При решении поставленных в диссертации задач использовались методы теории вероятностей, теории принятия решений, а также методы объектно-ориентированного программирования.

Тематика работы соответствует следующим пунктам паспорта специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»: п. 4. «Интеллектуальные системы машинного обучения, управления базами данных и знаний, инструментальные средства разработки цифровых продуктов»; п. 9. Модели, методы, алгоритмы, облачные технологии и программная инфраструктура организации глобально распределенной обработки данных.

Научная новизна работы. В диссертации получены следующие результаты, характеризующиеся научной новизной:

- мягкая максиминная оценка для гетерогенных данных, содержащих уникальные вариационные компоненты, отличающаяся извлечением надежных данных из разнородных групп, и обеспечивающая сохранение статистических свойств и лучшую вычислительную эффективность;

– архитектура программной системы управления гетерогенной информационной системой, отличающаяся использованием отношения эквивалентности на множестве объектов для измерения неопределенности системы, и обеспечивающая редукцию многомерных анализируемых атрибутов на основе грануляции информации и информационной энтропии;

– алгоритм выбора наилучшей модели машинного обучения для анализа гетерогенных данных, отличающийся применением активного обучения и ансамблевых методов для классификаторов при решении

проблем анализа разнородных данных, обеспечивающий сокращение объемов данных и повышение точности анализа совокупности данных;

- алгоритм идентификации нулевых значений в гетерогенных базах данных, отличающийся предварительной классификацией исходных данных на основе взвешенных значений и обеспечивающий более эффективную и точную оценку нулевых значений в среднем на 7.6%.

Теоретическая и практическая значимость исследования заключается в разработке моделей и методов управления гетерогенными данными информационных систем с многомерными атрибутами на основе мягкой максиминной оценки и активного обучения.

Теоретические результаты работы могут быть использованы в проектных и научно-исследовательских организациях, занимающихся проектированием программных систем с гетерогенными базами данных с многомерными атрибутами.

Положения, выносимые на защиту

1. Мягкая максиминная оценка для гетерогенных данных, содержащих уникальные вариационные компоненты, сохраняет статистические свойства наборов данных и обеспечивает лучшую вычислительную эффективность по сравнению с максиминной оценкой.

2. Архитектура программной системы управления гетерогенной информационной системой реализует эффективную редукцию многомерных анализируемых атрибутов на основе грануляции информации и информационной энтропии.

3. Алгоритм выбора наилучшей модели машинного обучения для анализа гетерогенных данных обеспечивает сокращение объемов анализируемых данных и повышение точности анализа совокупности данных.

4. Алгоритм идентификации нулевых значений в гетерогенных базах данных обеспечивает более эффективную и точную оценку нулевых

значений в среднем на 7.6%.

Результаты внедрения. Основные результаты внедрены в ООО М-Сервис (г. Воронеж) при проектировании систем управления гетерогенными программными системами, в учебный процесс Воронежского государственного технического университета в рамках дисциплин: «Вычислительные машины, системы и сети», «Информационные сети и телекоммуникационные технологии», а также в рамках курсового и дипломного проектирования.

Апробация работы. Основные положения диссертационной работы докладывались и обсуждались на следующих конференциях: XXVIII-th и XXX-th International Open Science Conference «Modern informatization problems in simulation and social technologies» (Yelm, WA, USA, 2023, 2025); Международной молодежной научной школе «Оптимизация и моделирование в автоматизированных системах» (Воронеж, 2023); XXIX-th International Open Science Conference «Modern informatization problems in the technological and telecommunication systems analysis and synthesis» (Yelm, WA, USA, 2024); Международной научно-практической конференции, «Интеллектуальные информационные системы» (Воронеж, 2024); VI Всероссийской научно-практической конференции «Информационные технологии в экономике и управлении» (Махачкала, 2024), а также на научных семинарах кафедры автоматизированных и вычислительных систем ВГТУ (2023-2026 гг.).

Достоверность результатов обусловлена корректным использованием теоретических методов исследования и подтверждена результатами сравнительного анализа данных вычислительных и натурных экспериментов.

Публикации. По результатам диссертационного исследования опубликовано 19 научных работ, в том числе 7 – в изданиях, рекомендованных ВАК РФ (из них 1 – в издании, индексируемых в WoS и

одно свидетельство о регистрации программы для ЭВМ). В работах, опубликованных в соавторстве и приведенных в конце автореферата, лично автором получены следующие результаты: [2, 3, 8, 10, 11] - мягкая максиминная оценка для гетерогенных данных, содержащих уникальные вариационные компоненты, отличающаяся извлечением надежных данных из разнородных групп, и обеспечивающая сохранение статистических свойств и лучшую вычислительную эффективность; [1, 6, 7] - архитектура программной системы управления гетерогенной информационной системой, отличающаяся использованием отношения эквивалентности на множестве объектов для измерения неопределенности системы, и обеспечивающая редукцию множества анализируемых атрибутов; [5, 12, 13, 15, 17] - алгоритм выбора наилучшей модели машинного обучения для анализа гетерогенных данных, отличающийся применением активного обучения и ансамблевых методов для классификаторов при решении проблем анализа разнородных данных, обеспечивающий сокращение объемов данных и повышение точность анализа совокупности данных; [4, 9, 16, 18] - алгоритм идентификации нулевых значений в гетерогенных базах данных, отличающийся предварительной классификацией исходных данных на основе взвешенных значений и обеспечивающий более эффективную и точную оценку нулевых значений.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав, заключения, списка литературы из 207 наименований. Работа изложена на 185 страницах.

Основное содержание работы

Во введении обоснована актуальность исследования, сформулированы его цель и задачи, научная новизна и практическая значимость полученных результатов, приведены сведения об апробации и внедрении работы.

В первой главе исследуются проблемы управления гетерогенными данными информационных систем с многомерными атрибутами на основе мягкой максиминной оценки и активного обучения. Отмечено, что повысить эффективность такого управления можно путем применения мягкой максиминной оценки для гетерогенных данных, содержащих уникальные вариационные компоненты, архитектуры системы управления гетерогенной информационной системой с редукцией многомерных анализируемых атрибутов, выбора наилучшей модели машинного обучения для анализа гетерогенных данных, создания алгоритмов идентификации нулевых значений в гетерогенных базах данных. Сформулирована цель и задачи исследования.

Вторая глава посвящена развитию инструментов интерполяции мягкой максиминной оценки для гетерогенных данных.

Рассматривается проблема извлечения общего сигнала из разнородных данных. Поскольку гетерогенность преобладает в крупномасштабных системах, цель - эффективный в вычислительном отношении оценщик (решение) с хорошими статистическими свойствами при различной степени неоднородности данных.

Извлечение общего надежного сигнала из данных, разделенных на разнородные группы, является сложной задачей, когда каждая группа - в дополнение к сигналу - содержит большие уникальные вариационные компоненты. Ранее максиминная оценка была предложена в качестве надежного метода при наличии неоднородного шума. Предлагается мягкая максиминная оценка максимального значения в качестве привлекательной с вычислительной точки зрения альтернативы, направленной на достижение баланса между объединенной оценкой и (жесткой) оценкой максимального значения. Метод мягкой максиминной оценки предоставляет диапазон оценок, управляемых параметром $\xi>0$, который интерполирует объединенную оценку наименьших квадратов и

максиминную оценку. Устанавливая соответствующие теоретические свойства, утверждается, что метод мягкой максиминной оценки является статистически обоснованным и привлекательным с точки зрения вычислений. Демонстрируется на реальных и смоделированных данных, что мягкая максиминная оценка может предложить улучшения как по сравнению с объединенными методами наименьших квадратов (МНК), так и по сравнению с жестким максимином с точки зрения производительности прогнозирования и вычислительной сложности.

Третья глава посвящена измерению неопределенности гетерогенных данных и редукции атрибутов в гетерогенных информационных системах.

Исследуется измерение неопределенности для разнородных данных и приводится его применение для редукции атрибутов. Сначала предлагается концепция гетерогенной информационной системы (HIS).

Затем строится отношение эквивалентности на множестве объектов. Затем исследуется измерение неопределенности для HIS, проводится численный эксперимент, в котором проведен дисперсионный анализ, корреляционный анализ, а также тест Фридмана и тест Бонферрони–Данна в статистике. В качестве применения предложенных мер изучается уменьшение атрибутов в HIS, и предлагаются соответствующие алгоритмы и их анализ.

Установлено что грануляция информации G и грубая энтропия E_r монотонно уменьшаются по мере увеличения отношений эквивалентности. В то же время количество информации E и информационная энтропия H монотонно увеличиваются с увеличением отношений эквивалентности. Это означает, что неопределенность нечеткого отношения уменьшается по мере увеличения отношений эквивалентности. Таким образом, грануляция информации G, грубая энтропия E_r, объем информации E и

информационная энтропия H могут быть применены для измерения неопределенности HIS.

Разработаны алгоритмы редукции HIS, основанные на грануляции информации и информационной энтропии.

В главе 4 представлены особенности управления гетерогенными данными.

В новом подходе, основанном на пуле, процесс маркировки выполняется пакетами, выбранными из пула немаркированных данных; после маркировки каждого пакета алгоритм обучается с использованием этих пакетов; и этот процесс повторяется с набором новых образцов до тех пор, пока эффективность обучения не улучшится.

Этап предварительной обработки включает в себя методы удаления избыточных и незначительных данных для минимизации объема объекта.

Этап извлечения данных использует методы являются BOW, TF-IDF и word2vec.

Используется выборка с неопределенностью, при которой из массива немаркированных данных отбираются наименее достоверные образцы. Затем эти образцы обрабатываются специалистом, и этот процесс повторяется до тех пор, пока не будут помечены все партии. После каждой партии точность проверяется до тех пор, пока не будет достигнута требуемая степень точности. Затем помеченные образцы данных классифицируются. Этот процесс повторяется для маркировки большего количества немаркированных образцов до тех пор, пока точность модели не повысится до приемлемого уровня.

Далее в работе проведена оценка неопределенностей нулевого значения гетерогенной базы данных на основе искусственного интеллекта.

В качестве инструмента для отображения реального мира база данных использует нулевые значения (null), чтобы выразить проблему отсутствия информации. Для решения проблемы нулевого значения в базе

данных с неопределенностью предлагается алгоритм оценки нулевого значения на основе искусственного интеллекта. Сначала анализируются характеристики неопределенной базы данных, затем строится модель поиска потерянной информации, а оценка пустого значения базы данных завершается выбором признаков и преобразованием данных, кластеризацией искусственного интеллекта, вычислением степени влияния, оценкой шага пустого значения и другими методами. Наконец, он анализирует временную сложность алгоритма и устраняет проблему низкого эффекта оценки традиционных алгоритмов. Результаты показывают, что предложенный алгоритм обладает более высокой точностью, чем традиционный алгоритм.

В заключении представлены основные результаты работы.

1. Проблемы управления гетерогенными информационными системами и данными

Как важный способ уменьшить информационную перегрузку, рекомендательная система направлена на то, чтобы отфильтровывать не относящуюся к делу информацию для пользователей и предоставлять им товары, которые могут их заинтересовать. В последние годы было предложено провести все больший объем работ по внедрению вспомогательной информации в рекомендательные системы, чтобы уменьшить разреженность данных и проблемы с холодным запуском. Среди них рекомендательные системы на основе гетерогенных информационных сетей (HIN) обеспечивают единый подход к объединению различной вспомогательной информации, которую можно комбинировать с основными алгоритмами рекомендаций для эффективного повышения производительности. В частности, сначала мы представляем концепции, связанные с рекомендательными системами, гетерогенными информационными сетями и рекомендациями на основе HIN. Во-вторых, представлено более 70 методов, классифицированных в соответствии с моделями или сценариями применения, и описаны репрезентативные методы символически. В-третьих, обобщены базовые наборы данных и открытый исходный код. Наконец, обсуждается несколько потенциальных направлений исследований и завершаем наш обзор.

1.1. Гетерогенные данные и рекомендательные системы

С быстрым развитием эпохи Интернета люди перегружены большим количеством не относящейся к делу информации, что также называется информационной перегрузкой [1.28]. Информационная перегрузка сильно влияет на эффективность людей в получении полезной информации, и

рекомендательные системы [1.50] стремятся решить эту проблему, предоставляя пользователям фильтр товаров, которые могут их заинтересовать. За десятилетия разработки рекомендательные системы успешно использовались во многих областях, таких как электронная коммерция [1.53] и мультимедиа [1.22].

Распространенным подходом рекомендательных систем к решению проблемы разреженности данных и холодного запуска является введение вспомогательной информации [1.65]. Например, на основе онлайн-сервиса социальных сетей мы можем интегрировать социальные интересы или социальное доверие между пользователями в качестве вспомогательной информации, которая также называется социальной рекомендацией [1.80]. Более того, социальные сети, основанные на местоположении (LBSN) [1.3], дополнительно добавляют информацию о местоположении в социальную структуру, которая также может быть использована для улучшения систем рекомендаций. Фактически, интеграция различной вспомогательной информации с рекомендательными системами унифицированным образом стала важной задачей.

Гетерогенные информационные сети (HIN) [1.58] представляют собой сложные сети, состоящие из множества типов узлов или ребер. Поскольку рекомендательную систему саму по себе можно рассматривать как двудольный граф, состоящий из пользователей и элементов, а множество вспомогательной информации также имеет сложную сетевую структуру, рекомендательную систему со вспомогательной информацией часто можно рассматривать как сложную систему взаимодействия. Следовательно, моделирование такой системы взаимодействия с разнородными информационными сетями не только естественным образом сохраняет сущности и взаимосвязи в рекомендательной системе, но также эффективно включает различную вспомогательную информацию, тем самым эффективно уменьшая разреженность данных и проблемы

холодного запуска, и в определенной степени улучшая интерпретируемость рекомендательных систем.

- Существует исследование, в котором представлены рекомендательные системы, объединяющие разнородную информацию [1.13].

1.2. Концепции рекомендательных систем для гетерогенных данных

1.2.1. Рекомендательные системы

Система рекомендаций [1.50] - это эффективный метод фильтрации информации, который предоставляет пользователям персонализированный контент, который может их заинтересовать.

В соответствии с различными стратегиями проектирования рекомендательные системы обычно можно разделить на рекомендации на основе совместной фильтрации, рекомендации на основе контента, рекомендации на основе знаний и гибридные рекомендации [1.29]. Рекомендация, основанная на совместной фильтрации (CF), относится к выработке рекомендаций, основанных на взаимодействии между пользователями и элементами, которые могут напрямую вычислять сходство между пользователями и элементами с помощью исторических моделей взаимодействия (т.е. CF на основе памяти) или использовать алгоритмы машинного обучения для получения представлений пользователя и элемента (т.е. CF на основе модели). Кроме того, рекомендация на основе контента относится к выработке рекомендаций на основе характеристик элементов, а рекомендация на основе знаний относится к выработке рекомендаций по поиску в базе знаний на основе формальных потребностей пользователей. Гибридная рекомендация относится к рекомендательной системе, которая использует несколько вышеупомянутых методов, тем самым сохраняя преимущества различных

методов одновременно.

Рекомендательные системы часто сталкиваются с нехваткой данных и проблемами холодного запуска. Разреженность данных относится к ограниченному взаимодействию между пользователями и элементами, в то время как холодный запуск относится к предоставлению рекомендаций для пользователей и элементов, у которых нет исторических записей о взаимодействии. Чтобы решить эти две проблемы, исследователи разработали различные алгоритмы совместной фильтрации, которые интегрируют вспомогательную информацию. Вспомогательная информация в основном делится на две категории: структурированная вспомогательная информация и неструктурированная вспомогательная информация [1.65]. Структурированная вспомогательная информация в основном относится к социальным сетям, графикам знаний, каталогам товаров и т.д., В то время как неструктурированная вспомогательная информация включает текст товара, изображения товара, видео товара и т.д.

Структурированная вспомогательная информация распространена повсеместно и содержит богатую семантическую информацию. Вопрос о том, как внедрить ее в рекомендательную систему, является предметом исследования. Многие исследователи разработали рекомендательные системы, которые подходят для получения определенной структурированной вспомогательной информации, такой как социальная рекомендация [1.79], социальная рекомендация на основе местоположения [1.3] и так далее. Было доказано, что эти методы значительно повышают производительность рекомендательной системы, но эти методы часто связаны с определенными типами вспомогательной информации и не являются универсальными. Разработка подхода к моделированию, который может интегрировать различную вспомогательную информацию, стала серьезной исследовательской задачей для рекомендательных систем.

1.2.2. Гетерогенные информационные сети

Многие реальные системы можно рассматривать как сложные сети, состоящие из множества типов сущностей и отношений. Традиционные методы исследования моделируют их как однородные сети, игнорируя неоднородность объектов и связей. Чтобы всесторонне моделировать богатую структурную и семантическую информацию в сложных системах, исследователи предложили концепцию гетерогенных информационных сетей [1.58] и успешно применили ее для моделирования различных задач интеллектуального анализа данных. Основные термины гетерогенных информационных сетей определяются следующим образом.

1.2.2.1. Информационная сеть

Информационную сеть [1.62] можно определить как ориентированный граф G={V, E}, где V представляет узлы, E представляет собой отношения (связи), и есть функции сопоставления φ: V→A и φ: E→R. Гетерогенная информационная сеть [1.62] относится к информационной сети с количеством типов сущностей |A|>1 или количеством типов отношений |R|>1. Рассмотрим пример моделирования гетерогенной информационной сети для систем рекомендаций, в которые включены пять типов сущностей (т.е. пользователь (U), группа (G), фильм (M), режиссер (D) и тип фильма (T)); и пять типов отношений и включены (т.е. “пользователь-группа" (UG), “пользователь-юзер” (UU), “пользователь-фильм” (UM), “тип фильма” (MT) и “режиссер фильма” (MD)). Гетерогенные информационные сети также часто называют гетерогенными графами. Напротив, информационная сеть с количеством типов сущностей |A|=1 и количеством типов отношений |R|=1 называется однородной информационной сетью или однородным графом.

1.2.2.2. Схема сети

Схема сети [1.62] относится к мета-шаблону информационной сети, которая также является ориентированным графом, и удовлетворяет функции отображения φ: V→A и φ: E→R, можно отметить как T_G=(A,R). Информационная сеть, соответствующая определенному металшаблону, является сетевым экземпляром сетевой схемы.

1.2.2.3. Мета-путь

Мета-путь [1.58] относится к определенному пути в сетевую схему, которая может быть записана в виде A_1 ¾® A_2 ¾® ... ¾® A_{l+1}. Мета-путь определяет составной путь R = R_1∘R_2...∘R_l от объекта A_1 к объекту A_l+1, где ∘ представляет операцию комбинирования отношений. Следовательно, по сравнению с отношением, метапуть описывает семантическую информацию более высокого уровня между двумя объектами. Ограниченный метапуть [1.57] относится к метапути с добавленными ограничениями объекта или отношения.

1.2.2.4. Метаструктура

Метаструктура [1.82] относится к направленному ациклическому графу в сетевой схеме T_G=(A,R), который определяет начальную сущность A_1 и целевую сущность A_{l+1}, и состоит из нескольких мета-путей, которые также называются мета-графом. Следовательно, по сравнению с мета-путем, мета-структура описывает семантическую информацию более высокого уровня между двумя объектами.

Благодаря очевидным преимуществам в интеллектуальном анализе богатой структуры и семантической информации, лежащей в основе реальных сложных систем, гетерогенные информационные сети широко

использовались для измерения сходства, кластеризации узлов, классификации, прогнозирования ссылок, ранжирования, объединения информации и других задач интеллектуального анализа данных [1.62] и достигли хороших результатов в вышеуказанных задачах.

1.2.3. Рекомендательные системы на основе HIN

Рекомендательную систему можно рассматривать как задачу прогнозирования связей в информационной сети. Некоторые традиционные рекомендательные системы основаны на моделировании двудольных графов, что затрудняет использование различной вспомогательной информации. Некоторые исследователи предложили методы объединения определенных типов вспомогательной информации, но они часто не являются универсальными. Рекомендательные системы на основе HIN успешно решили проблему моделирования разнородной вспомогательной информации и поведения при взаимодействии с пользователем унифицированным образом, что может не только эффективно уменьшить разреженность данных и проблемы холодного запуска в рекомендательных системах, но и значительно улучшить интерпретируемость рекомендательных систем, поэтому получило широкое внимание и применение.

В табл. 1.1 показаны значения различных мета-путей. Из приведенных выше примеров видно, что рекомендательная система, основанная на гетерогенной информационной сети, может полностью интегрировать различную структурированную вспомогательную информацию: интерактивную (U-M), социальную (U-U), атрибутивную (M-T).

Таблица 1.1

Значения и соответствующие рекомендательные модели метапутей

Метапуть	Семантическое значение	Модели рекомендации
UU	друзья целевого пользователя	Социальная рекомендация
UGU	пользователи из той же группы, что и целевой пользователь	Рекомендация участника
UMU	пользователи, которые применяют одну и ту же информацию с целевым пользователем	Рекомендация совместной работы
UMTMU	пользователи, которые применяют одну и ту же информацию, что и у целевого пользователя	Рекомендация контента

1.3. Таксономия на основе моделей

В этом разделе классифицируются существующие работы по трем основным категориям: измерение сходства, факторизация матриц и изучение графического представления. Кроме того, работы разделены на несколько подкатегорий, как показано в табл. 1.2.

Таблица 1.2

Таксономия, основанная на моделях

Систематика			**Исследования**
Измерение подобия	На основе отношений		21, 22, 31, 76, 69, 120
	На основе мета-путей		6, 27, 28, 58, 60, 81, 83
Матрица факторизации	На основе регуляризации		44, 65,67, 88, 109
	На основе нейронной матрицы факторизации		11, 31, 32
Граф представительства обучения	На основе двухступенчатой подготовки		53, 86, 122, 123
	На основе сквозного обучения	Работа с реляционными отношениями	8, 16, 50, 62, 63, 96, 104
		На основе мета-путей	10, 51, 55, 110, 111

1.3.1. Измерение сходства

Персонализированное сопоставление рекомендательных систем часто основано на измерении сходства объектов, где совместная

фильтрация вычисляет сходство на основе истории взаимодействия между пользователем и товаром. Ранние алгоритмы измерения сходства были определены только для однородных информационных сетей. Например, P-PageRank [1.31] оценивает вероятность перехода от исходного объекта к целевому объекту путем перезапуска случайных блужданий, а SimRank [1.30] оценивает сходство объектов по сходству соседей двух объектов.

Однако эти алгоритмы игнорируют различные типы объектов и соединений и не подходят для рекомендательных систем, моделируемых как гетерогенные информационные сети. Чтобы решить эту проблему, исследователи предложили серию алгоритмов для измерения сходства объектов в гетерогенных информационных сетях, в основном включая методы, основанные на отношениях, и методы, основанные на метапутях. Основываясь на этих двух типах алгоритмов измерения сходства для гетерогенных информационных сетей, исследователи предложили множество вариантов алгоритмов совместной фильтрации. В этой статье эти методы в совокупности называются методами, основанными на измерении сходства.

1.3.1.1. Методы измерения сходства, основанные на отношениях

Методы измерения сходства, основанные на отношениях, всегда используют алгоритм случайного блуждания, который предполагает, что релевантность элемента для пользователя может быть получена из вероятности случайного блуждания пользователя по элементу, а более высокая вероятность означает более высокую их релевантность. Например, RHSN [1.83] использует модель случайного блуждания для оценки глобальной значимости каждого объекта и предлагает алгоритм парного обучения для определения веса каждого типа отношений. Чтобы быть более конкретным, RHSN вычисляет оценку глобальной важности s на

основе матрицы переходов гетерогенного графа:

$$s_Y = \alpha \cdot E + (1-\alpha) \cdot \sum_{\lambda_{XY} \in \Lambda} \lambda_{XY} M_{XY}^T s_X \quad (1.1)$$

где X и Y - типы узлов, s_X и s_Y - векторы ранга для типов X и Y, α - параметр случайного перехода, E = (1/n,...,1/n)^T(1,...,1), λ_{XY} - вероятность перехода из типа X в тип Y, Λ - множество λ_{XY}, а M_{XY} - матрица перехода, соответствующая типу отношения XY.

Чтобы узнать значение каждого λ_{XY}, RHSN проектирует следующую целевую функцию:

$$\max L = \sum_{i,j \in A} \hat{y}_{ij} y_{ij}$$

$$\lambda_{UT} + \lambda_{UC} + \lambda_{UR} + \lambda_{UU} = 1,$$

$$\lambda_{TR} + \lambda_{TU} + \lambda_{TC} = 1,$$

$$\lambda_{CT} + \lambda CU + \lambda CR = 1, \quad (1.2)$$

$$\lambda_{RT} + \lambda_{RU} + \lambda_{RC} = 1,$$

$$\lambda_{XY} > 0$$

где y_{ij} - индикаторная функция для оцененного значения алгоритмом случайного блуждания, а $\hat{y}_{ij}$ - индикаторная функция для истинного значения обучающих данных.

После вычисления показателя глобальной важности с использованием модели случайного блуждания RHSN использует языковую модель для вычисления показателя релевантности между объектом o и запросом q для рекомендации:

$$P(q|o) = \prod_{t_i \in q} \left(w \cdot \frac{tf(t_i, o)}{|o|} + (1-w) \cdot \frac{tf(t_i, O)}{|O|} \right) \quad (1.3)$$

где o - профиль объекта, |*| - норма *, tf(t_i,*) - частота встречаемости t_i в *, $w = \frac{|o|}{|o| + |v|}$ - параметр в диапазоне [0, 1], v - средняя длина профиля объекта в O.

Кроме того, ECTD [1.17] вычисляет среднее время в пути между любой парой объектов в качестве измерения сходства, которое определяется как симметричная величина среднего числа шагов, необходимых случайному прохожему для достижения объекта в первый раз. OptRank [1.16] изучает как веса ребер, так и узлов, максимизируя среднее значение AUC. HeteRC [1.51] представляет гетерогенную информационную сеть в виде множественных матриц преобразования на основе отношений и использует многомерную цепочку Маркова для получения результата рекомендации для данного узла запроса. Div-HeteRec [1.46] использует случайное блуждание с улучшенными вершинами, чтобы избежать влияния множества соседних и влиятельных узлов на разнообразие рекомендаций. HeteLearn [1.35] изучает веса ссылок на основе случайного блуждания и байесовской технологии персонализированного ранжирования для достижения персонализированного моделирования пользовательских предпочтений.

1.3.1.2. Методы измерения сходства, основанные на мета-пути

Методы измерения сходства, основанные на отношениях, направлены на изучение вероятности перехода на уровень отношений, без явного использования семантической информации более высокого уровня. По сравнению с ними методы измерения сходства, основанные на мета-пути, вводят дополнительные предварительные знания, вводя значимые мета-пути вручную. Разработанные метапути удовлетворяют некоторым хорошим свойствам (таким как симметрия и самодиагностика), что полезно для сбора семантической информации на большом расстоянии. Существует три вида классических метрик сходства, основанных на мета-пути, то есть количество путей (например, PathSim [1.63]), случайное блуждание на основе пути (например, PCRW [1.47]) и попарное случайное

блуждание на основе пути (например, HeteSim [1.56]).

Чтобы применить вышеупомянутые метрики сходства на основе мета-пути к совместной фильтрации, основная идея состоит в том, чтобы использовать разные мета-пути для генерации большого количества возможных сходств, а затем вручную назначить или автоматически узнать веса возможных сходств. Следовательно, благодаря удобочитаемому мета-пути и весу различных мета-путей методы, основанные на мета-пути, также обладают преимуществом высокой интерпретируемости [1.26]. Например, как показано на рис. 1.1, SemRec [1.57] вычисляет прогнозируемые рейтинги $\hat{R}_{u,i}$ на основе присвоенных им весовых векторов на мета-путях следующим образом:

$$\hat{R}_{u,i} = \sum_{l=1}^{|P|} W_u^{(l)} \cdot \hat{R}_{u,i}^{(l)} \tag{1.4}$$

где $W_u^{(l)}$ означает вес предпочтений пользователя u на пути P_l. Следовательно, цель оптимизации определяется следующим образом:

$$\min L_3(W) = \frac{1}{2}\left\| Y \odot \left(R - \sum_{l=1}^{|P|} \mathrm{diag}\left(W^{(l)}\right) \hat{R}^{(l)} \right) \right\|_2^2 +$$

$$+ \frac{\lambda_1}{2} \sum_{l=1}^{|P|} \left\| W^{(l)} - \bar{S}^{(l)} W^{(l)} \right\|_2^2 + \frac{\lambda_0}{2} \|W\|_2^2 \tag{1.5}$$

$$W \geq 0$$

где W - рейтинг индикатора матрицы, $\odot$ - произведение Адамара, $\|\cdot\|_p$ - L^p-норма матрицы, λ_0 и λ_1 контролируемые параметры, $\bar{S}^{(l)}$ - нормированная схожесть пользователя по пути P_l, $\sum_{l=1}^{|P|} \left\| W^{(l)} - \bar{S}^{(l)} W^{(l)} \right\|_2^2$ - вес регуляризации, который требует весов пользователей соответствии со средней весов похожих пользователей.

Аналогично, WHyLDR [1.5] включает несколько типов мета-путей при взвешенном микшировании. X-MAP [1.20] вычисляет транзитивное замыкание сходства между элементами в нескольких доменах на основе

метапути. PathRank [1.39] расширяет PCRW до более общей формулы и применяет ее к системам рекомендаций. HeteRecom [1.55] вычисляет взвешенное сходство пользователей по метапути на основе HeteSim [1.56], а затем использует эвристический метод изучения веса для вычисления веса. MP-PRF [1.40] переупорядочивает бумажные объекты в гетерогенной сети цитирования на основе ограниченного метапути. AFG [1.21] извлекает все возможные мета-пути из сетевой схемы для рекомендации. ClusCite [1.52] использует как PathSim [1.63], так и метрику на основе случайного блуждания для генерации объектов и предлагает функцию оценки на основе кластера.

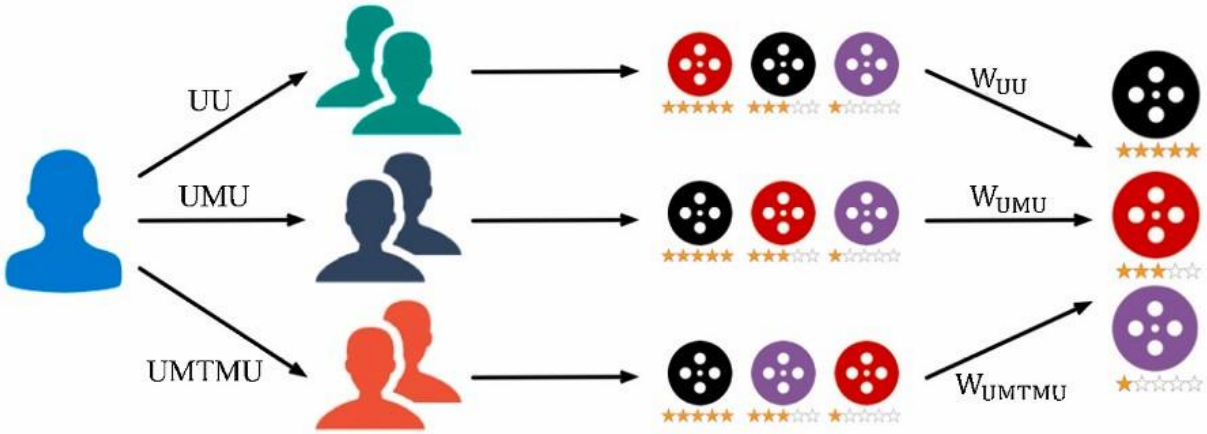


Рис. 1.1. Архитектура SemRec

1.3.1.3. Измерение гетерогенного сходства для рекомендательных систем на основе краткого содержания

Алгоритмы рекомендаций, обсуждаемые в этом разделе, являются самыми ранними работами, использующими измерение гетерогенного сходства для рекомендательных систем на основе CF, которым требуется большой объем интерактивной информации, и трудно предоставлять рекомендации пользователям и элементам, не имеющим подключения. Однако гетерогенные показатели подобия, представленные в этом разделе (например, Path-Sim [1.63] и HeteSim [1.56]), все еще влияют на последующие модели (например, матричную декомпозицию и нейронные

сети) и могут быть объединены с ними.

1.3.2. Факторизация матрицы

Чтобы предоставить рекомендации пользователям и элементам, которым не хватает подключения, исследователи рекомендательных систем предложили модель матричной факторизации. Основная идея состоит в том, чтобы извлечь скрытые векторы пользователей и товаров путем декомпозиции рейтинговой матрицы, а затем дать рекомендации, основанные на сходстве скрытых векторов.

Традиционные модели матричной факторизации используют неявную матрицу совпадения векторных реконструкций в качестве цели оптимизации и не могут использовать богатую семантическую информацию в гетерогенных информационных сетях. Многие исследователи предложили методы матричной факторизации, подходящие для моделирования гетерогенных информационных сетей, которые можно разделить на две категории: методы, основанные на регуляризации, и методы, основанные на факторизации нейронных матриц. По сравнению с методами, основанными на измерении подобия, представленными в разделе 1.3.1, методы, представленные в этом разделе, не полагаются на явную достижимость пути и не дадут сбоя, если соединение по пути разреженное или зашумленное.

1.3.2.1. Матричные факторизации, основанные на регуляризации

Традиционная модель матричного разложения восстанавливает только матрицу оценки пользователем элемента и не может использовать другие типы объектов и отношений в гетерогенной информационной сети. Многие улучшения привели к появлению разнородных условий регуляризации в задаче оптимизации чтобы восполнить этот недостаток.

Мы называем этот тип методов матричными факторизациями, основанными на регуляризации. В частности, методы, основанные на регуляризации, используют различные методы измерения сходства для вычисления матрицы сходства, а затем используют неявный вектор для восстановления матрицы сходства как отдельного элемента регуляризации или объединяют матрицу сходства и матрицу оценок в единую матрицу, подлежащую разложению.

Некоторые методы учитывают различные типы гетерогенных отношений при разработке условий регуляризации. Например, CMF [1.61] разрабатывает модель совместной факторизации матриц, разлагающую несколько матриц одновременно и совместно использующую параметры. Учитывая, что разные отношения должны оказывать разное воздействие, HeteroMF [1.28] разрабатывает контекстно-зависимую матричную модель факторизации. Термин регуляризации рассматривает общее встраиваемое представление каждой сущности и контекстно-зависимое встраиваемое представление каждого отношения.

Чтобы использовать высокоуровневую семантическую информацию метапутей, существуют некоторые методы, которые используют меры подобия, основанные на метапутях, для генерации матриц подобия в качестве элементов регуляризации при факторизации матрицы. Например, Hete-MF [1.75] сначала использует PathSim для вычисления сходства между элементами на основе мета-путей, а затем использует линейную регрессию для взвешивания показателей сходства различных мета-путей. Он использует модель линейной регрессии для различения различных предпочтений в семантике разных метапутей, что приводит к интегрированной матрице подобия $S = \sum_{l=1}^{L} \theta_l S^{(1)}$, где значение θ_l обозначает важность l-го мета-пути среди всех L метапутей. Его целевая функция состоит из двух частей. Первая часть направлена на изучение

представлений низкого ранга U для пользователей и V для элементов с учетом рейтинговой матрицы R, а вторая часть направлена на изучение сильных сторон пути $\boldsymbol{\theta} = [\theta_1,\theta_2,...,\theta_l]^T$, который генерирует составленные отношения подобия между элементами.

$$\begin{aligned}
&\min_{U,V,q} \left\| Y - \left(R - UV^T \right) \right\|_F^2 + \lambda_0 \left(\|U\|_F^2 + \|V\|_F^2 \right) + \\
&\quad + \frac{\lambda_1}{2} \sum_{i,j} \sum_{l=1}^{L} q_l S_{ij}^{(l)} \left\| V_i - V_j \right\|_2^2 + \lambda_2 \|q\|_2^2 \\
&U \geq 0, V \geq 0, q \geq 0 \\
&\sum_{l=1}^{L} q_l = 1
\end{aligned} \tag{1.6}$$

где λ_0 - параметр настройки, который управляет регуляризацией, чтобы избежать чрезмерной подгонки при изучении U и V. λ_1 управляет компромиссом между рейтингом пользователя и матрицами подобия. λ_2 - параметр регуляризации на θ.

Кроме того, Hete-CF [1.45] расширяет Hete-MF [1.75], используя PathSim [1.63] для измерения отношений между пользователем, товаром-товаром и пользователем-товаром, а затем использует унифицированную матричную модель факторизации для интеграции вышеупомянутых трех видов разнородной информации в задачу социальной рекомендации. HeteRec [1.76] сначала использует мета-путь для вычисления сходства товара с товаром, а затем выполняет произведение с матрицей оценок пользователей, чтобы сгенерировать матрицу распространения пользовательских предпочтений, и использует неотрицательную матрицу в матрице распространения, чтобы узнать потенциальные характеристики пользователей и товаров. В [1.77] рассматривается персонализация на основе HeteRec [1.76] и использует PathSim для вычисления сходства пользовательских элементов на основе мета-путей, которое называется оценкой распространения пользовательских предпочтений. Amp-MF [1.43] предлагает усовершенствованный метод измерения сходства мета-путей,

который фиксирует богатую семантику сходства между объектами за счет рассмотрения структуры ссылок и улучшенных атрибутов ссылок. SimMF [1.50] разрабатывает структуру двойной регуляризации для интеграции разнородной информации с использованием средних или индивидуальных элементов регуляризации сходства пользователей и предметов.

1.3.2.2. Факторизация нейронных матриц в рекомендательных задачах

Способность матричных факторизаторов к линейному моделированию не позволяет им эффективно выражать сложные предпочтения пользователей, поэтому результаты рекомендаций недостаточно хороши. С развитием глубокого обучения факторизация нейронных матриц [1.25] и ее варианты достигли хороших результатов в рекомендательных задачах благодаря мощной способности нейронных сетей соответствовать любой функции.

Однако традиционный метод факторизации нейронных матриц использует только одноразовое представление идентификаторов пользователя и элемента в качестве входных признаков и не может использовать богатую семантическую информацию гетерогенных информационных сетей. Чтобы объединить структурную информацию, исследователи используют метод измерения сходства, представленный в разделе 1.3.1, для генерации матрицы сходства, матрицы обмена или последовательности случайного блуждания, которые вводятся в модель факторизации нейронной матрицы в качестве структурных признаков, а затем эти признаки дополнительно объединяются.

Например, как показано на рис. 1.2, NeuACF [1.23] использует PathSim для вычисления матриц подобия на уровне аспектов для различных мета-путей, где каждая матрица представляет предпочтения

пользователя в определенном аспекте (таком как бренд и категория и т.д.), А затем использует многослойные персептроны (MLP) для получения представлений на уровне аспектов для пользователей и товаров. NeuACF использует механизм внимания для объединения этих представлений на уровне аспектов. Например, учитывая представления пользователя в аспекте бренда u^B_i , двухуровневая сеть используется для вычисления показателя внимания s^B_i с помощью уравнения (1.7),

$$s_i^B = W_2^T f\left(W_1^T u_i^B + b_1\right) + b_2 \qquad (1.7)$$

где W_* и b_* - весовые матрицы и отклонения соответственно, а f(x)=max(0,x) - функция активации.

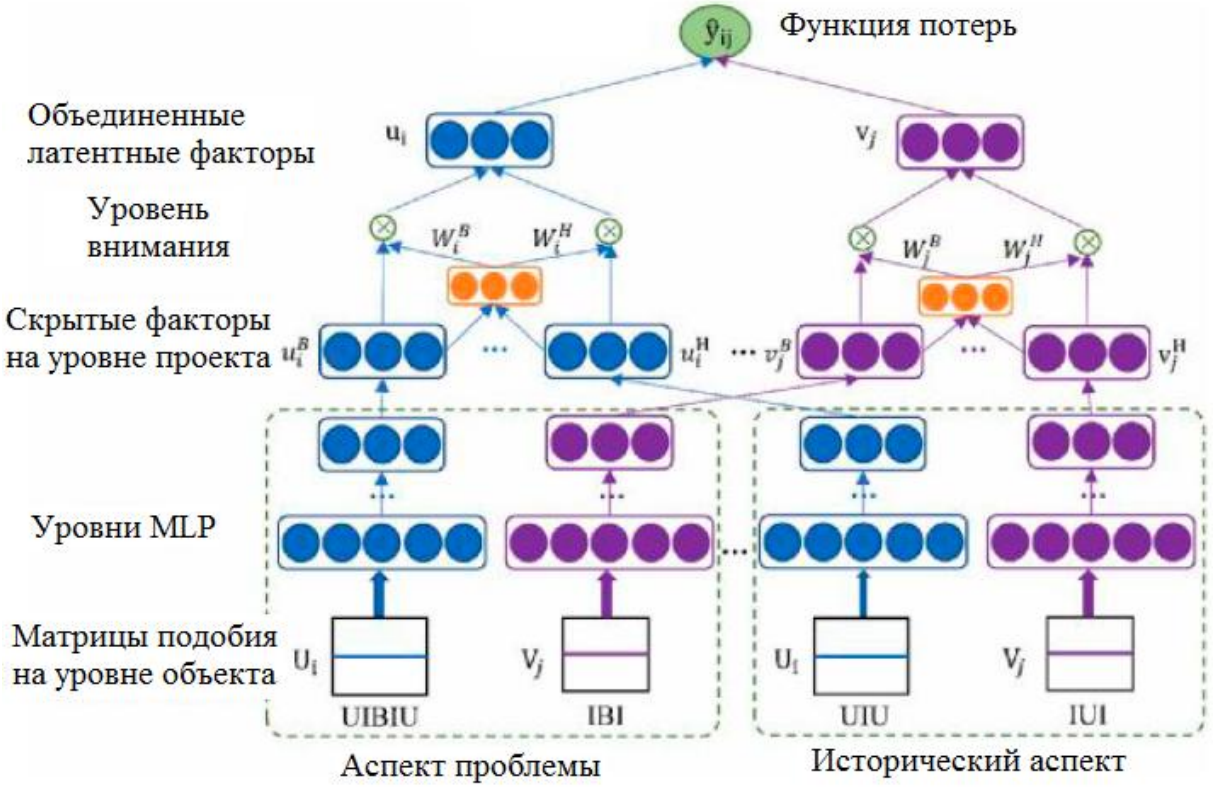


Рис. 1.2. Архитектура NeuACF

Окончательные значения внимания W^B_i для представлений на уровне аспектов получены путем нормализации приведенных выше показателей внимания s^B_i с помощью функции Softmax, которую можно

интерпретировать как вклад различных аспектов B в совокупный скрытый фактор U_i пользователя. Затем агрегированное представление u_i представляется как взвешенная сумма каждого аспекта u_i^B, и вероятность y_{ij} взаимодействия между пользователем U_i и элементом I_j вычисляется путем применения функции sigmod(·) к их агрегированным представлениям u_i и v_j. Поскольку основная истина y_{ij} находится во множестве {0, 1}, общая целевая функция записывается следующим образом:

$$Loss = - \sum_{i,j \in Y^+ \cup Y^-} \left(y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log (1 - \hat{y}_{ij}) \right) \quad (1.8)$$

где Y^+ и Y^- наборы положительных и отрицательных экземпляров соответственно. Набор отрицательных экземпляров Y^- выбирается из ненаблюдаемых данных для обучения.

Аналогично, HNAFM [1.9] использует различные мета-пути для вычисления матрицы обмена и использует многослойные персептроны для изучения представления пользователя и элемента, а затем использует иерархический механизм внимания для объединения нескольких представлений из мета-путей. RLDB [1.24] сначала использует случайное блуждание на основе мета-пути для получения последовательности соединений пользователей и элементов, затем использует сверточную нейронную сеть для изучения структурного представления и текстового представления и, наконец, использует многовидовую машину для эффективного объединения структурного представления и текстового представления.

1.3.2.3. Технология обучения графическому представлению

Методы, основанные на матричной факторизации, имеют много преимуществ. Поскольку результатом матричной факторизации являются

скрытые векторы пользователей и элементов, любая пара пользователь-элемент может предсказать результат, что улучшает способность к обобщению и снижает сложность пространства. Однако матричная факторизация также имеет недостатки, такие как сложность моделирования взаимодействий высокого порядка между пользователями и элементами. Чтобы решить эти проблемы, технология обучения графическому представлению внедрена в рекомендательную модель и постепенно получила широкое применение в рекомендательных системах, основанных на HIN.

1.3.3. Изучение графического представления

С развитием глубокого обучения рекомендательная модель, основанная на нейронной сети, достигла лучших результатов рекомендаций благодаря своей высокой способности к перекрестному использованию функций и гибкости проектирования архитектуры модели. Однако традиционные нейронные сети не могут напрямую моделировать графовые структуры. С развитием технологии обучения представлению графов исследователи попытались разработать рекомендательные модели, которые включают технологию обучения представлению графов, чтобы лучше усваивать богатую структурную и семантическую информацию, содержащуюся в графовых данных. В этом разделе эти методы в совокупности называются методами обучения графическому представлению и далее подразделяются на методы, основанные на двухэтапном обучении, и методы, основанные на сквозном обучении.

1.3.3.1. Двухэтапное обучение как предварительная подготовка и тонкая настройка

Обучение представлению графов без учителя направлено на

изучение низкоразмерных векторных представлений структуры графа, чтобы их можно было удобно хранить и применять для различных последующих задач. Методы, основанные на двухэтапном обучении, используют неконтролируемое встраивание графов для генерации структурных характеристик и использования их в качестве входных данных рекомендательной модели. Двухэтапное обучение также можно рассматривать как предварительную подготовку и тонкую настройку, где модель второго этапа играет роль точной настройки встраивания на основе контрольного сигнала рекомендательной задачи.

Фактически, неконтролируемая матричная факторизация является типичной моделью вложения в двудольный граф. Разница с рекомендательной моделью матричной факторизации заключается в том, что разложенная матрица неконтролируемой матричной факторизации основана на матрице подобия, полученной с помощью метапути, а не на рейтинговой матрице, содержащей надзорную информацию. Например, FMG [1.85] разрабатывает метод измерения сходства, основанный на мета-графах, и выполняет неконтролируемую матричную факторизацию матрицы сходства пользователя и элемента, полученной из мета-графов, для изучения векторов встраивания пользователей и элементов; затем вектор встраивания вводится в машину факторизации в качестве структурного признака, тем самым моделируя взаимосвязь высокого порядка между признаками. Подобно FMG, MoHINRec [1.86] предлагает метапуть с улучшенным мотивом, который дополнительно фиксирует отношения высокого порядка между узлами одного и того же типа, а затем вводит представление встраивания в машину факторизации для обучения.

Однако стоимость разложения крупномасштабных матриц обычно очень высока. Вдохновленные DeepWalk [1.4], node2vec [1.19] и другими методами, исследователи разработали множество неконтролируемых методов встраивания гетерогенных графов, основанных на случайных

блужданиях, таких как metapath2vec [1.14] и HIN2Vec [1.18], и применили их к генерации структурных признаков рекомендательных систем. Например, HERec [1.60] использует случайные блуждания на основе мета-пути для генерации последовательностей объектов и использует node2vec для изучения вложений объектов, а затем вычисляет сходство вложений для рекомендации. Чтобы быть более конкретным, HERec сначала генерирует траекторию обхода в соответствии со следующим распределением:

$$P(n_{t+1}=x \mid n_t=v, r)=\begin{cases}\dfrac{1}{\left|N^{A_{t+1}}(v)\right|}, (v,x)\in E \text{ и } j(x)=A_{t+1} \\ 0, \text{ в противном случае}\end{cases} \quad (1.9)$$

где n_T является T-м узлом маршрута, ρ представляет собой мета-путь, тип v из A_T, $\varphi(\cdot)$ является объектом типа отображения функции, а $N^{A_{t+1}}(v)$ множество первоочередной окрестности для узла v типа A_{T+1}.

Учитывая обработанную последовательность узлов, HERec определяет окрестность для узла u на основе совместного появления в окне фиксированной длины в последовательности, обозначается как N_u. Следуя node2vec, окончательный вклад узлов изучается путем решения следующей задачи оптимизации:

$$\max_f \sum_{u\in V} \log \Pr\left(N_u \mid f(u)\right) \quad (1.10)$$

где f: $V \rightarrow R^d$ - функция, отображающая каждый узел в d-мерное пространство вложения.

$\Pr\left(N_u \mid f(u)\right)$ оценивает вероятность соседей u с учетом вложения u. После оптимизации с помощью уравнения (10) мы можем получить набор вложений $\left\{e_v^{(l)}\right\}_{l=1}^{|P|}$, где P - набор мета-путей. Затем HERec использует общую функцию $g(\cdot)$, чтобы объединить изученные вложения узлов, чтобы получить $e_u^{(U)}$ и $e_i^{(I)}$ для пользователей и элементов соответственно.

Наконец, рейтинг $r_{u,i}$ пользователя u по позиции i рассчитывается следующим образом:

$$\hat{r}_{u,i} = x_u^T y_i + \alpha e_u^{(U)^T} g_i^{(I)} + \beta g_u^{U^T} e_i^I \tag{1.11}$$

где $x_u \in R^D$ и $y_i \in R^D$ обозначают скрытые факторы, соответствующие пользователю u и i позиции путем факторизации матрицы оценок пользовательских позиций, $\gamma_u^{(U)}$ и $\gamma_i^{(I)}$ являются скрытыми факторами, специфичными для пользователя и конкретной позиции, которые необходимо сопоставить с HG вложения $e_i^{(I)}$ и $e_u^{(U)}$ соответственно, а также α и β являются параметрами настройки для объединения трех членов.

Общая цель сформулирована следующим образом:

$$\begin{aligned} \Omega = & \sum_{(u,i,r_{u,i}) \in R} \left(r_{u,i} - \hat{r}_{u,i} \right)^2 + \\ & + \lambda \sum_u \left(\|x_u\|_2 + \|y_i\|_2 + \|g_u^{(U)}\|_2 + \|g_i^{(I)}\|_2 + \|\Theta^{(U)}\|_2 + \|\Theta^{(I)}\|_2 \right) \end{aligned} \tag{1.12}$$

где $\hat{r}_{u,i}$ - прогнозируемый рейтинг HERec с использованием уравнения (11), λ - параметр регуляризации, а $\Theta^{(U)}$ и $\Theta^{(I)}$ - параметры функции $g(\cdot)$ для пользователей и товаров соответственно.

Аналогично, IF-BPR [1.81] использует мета-пути для идентификации неявных друзей в системе социальных рекомендаций и разрабатывает метод предвзятого случайного блуждания для учета шума в социальных отношениях. HueRec [1.72] предполагает, что пользователи или элементы имеют общую семантику в разных мета-путях, а затем использует все мета-пути для изучения единого представления пользователей и элементов. GetHERec [1.87] преобразует гетерогенную сеть в несколько подсетей на основе метапутей и использует изученные вложения для упорядочивания факторизации матрицы.

Кроме того, некоторые методы основаны на реляционном моделировании и не требуют ручного указания мета-пути. Например, HRLHG [1.35] предлагает иерархический алгоритм случайного блуждания,

который реализует двухуровневое случайное блуждание под руководством двух разных наборов распределений. Локальное случайное блуждание основано на определении распределения реляционных преобразований и используется для моделирования локальной семантической информации, которая не имеет отношения к задаче. Глобальное случайное блуждание определяется на основе распределения полезности типов отношений и используется для моделирования сетевых шаблонов, специфичных для конкретной задачи. HIGE [1.70] фиксирует совпадение на уровне локального контекста, на уровне пользователя и на уровне метаданных на этапе внедрения, а затем моделирует глобальные и контекстуальные предпочтения пользователей для рекомендаций. NREP [1.80] совместно оптимизирует корреляцию содержимого узла, сохранение структуры и прогнозирование границ для создания встраивания объектов. WHIN-CSL [1.10] использует node2vec непосредственно для генерации встраивания узлов и рекомендует использовать линейную комбинацию мультимодальных сходств. Кроме того, GAN-HBNR [1.6] и ECHCDR [1.37] предлагают использовать методы на основе GAN для изучения встраивания узлов, CFKG [1.84] и ECFKG [1.1] используют модель на основе трансляции [1.2] для генерации встраивания узлов, а HRec [1.66] использует PTE [1.67] для генерации встраивания узлов.

1.3.3.2. Комплексное обучение на основе сквозного встраивания графа

Двухэтапные методы обучения не используют контролируемую информацию на этапе встраивания графа, поэтому сгенерированное представление встраивания графа трудно подходит для различных рекомендательных задач. Напротив, метод сквозного встраивания графа может использовать информацию о наблюдении при изучении встраивания

графа, чтобы улучшить степень соответствия встраивания графа и рекомендательных задач.

Графовая нейронная сеть - это репрезентативная нейронная сеть, которая обрабатывает графовые структуры и поддерживает сквозное обучение. Нейронные сети ранних графов подходят только для однородных графов, включая **G**raph **C**onvolution **N**etwork (**GCN**), **G**raph **A**ttention **N**etwork [1.68] и др. Чтобы применить графовые нейронные сети к гетерогенным информационным сетям, исследователи предложили множество гетерогенных графовых нейронных сетей и применили их к рекомендательным системам. В этом разделе будут представлены комплексные методы обучения, представленные гетерогенными графовыми нейронными сетями.

1.3.3.2.1. Нейронные сети на графах с учетом отношений

Для гетерогенных информационных сетей с богатыми типами отношений исследователи обычно используют нейронные сети на графах с учетом отношений (например, RGCN [1.54]) и применяют их к рекомендательным системам.

Репрезентативной работой является IntentGC [1.86], которая преобразует сложные гетерогенные графы в двудольные графы, состоящие только из пользователей и элементов, на основе отношений второго порядка, и выполняет векторную свертку вместо побитовой, чтобы избежать ненужного взаимодействия объектов и повысить надежность модели. IntentGC сначала проектирует следующую векторную функцию свертки, учитывающую только один тип вспомогательных отношений.

$$\begin{aligned} g_u^{k-1}(i) &= s\left(w_u^{k-1}(i,1)h_u^{k-1} + w_u^{k-1}(i,2)h_{N(u)}^{k-1}\right) \\ h_k^u &= s\left(\sum_{i=1}^{L} q_i^{k-1} g_u^{k-1}(i)\right) \end{aligned} \tag{1.13}$$

где h^k_u - вектор вложения пользователя u после k-го сверточного слоя,

$w_u^{k-1}(i,1)$ и $w_u^{k-1}(i,2)$ обозначают веса i-го локального фильтра для собственного узла и окрестности соответственно.

Чтобы охватить более разнородные взаимосвязи вспомогательной информации, IntentGC обобщается следующим образом:

$$g_u^{k-1}(i) = s\left(w_u^{k-1}(i,1)h_u^{k-1} + \sum_{r=1}^{R-2} w_u^{k-1}(i,r+1)h_{N^{(r)}(u)}^{k-1}\right) \quad (1.14)$$

где $h_{N^{(r)}(u)}^{k-1}$ - агрегированный вектор через *r*-й тип окрестности в соответствии с $S_U^{(r)}$.

Аналогично, веса $\left\{w_u^{k-1}(i,r+1)\right\}_{r=0}^{r=R-2}$ локального фильтра i являются общими на графе, где R−2 - количество типов отношений.

Наконец, IntentGC минимизирует следующую функцию тройных потерь, которая разработана как подход с максимальной маржой:

$$L\left(X_u, X_v, X_{neg}\right) = \max\left\{0, z_u, z_{neg} - z_u z_v + d\right\} \quad (1.15)$$

где δ - гиперпараметр прибылии, а внутреннее произведение используется для измерения показателя сходства между узлом пользователя и узлом товара.

Кроме того, несколько работ основаны на двухуровневом механизме агрегирования, который объединяет информацию внутри отношений и между различными отношениями соответственно. Например, DisenHAN [1.73] проецирует атрибуты узла в разные подпространства и использует двухуровневый механизм внимания для агрегирования внутри- и взаимосвязей для изучения представлений узлов. SIAN [1.44] использует двухуровневый механизм внимания для двухуровневой агрегации для изучения представлений узлов и использует соединитель социального влияния для моделирования степени влияния связи между друзьями и предметами. SHCF [1.12] использует двухуровневое внимание для изучения встраивания предметов и механизм самоконтроля с учетом положения для изучения встраивания пользователей. IPRec [1.13]

разрабатывает сеть внимания внутри и между пакетами для моделирования пакетов и двухуровневую сеть внимания для моделирования пользователей. THIGE [1.32] использует двухуровневое внимание для долгосрочного моделирования предпочтений пользователя и моделирования предметов, а также как закрытую повторяющуюся единицу, так и уровень внимания для краткосрочного моделирования предпочтений пользователя. DHGAT [1.48] использует двухуровневую сеть внимания для изучения надежного представления запроса и хранилища, затем объединяет графическое представление с дополнительным текстовым представлением и передает их в двухбашенную сеть для вычисления сходства. GAMMA [1.69] использует двухуровневый механизм внимания для агрегирования на уровне компонентов и представлений, а также дополнительный уровень внимания в многовидовой памяти с несколькими головками для внимательного извлечения компонентов, специфичных для пользователя, из каждого представления. TreeMS [1.49] использует агрегирование внутри представления на основе внимания и концентрирует каждое представление для получения пользовательского представления, а также использует сеть LSTM и гибридную экспертную сеть для моделирования древовидного иерархического представления элементов. HHFAN [1.7] использует внутривидовую агрегацию на основе выборки и межтиповую агрегацию на основе внимания. Comb-K [1.34] использует GCN для изучения пользовательских вложений, зависящих от отношений, объединяет несколько пользовательских вложений, зависящих от отношений, для получения конечных пользовательских вложений и использует гетерогенную сеть объединения графов для кластеризации пользователей и оценки групповых предпочтений.

Кроме того, существуют также некоторые другие методы вложения графов, основанные на отношениях. GHCF [1.11] явно изучает

представление отношений и использует произведение узла на уровне элемента и вектора отношений для агрегирования соседей, а затем усредняет агрегированные векторы на каждом уровне, тем самым явно объединяя информацию о прямом взаимодействии и взаимодействии высокого порядка. HGNR [1.41] строит гетерогенный граф путем добавления социальных ссылок и семантических связей, предсказанных на основе социальных сетей и текстовых обзоров, в матрицу взаимодействия пользователя и элемента и использует GCN [1.38] для одновременного изучения встраивания пользователя и элемента. BasConv [1.42] разрабатывает три типа агрегаторов, специально разработанных для трех типов узлов в гетерогенном графе пользовательская корзина-товар, включая слои самораспространения и интерактивные слои. User-as-Graph [1.74] использует стратегию объединения разнородных графов для обобщения информации о графе для конкретного типа из всего графика, чтобы изучить встраивание пользователя, и внимательную платформу с несколькими представлениями для изучения встраивания элементов.

1.3.3.2.2. Нейронная сеть с гетерогенным графом, основанная на моделировании мета-пути

Другой тип нейронной сети с гетерогенным графом основан на моделировании мета-пути (например, HAN [1.71], который может эффективно вводить предварительные знания и фиксировать высокоуровневую семантическую информацию в гетерогенных графах.

Типичной работой по применению такого рода гетерогенной графовой нейронной сети к системам рекомендаций является MEIRec [1.15], которая агрегирует гетерогенных соседей каждого порядка вдоль мета-пути, а затем агрегирует различные мета-пути для получения пользовательского встраивания и встраивания запросов. MEIRec сначала извлекает термины из запросов и заголовков элементов и создает словарь

терминов W и отображает все термины в d-мерные векторы с помощью функции отображения $f:M\rightarrow R^d$, где M представляет словарь терминов. Затем встраивания терминов агрегируются для получения встраиваний запросов или элементов в единое пространство встраивания следующим образом:

$$\begin{aligned} E_{q_2} &= g\left(e_{w_1}, e_{w_n}\right) \\ E_{i_2} &= g\left(e_{w_1}, e_{w_{n-1}}, e_{w_n}\right) \end{aligned} \tag{1.16}$$

где e_{wi} - вложение терма w_i, а $g(\cdot)$ означает среднюю функцию, применяемую к термам.

После этого MEIRec агрегирует соседей, управляемых метапутью, чтобы получить пользовательские вложения. Взяв мета-путь UIQ в качестве примера, MEIRec сначала получает встраивание элемента i_j путем агрегирования его соседей с одним переходом следующим образом:

$$I_j^{UIQ} = g\left(E_{q_1}, E_{q_2}, ...\right) \tag{1.17}$$

где $g(\cdot)$ - функция усреднения агрегирования, а запросы $\{q_1, q_2, \cdots\}$ являются соседями элемента i_j. Затем MEIRec агрегирует вложения соседей (элементов) первого шага, чтобы получить встраивание U^{UIQ}_i пользователя u_i:

$$U_i^{UIQ} = g\left(I_1^{UIQ}, I_2^{UIQ}, ...\right) \tag{1.18}$$

где элементы $\{i_1, i_2, \ldots\}$ являются соседями пользователя u_i, а функция агрегирования $g(\cdot)$ есть LSTM. Затем MEIRec получает объединенное встраивание пользователя и встраивание запроса путем агрегирования встраиваний на основе различных метапутей:

$$\begin{aligned} U_i &= g\left(U_i^{r_1}, U_i^{r_2}, ..., U_i^{r_k}\right) \\ Q_i &= g\left(Q_i^{r_1}, Q_i^{r_2}, ..., Q_i^{r_k}\right) \end{aligned} \tag{1.19}$$

где ρ - метапуть, начинающаяся от пользователя или запроса, а $g(\cdot)$ - функция агрегирования.

Наконец, MEIRec объединяет вложения пользовательских функций, запросов и статических функций S_{ij} , чтобы объединить их, и передает объединенные вложения в слои MLP, чтобы получить оценку прогнозирования $\hat{y}_{ij}$. Наконец, функция потерь MEIRec является точечной функцией потерь в уравнении (1.20).

$$J = \sum_{i,j \in Y^+ \cup Y^-} \left(y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log (1 - \hat{y}_{ij}) \right) \quad (1.20)$$

где y_{ij} − метка экземпляра (т.е. 1 или 0), а Y^+ и Y^- - это набор положительных и отрицательных экземпляров соответственно.

Кроме того, некоторые методы явно изучают встраиваемое представление метапути во время обучения модели, которые были предложены до появления графовых нейронных сетей. Например, MCRec [1.27] использует сверточные нейронные сети для изучения представлений встраивания мета-пути и предлагает общий механизм внимания для взаимного улучшения представлений контекста, пользователя и элемента на основе мета-пути. RKGE [1.64] использует несколько RNN для генерации вложений различных путей, связывающих одну и ту же пару объектов, и объединяет вложения путей с помощью операции объединения, чтобы представить полное предпочтение пользователя в отношении элемента.

1.3.3.3. Рекомендательная модель, основанная на обучении графическому представлению

Рекомендательная модель, основанная на обучении графическому представлению, позволяет эффективно усваивать богатую структуру и семантическую информацию на графике и может гибко комбинироваться с другими технологиями глубокого обучения. С одной стороны, неконтролируемое обучение представлению графов может использоваться

в качестве предварительной подготовки для генерации векторов признаков, затем объединяться с другими векторами признаков и вводиться в последующие модели; с другой стороны, контролируемое обучение представлению графов может использоваться в качестве уровня встраивания рекомендательной модели для достижения сквозного обучения. Однако существующая технология обучения графическому представлению по-прежнему сталкивается со многими проблемами, такими как масштабируемость и недостаточная интерпретируемость, которые сейчас находятся в центре внимания исследований.

1.4. Постановка задач работы

Метод описания гетерогенных информационных сетей позволяет эффективно использовать богатую структурную и семантическую информацию во вспомогательной информации и применять ее для моделирования рекомендательных систем. Это может не только эффективно решить проблему холодного запуска, вызванную разреженностью данных, но и повысить точность и интерпретируемость рекомендательных моделей. Систематически исследованы соответствующие работы и ресурсы рекомендательных систем, основанных на HIN.

Результат анализа потребовал формализации данных задач, а также алгоритмизации их решения с учетом особенностей, отраженных на рис. 1.3.

Целью работы является разработка моделей и алгоритмов управления процессами обработки гетерогенных данных в рамках информационных систем с многомерными атрибутами на основе мягкой максиминной оценки и активного обучения.

Задачи исследования. Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ проблем управления гетерогенными данными информационных систем с многомерными атрибутами на основе мягкой максиминной оценки и активного обучения.

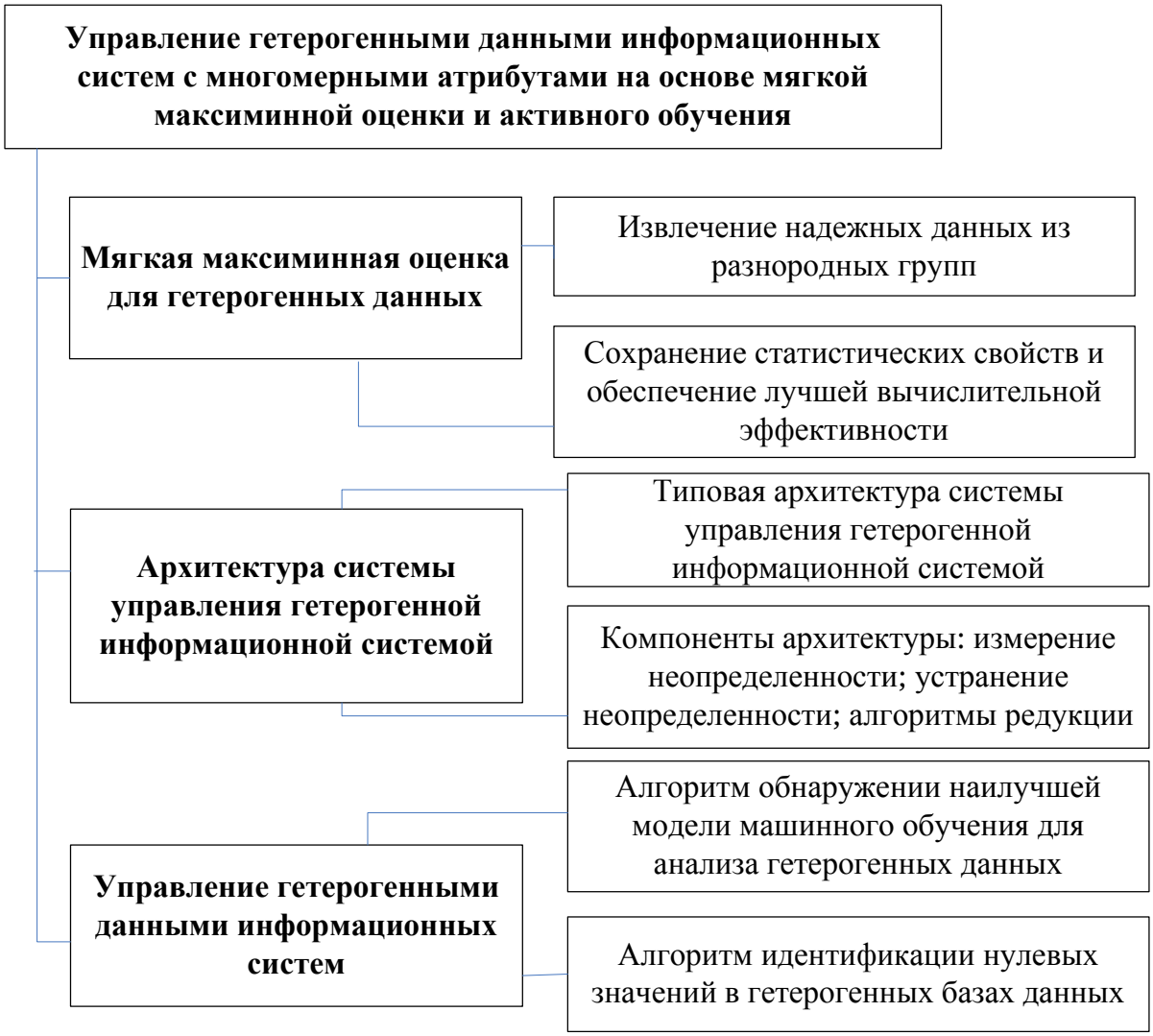


Рис. 1.3. Дизайн исследования

2. Разработать мягкую максиминную оценку для гетерогенных данных, содержащих уникальные вариационные компоненты, обеспечивающую сохранение статистических свойств и лучшую вычислительную эффективность.

3. Предложить архитектуру системы управления гетерогенной информационной системой, обеспечивающую эффективную редукцию многомерных анализируемых атрибутов.

4. Создать алгоритм обнаружения наилучшей модели машинного обучения для анализа гетерогенных данных, обеспечивающий сокращение объемов данных и повышение точности анализа совокупности данных.

5. Разработать алгоритм идентификации нулевых значений в гетерогенных базах данных, обеспечивающий более эффективную и точную оценку нулевых значений.

Источники к главе 1

1.1. Ai Q., Azizi V., Chen X., Zhang Y. 2018. Learning heterogeneous knowledge base embeddings for explainable recommendation// Algorithms 11 (9), 137.

1.2. Antoine B. et al. 2013. Translating embeddings for modeling multi-relational data// Adv. Neural Inf. Process. Syst. 26.

1.3. Bao J., Zheng Yu, Wilkie D., Mokbel M. 2015. Recommendations in location-based social networks: a survey// GeoInformatica 19 (3), 525–565.

1.4. Bryan P., Al-Rfou R., Skiena S. 2014. Deepwalk: online learning of social representations// Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 701–710.

1.5. Burke R., Vahedian F., Mobasher B. 2014. Hybrid recommendation in heterogeneous networks// Int. Conf. on User Modeling, Adaptation, and Personalization. Springer, pp. 49–60.

1.6. Cai X., Han J., Yang L. 2018. Generative adversarial network based heterogeneous bibliographic network representation for personalized citation recommendation// Thirty-second AAAI Conf. on Artificial Intelligence.

1.7. Cai D., Qian S., Quan F., Xu C. 2021. Heterogeneous hierarchical feature aggregation network for personalized micro-video recommendation// IEEE Trans. Multimed. 24, 805–818.

1.8. Chang J. et al. 2020. Bundle recommendation with graph convolutional networks// Proc. of the 43rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 1673–1676.

1.9. Chen L. et al. 2018. Heterogeneous neural attentive factorization machine for rating prediction// Proc. of the 27th ACM Int. Conf. on Information and Knowledge Management, pp. 833–842.

1.10. Chen J. et al. 2019. Citation recommendation based on weighted heterogeneous information network containing semantic linking// 2019 IEEE Int. Conf. on Multimedia and Expo (ICME). IEEE, pp. 31–36.

1.11. Chen C. et al. 2021. Graph heterogeneous multi-relational recommendation// Proc. of the AAAI Conf. on Artificial Intelligence, 35, pp. 3958–3966.

1.12. Chen Li et al. 2021. Sequence-aware heterogeneous graph neural collaborative filtering// Proc. of the 2021 SIAM Int. Conf. on Data Mining (SDM). SIAM, pp. 64–72.

1.13. Chen Li, et al. 2021. Package recommendation with intra- and inter-package attention networks// The 44th Int. ACM SIGIR Conf. on Research & Development in Information Retrieval.

1.14. Dong Y., Chawla N.V., Swami A. 2017. metapath2vec: scalable representation learning for heterogeneous networks// Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 135–144.

1.15. Fan S. et al. 2019. Metapath-guided heterogeneous graph neural

network for intent recommendation// Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 2478–2486.

1.16. Feng W., Wang J. 2012. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems// Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1276–1284.

1.17. Fouss F. et al. 2007. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation// IEEE Trans. Knowl. Data Eng. 19 (3), 355–369.

1.18. Fu T.-y., Lee W.-C., Lei Z. 2017. Hin2vec: explore meta-paths in heterogeneous information networks for representation learning// Proc. of the 2017 ACM on Conf. on Information and Knowledge Management, pp. 1797–1806.

1.19. Grover A., Leskovec J. 2016. node2vec: scalable feature learning for networks// Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 855–864.

1.20. Guerraoui R. et al. 2017. Heterogeneous recommendations: what you might like to read after watching interstellar// Proc. of the VLDB Endowment, 10, pp. 1070–1081, 10.

1.21. Guo C., Liu X. 2015. Automatic feature generation on heterogeneous graph for music recommendation// Proc. of the 38th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 807–810.

1.22. Guy I. et al. 2010. Social media recommendation based on people and tags// Proc. of the 33rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 194–201.

1.23. Han X. et al. 2018. Aspect-level deep collaborative filtering via heterogeneous information networks// IJCAI, pp. 3393–3399.

1.24. Han X. et al. 2018. Representation learning with depth and breadth for recommendation using multi-view data// Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Int. Conf. on Web and Big Data. Springer, pp. 181–188.

1.25. He X. et al. 2017. Neural collaborative filtering// Proc. of the 26th Int. Conf. on World Wide Web, pp. 173–182.

1.26. Hu J. et al. 2016. Recexp: a semantic recommender system with explanation based on heterogeneous information network// Proc. of the 10th ACM Conf. on Recommender Systems, pp. 401–402.

1.27. Hu B. et al. 2018. Leveraging meta-path based context for top-n recommendation with a neural co-attention model// Proc. of the 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 1531–1540.

1.28. Jamali M., Lakshmanan L. 2013. Heteromf: recommendation in heterogeneous information networks using context dependent factor models// Proc. of the 22nd Int. Conf. on World Wide Web, pp. 643–654.

1.29. Jannach D., Zanker M., Felfernig A., Gerhard F. 2010. Recommender Systems - an Introduction. - Cambridge University Press.

1.30. Jeh G., Widom J. 2002. Simrank: a measure of structural-context similarity// Proc. of the Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 538–543.

1.31. Jeh G., Widom J. 2003. Scaling personalized web search// Proc. of the 12th Int. Conf. on World Wide Web, pp. 271–279.

1.32. Ji Y. et al. 2020. Temporal Heterogeneous Interaction Graph Embedding for Next-Item Recommendation// Lecture Notes in Computer Science, vol 12459. Springer, Cham. https://doi.org/10.1007/978-3-030-67664-3_19.

1.33. Ji H. et al. 2021. Large-scale comb-k recommendation// Proc. of Web Conf. 2021, pp. 2512–2523.

1.34. Ji H. et al. 2021. Who you would like to share with? a study of share recommendation in social e-commerce// Proc. of the AAAI Conf. on Artificial Intelligence, vol. 35, pp. 232–239.

1.35. Jiang Z. et al. 2018b. Cross-language citation recommendation via hierarchical representation learning on heterogeneous graph// The 41st Int. ACM SIGIR Conf. on Research & Development in Information Retrieval, pp. 635–644.

1.36. Jin J. et al. 2020. An efficient neighborhood-based interaction model for recommendation on heterogeneous graph// Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 75–84.

1.37. Jin L. et al. 2020. Heterogeneous graph embedding for cross-domain recommendation through adversarial learning// Int. Conf. on Database Systems for Advanced Applications. Springer, pp. 507–522.

1.38. Kipf T.N., Welling M. 2016. Semi-Supervised Classification with Graph Convolutional Networks. - arXiv preprint arXiv:1609.02907.

1.39. Lee S. et al. 2013. Pathrank: ranking nodes on a heterogeneous graph for flexible hybrid recommender systems// Expert Syst. Appl. 40 (2), 684–697.

1.40. Liu X. et al. 2014. Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation// Proc. of the 23rd Acm Int. Conf. on Conf. on Information and Knowledge Management, pp. 121–130.

1.41. Liu S. et al. 2020. A heterogeneous graph neural model for cold-start recommendation// Proc. of the 43rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 2029–2032.

1.42. Liu Z. et al. 2020. Basconv: aggregating heterogeneous interactions for basket recommendation with graph convolutional neural network// Proc. of the 2020 SIAM Int. Conf. on Data Mining. SIAM, pp. 64–72.

1.43. Lu C.-T. et al. 2016. Item Recommendation for Emerging Online Businesses// IJCAI, pp. 3797–3803.

1.44. Lu Y. et al. 2020. Social Influence Attentive Neural Network for Friend-Enhanced Recommendation// Lecture Notes in Computer Science, vol 12460. Springer, Cham. https://doi.org/10.1007/978-3-030-67667-4_1.

1.45. Luo C. et al. 2014. Hete-cf: social-based collaborative filtering recommendation using heterogeneous relations// 2014 IEEE Int. Conf. on Data Mining. IEEE, pp. 917–922.

1.46. Nandanwar S. et al. 2018. Fusing diversity in recommendations in heterogeneous information networks// Proc. of the Eleventh ACM Int. Conf. on Web Search and Data Mining, pp. 414–422.

1.47. Ni L., Cohen W.W. 2010. Relational retrieval using a combination of path-constrained random walks// Mach. Learn. 81 (1), 53–67.

1.48. Niu X. et al. 2020. A dual heterogeneous graph attention network to improve long-tail performance for shop search in e-commerce// Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 3405–3415.

1.49. Niu X. et al. 2021. A Dual Heterogeneous Graph Attention Network to Improve Long-Tail Performance for Shop Search in E-Commerce// KDD '20: Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining. – P. 3405 – 3415.

1.50. Resnick P., Varian H.R., 1997. Recommender systems// Commun. ACM 40 (3), 56–58.

1.51. Nguyen P.T.-A. et al. 2016. A general recommendation model for heterogeneous networks. IEEE Trans. Knowl. Data Eng. 28 (12), 3140–3153.

1.52. Ren X. et al. 2014. Cluscite: effective citation recommendation by information network-based clustering// Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 821–830.

1.53. Schafer J.B., Konstan J.A., Riedl J. 2001. E-commerce recommendation applications// Data Min. Knowl. Discov. 5 (1), 115–153.

1.54. Schlichtkrull M. et al. 2018. Modeling relational data with graph convolutional networks// European Semantic Web Conf.. Springer, pp. 593–607.

1.55. Shi C. et al. 2012. Heterecom: a semantic-based recommendation system in heterogeneous networks// Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1552–1555.

1.56. Shi C. et al. 2014. Hetesim: a general framework for relevance measure in heterogeneous networks// IEEE Trans. Knowl. Data Eng. 26 (10), 2479–2492.

1.57. Shi C. et al. 2015. Semantic path based personalized recommendation on weighted heterogeneous information networks// Proc. of the 24th ACM Int. on Conf. on Information and Knowledge Management, pp. 453–462.

1.58. Shi C. et al. 2016. A survey of heterogeneous information network analysis// IEEE Trans. Knowl. Data Eng. 29 (1), 17–37.

1.59. Shi C. et al. 2016. Integrating heterogeneous information via

flexible regularization framework for recommendation// Knowl. Inf. Syst. 49 (3), 835–859.

1.60. Shi C. et al. 2018. Heterogeneous information network embedding for recommendation// IEEE Trans. Knowl. Data Eng. 31 (2), 357–370.

1.61. Singh A.P., Gordon J.G. 2008. Relational learning via collective matrix factorization// Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 650–658.

1.62. Sun Y., Yu Y., Han J. 2009. Ranking-based clustering of heterogeneous information networks with star network schema// Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 797–806.

1.63. Sun Y. et al. 2011. Pathsim: meta path-based top-k similarity search in heterogeneous information networks// Proc. of the VLDB Endowment, 4, pp. 992–1003, 11.

1.64. Sun Z. et al. 2018. Recurrent knowledge graph embedding for effective recommendation// Proc. of the 12th ACM Conf. on Recommender Systems, pp. 297–305.

1.65. Sun Z. et al. 2019. Research commentary on recommendations with side information: a survey and research directions// Electron. Commer. Res. Appl. 37, 100879.

1.66. Su Y. et al. 2019. Hrec: heterogeneous graph embedding-based personalized point-of-interest recommendation// Int. Conf. on Neural Information Processing. Springer, pp. 37–49.

1.67. Tang J. et al. 2015. Pte: predictive text embedding through large-scale heterogeneous text networks// Proc. of the 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1165–1174.

1.68. Velickovic P. et al. 2017. Graph Attention Networks// arXiv preprint arXiv:1710.10903.

1.69. Vijaikumar M., Shevade S,, Narasimha M.M. 2020. Gamma: a graph and multi-view memory attention mechanism for top-n heterogeneous recommendation. Adv. Knowl. Discov// Data Mining 28, 12084.

1.70. Wang, Do., Xu G., Deng S. 2017. Music recommendation via heterogeneous information graph embedding// 2017 Int. Joint Conf. on Neural Networks (IJCNN). IEEE, pp. 596–603.

1.71. Wang X. et al. 2019. Heterogeneous graph attention network// The World Wide Web Conf., pp. 2022–2032.

1.72. Wang Z. et al. 2019. Unified embedding model over heterogeneous information network for personalized recommendation// IJCAI, pp. 3813–3819.

1.73. Wang Y. et al. 2020. Disenhan: disentangled heterogeneous graph attention network for recommendation// Proc. of the 29th ACM Int. Conf. on Information & Knowledge Management, pp. 1605–1614.

1.74. Wu C. et al. 2021. User-as-graph: User Modeling with Heterogeneous Graph Pooling for News Recommendation// Proc. of the

Thirtieth Int. Joint Conf. on Artificial Intelligence. 10.24963/ijcai.2021/224

1.75. Xiao Y. et al. 2013. Collaborative filtering with entity similarity regularization in heterogeneous information networks// IJCAI HINA 27.

1.76. Xiao Y. et al. 2013. Recommendation in heterogeneous information networks with implicit user feedback// Proc. of the 7th ACM Conf. on Recommender Systems, pp. 347–350.

1.77. Xiao Y. et al. 2014. Personalized entity recommendation: a heterogeneous information network approach// Proc. of the 7th ACM Int. Conf. on Web Search and Data Mining, pp. 283–292.

1.78. Xu F. et al. 2019. Relation-aware graph convolutional networks for agent-initiated social e-commerce recommendation// Proc. of the 28th ACM Int. Conf. on Information and Knowledge Management, pp. 529–538.

1.79. Yang X. et al. 2014. A survey of collaborative filtering based social recommender systems// Comput. Commun. 41 (1–10).

1.80. Yang L., Zhang Z., Cai X., Guo L. 2019. Citation recommendation as edge prediction in heterogeneous bibliographic network: a network representation approach// IEEE Access 7, 23232–23239.

1.81. Yu J. et al. 2018. Adaptive implicit friends identification over heterogeneous network for social recommendation// Proc. of the 27th ACM Int. Conf. on Information and Knowledge Management, pp. 357–366.

1.82. Yuan F. et al. 2016. Semantic proximity search on graphs with metagraph-based learning// 2016 IEEE 32nd Int. Conf. on Data Engineering (ICDE). IEEE, pp. 277–288.

1.83. Zhang J/ et al. 2008. Recommendation over a heterogeneous social network// 2008 the Ninth Int. Conf. on Web-Age Information Management. IEEE, pp. 309–316.

1.84. Zhang Y. et al. 2018. Learning over Knowledge-Base Embeddings for Recommendation// arXiv preprint arXiv:1803.06540.

1.85. Zhao H. et al. 2017. Meta-graph based recommendation fusion over heterogeneous information networks// Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 635–644.

1.86. Zhao H. et al. 2019. Motif enhanced recommendation over heterogeneous information network// Proc. of the 28th ACM Int. Conf. on Information and Knowledge Management, pp. 2189–2192.

1.87. Zhao Z. et al. For recommendation// Proc. of the 25th ACM SIGKDD Int. 2020. Hetnerec: heterogeneous network embedding based recommendation. Knowl. Conf. on Knowledge Discovery & Data Mining, pp. 2347–2357.

2. Пути интерполяции мягкой максиминной оценки для гетерогенных данных

Извлечение общего надежного сигнала из данных, разделенных на разнородные группы, является сложной задачей, когда каждая группа - в дополнение к сигналу - содержит большие уникальные вариационные компоненты. Ранее максиминная оценка была предложена в качестве надежного метода при наличии неоднородного шума. Предлагается модификация мягкой максиминной оценки максимального значения в качестве привлекательной с вычислительной точки зрения альтернативы, направленной на достижение баланса между объединенной оценкой и (жесткой) оценкой максимального значения [2.11]. Метод мягкой максиминной оценки предоставляет диапазон оценок, управляемых параметром $\xi>0$, который интерполирует объединенную оценку наименьших квадратов и максиминную оценку. Устанавливая соответствующие теоретические свойства, утверждается, что метод мягкой максиминной оценки является статистически обоснованным и привлекательным с точки зрения вычислений. Демонстрируется на реальных и смоделированных данных, что мягкая максиминная оценка может предложить улучшения как по сравнению с объединенными OLS, так и по сравнению с жестким максимином с точки зрения производительности прогнозирования и вычислительной сложности. Эффективная по времени и памяти реализация предусмотрена в пакете R SMME, доступном на CRAN.

2.1. Проблема извлечения общего сигнала из разнородных данных

Рассматривается проблема извлечения общего сигнала из разнородных данных. Поскольку гетерогенность преобладает в

крупномасштабных системах, цель - эффективный в вычислительном отношении оценщик (решение) с хорошими статистическими свойствами при различной степени неоднородности данных.

Чтобы конкретизировать концепцию неоднородности, рассмотрим линейную модель смеси с одномерными переменными отклика Y_1,..., Y_n, сгенерированными как

$$Y_i = X_i^{\top} B_i + \varepsilon_i,\ i = 1,\ldots,\ n. \qquad (2.1)$$

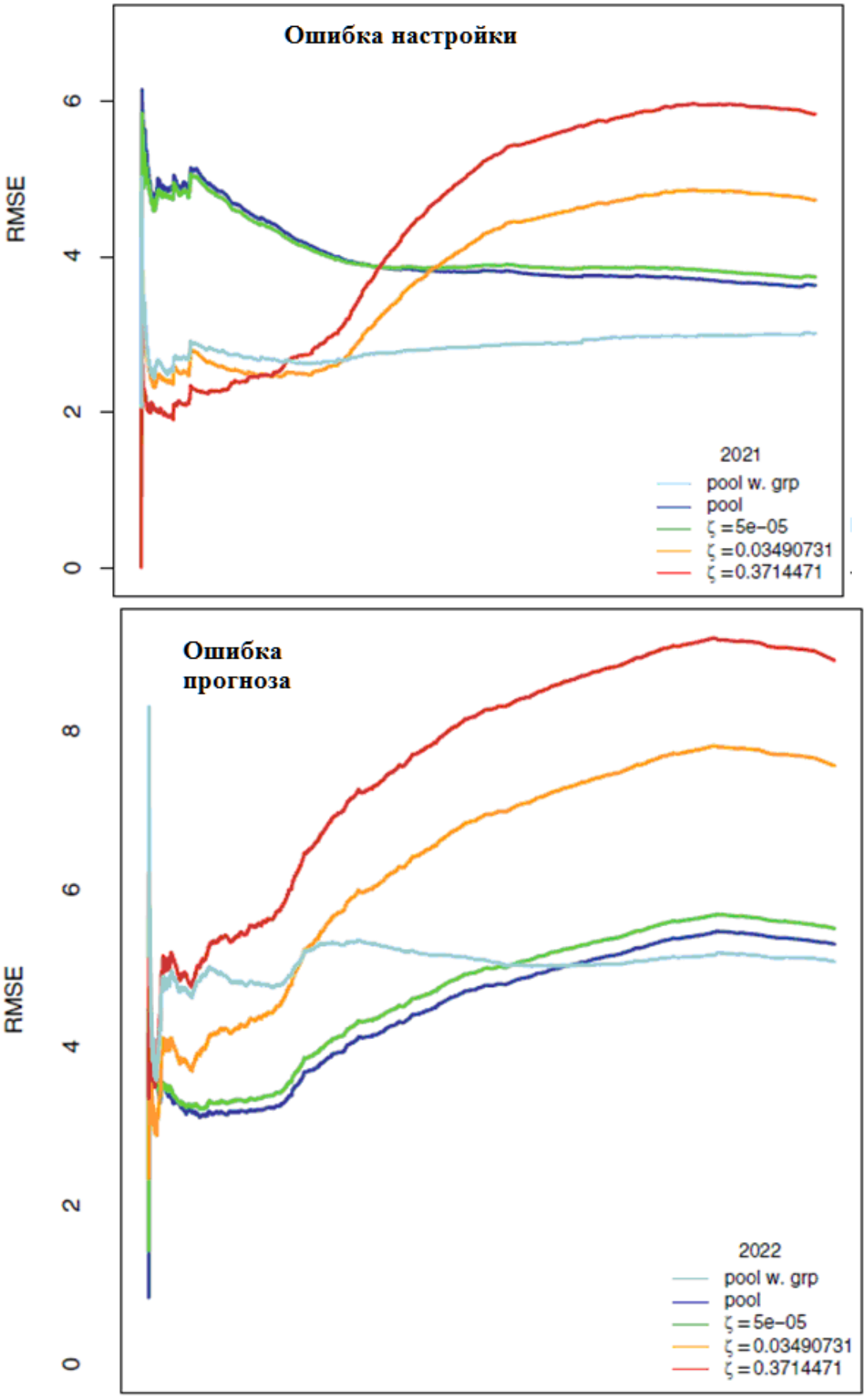
Здесь ε_1, …, ε_n и векторы признаков X_1, …, X_n являются p-мерными случайными величинами, а B_1, …, B_n - одномерными шумовыми переменными. Векторы признаков наблюдаются и предполагаются независимыми одинаково распределёнными случайными величинами, а шумовые переменные равны и предполагаются независимыми одинаково распределёнными случайными величинами. Ненаблюдаемые переменные B_1,…, B_n распределены идентично распределению F_B, но не обязательно независимы [2.1].

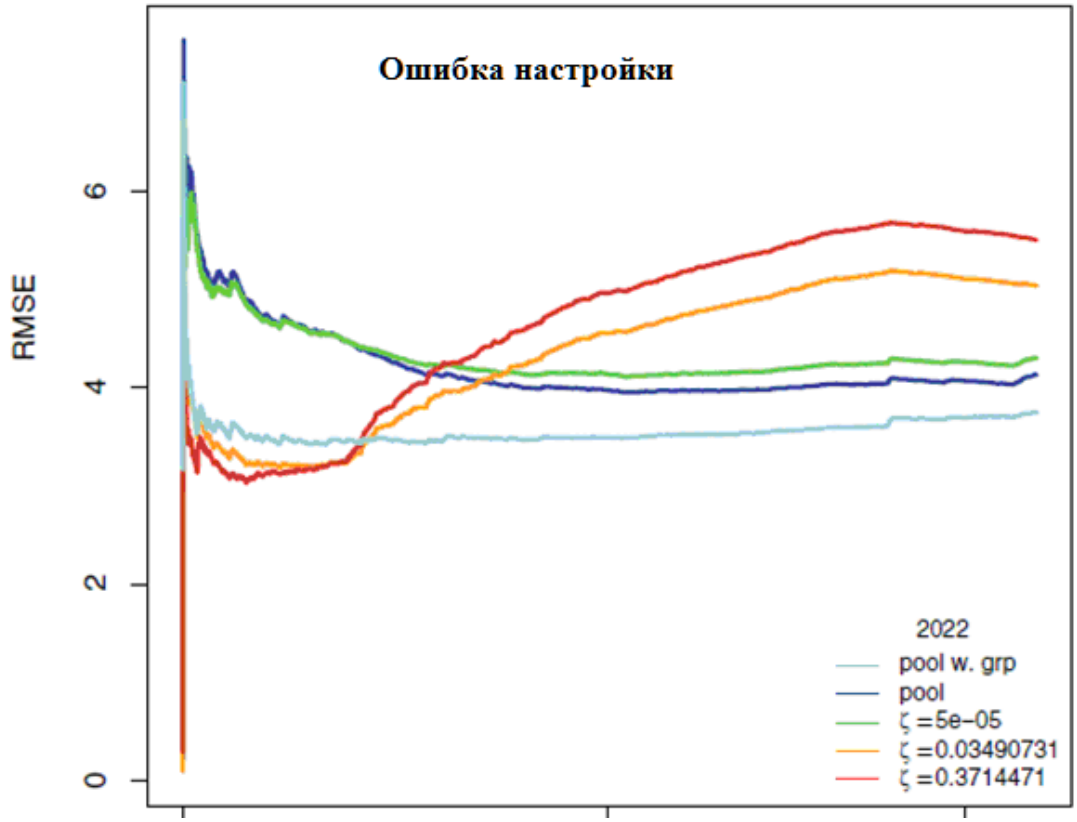
Неоднородность в модели, приведенной в (2.1), обусловлена изменением смещения, регулируемым F_B. Поскольку B_i может быть зависимым, модель (2.1) может фиксировать неоднородность, вызванную структурой группы, то есть, когда данные поступают с естественной группировкой, а B_i постоянен внутри групп, но различается между группами. Даже если данные не сгруппированы или если структура группы неизвестна, полезно изучить настройку с известной структурой группы. В приведенном в [2.1] примере совместного использования объектов групповая структура вводится для представления временной неоднородности и [2.1] демонстрирует, как создавать групповые структуры как часть вывода, когда группировка не задана.

Поэтому сосредоточимся на настройке с группами G и с постоянными B_i внутри групп. Цель состоит в том, чтобы изучить один

$\beta \in R^p$, который можно разумно рассматривать как общий сигнал B_i. Объединение данных по группам и вычисление обычной оценки наименьших квадратов (OLS) может быть ненадежным, в зависимости от F_B, и в [2.1] введена максиминная оценка в качестве надежной альтернативы OLS для разнородных данных из модели (2.1). Общим сигналом, оцениваемым с помощью оценки максимина, является количество населения, называемое максиминным эффектом.

Хотя максиминная оценка надежна, она также может быть консервативной, и предлагается мягкая максиминная оценка, чтобы обеспечить хороший баланс между максиминной оценкой и объединенной оценкой OLS. Баланс контролируется параметром настройки $\xi>0$, при этом $\xi \to \infty$ соответствует максиминной оценке. На рис. 2.1 показан результат применения мягкой максиминной оценки для трех значений ξ, а также объединенной оценки OLS к реальному набору данных. Это иллюстрирует, как прогностическая эффективность двух экстремальных оценок интерполируется с помощью мягкой максиминной оценки, определяемой количественно как кумулятивная среднеквадратичная ошибка (RMSE) с течением времени.


Ошибка настройки
RMSE
0
2
4
6
2021
pool w. grp
pool
ζ =5e−05
ζ =0.03490731
ζ =0.3714471
Ошибка прогноза
RMSE
0
2
4
6
8
2022
pool w. grp
pool
ζ =5e−05
ζ =0.03490731
ζ =0.3714471


Ошибка настройки
RMSE
0
2
4
6
2022
pool w. grp
pool
ζ =5e−05
ζ =0.03490731
ζ =0.3714471

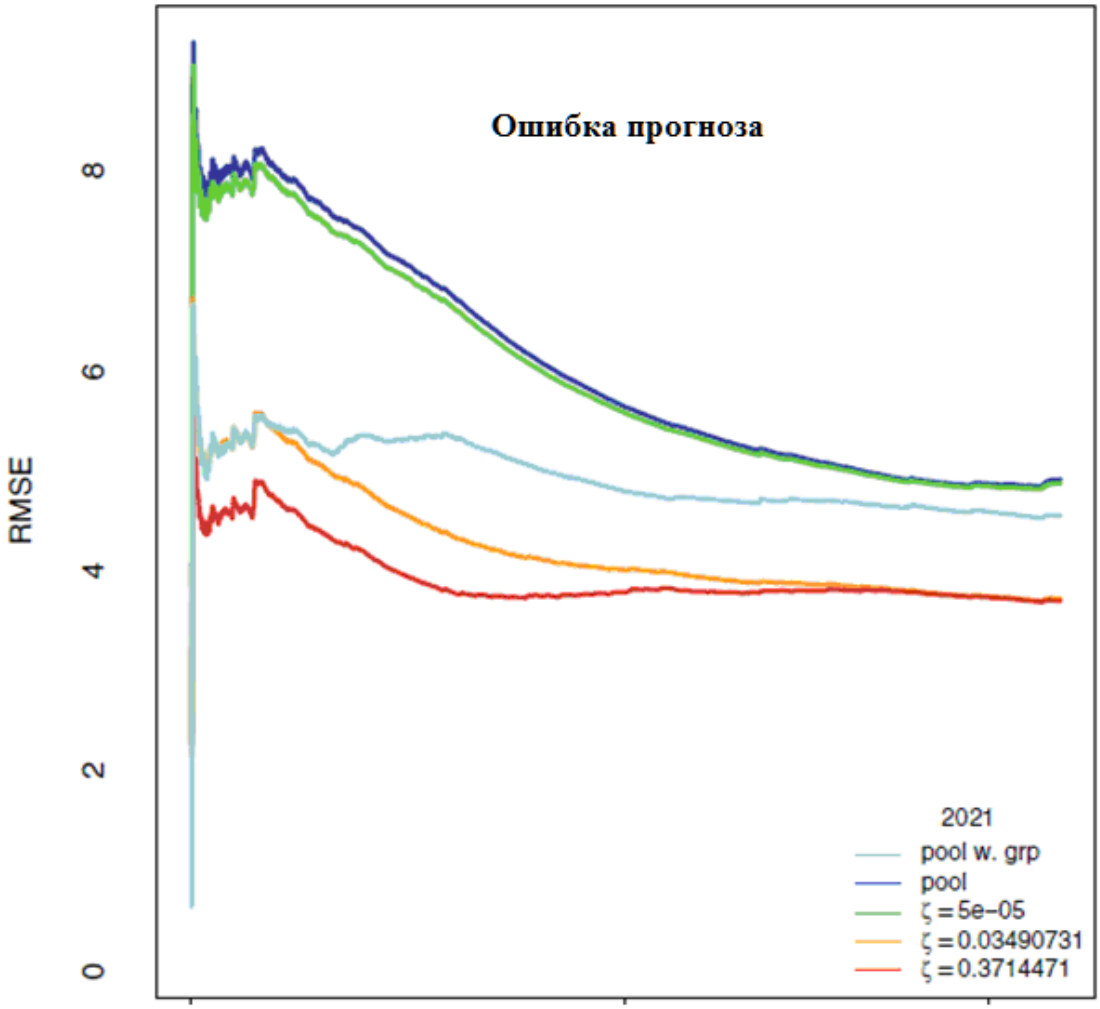

Ошибка прогноза
RMSE
0
2
4
6
8
2021
pool w. grp
pool
ζ =5e−05
ζ =0.03490731
ζ =0.3714471

Рис. 2.1. Пара вверху: RMSE, вычисленный для обучающих данных 2021 года и данных проверки 2022 года для мягкой максиминной оценки

Конкретное применение, проиллюстрированное на рис. 2.1, будет подробно описано в дальнейших фазах исследования. Данные основаны на почасовом количестве объектов [2.2] за 2 года (2021 и 2022). В этом случае лучше всего работают объединенные OLS или мягкие максиминные оценки с низким коэффициентом полезного действия. Однако обучение на данных за 2022 год приводит к тому, что объединенный OLS-оценщик перегружен, а мягкая максиминная оценка с большим значением имеет лучшую прогностическую производительность.

Исследование организовано следующим образом. Модель и структура оценки будут описаны на этапе 2, и будут обсуждаться статистические свойства мягкой максиминной оценки. Этот этап включает теоретические результаты, подтверждающие, что мягкая максиминная оценка интерполирует максиминную и объединенную оценку OLS. На этапе 3 будет предложен алгоритм для вычисления общей мягкой максиминной оценки. Алгоритм решает недифференцируемую задачу выпуклой оптимизации, но в отличие от максиминной оценки [2.1], решаемая задача является разделимой в смысле [2.3]. Это значительно облегчает построение эффективных алгоритмов с гарантиями сходимости. Этап 3 также будет включать теоретические оценки констант Липшица, которые могут быть использованы для выбора эффективного размера шага в алгоритме, и обсуждение того, как результаты могут быть эффективно применены к сглаживанию тензорной матрицы. На этапе 4 будет представлено приложение для обмена данными Objekt и результатами имитационного исследования по сглаживанию тензорной матрицы. Имитационное исследование было вдохновлено приложением к данным о нейронной активности.

2.2. Мягкий максиминный оценщик

Здесь мы представляем методологию в постановке с заданной структурой группы и воздействуем на константу B_i внутри каждой группы. Это, в свою очередь, подразумевает конечную поддержку F_B, и эту модель, возможно, лучше понимать как тип линейной смешанной модели, поскольку группировка доступна, следовательно, больше не является частью вывода. Однако, в отличие от традиционной смешанной модели, мы избегаем явного моделирования фиксированных и случайных эффектов, поскольку цель не состоит в том, чтобы делать выводы о них. Вместо этого мы стремимся получить только оценку возможных общих эффектов, присутствующих в данных.

Пусть в (2.1) $G \in N$ групп. Определим $I_1,..., I_G$ множества $\{1,..., n\}$ так, что $|Ig|=n_g$, а $n=\sum_g n_g$, где $\{" g, i \hat{I} Ig : B_i = Bg\}$ в g. Тогда

$supp(F_B) = \{b_1,..., b_G\} \subset Rp$, $b_g := B_g(\omega)$.

Пусть $\mathbf{Y}_g=(Y_{g,1},..., Y_{g,ng})^\top$ - вектор отклика $n_g \times 1$, $\mathbf{X}_g=(X_{g,1},..., X_{g,ng})^\top$ - след $n_g \times p$, $e_g=(e_{g,1},..., e_{g,ng})^\top$ - вектор ошибки $n_g \times 1$. Тогда

$$\mathbf{Y}_g=\mathbf{X}_g b_g+e_g, \quad g \in \{1,\ldots, G\}. \tag{2.2}$$

Общий сигнал в этой структуре представлен как $b \in R_p$ таким образом, что X_g является хорошей и надежной аппроксимацией $X_g b_g$ во всех группах G.

Чтобы оценить качество аппроксимации, мы используем критерий оптимальности [2.12]. Там объясненная дисперсия в группе g при использовании некоторого $b \in R_p$ в (2.2) определяется как

$$V_{bg}(b) := 2b^\top \Sigma b_g - b^\top \Sigma b. \tag{2.3}$$

Таким образом, оптимальное приближение есть

$b^* \in R^p$: $b^* := \arg\max_b \min_g V_{bg}(b)$.

Т.к. b_g не определен, то чтобы сделать этот критерий работоспособным, обозначим через $\hat{\Sigma}_g = X_g^T X_g / n_g$ эмпирическую матрицу Грама в группе g. Заменив Σ на Σ_g в (2.3), мы получим эмпирически объясненную дисперсию в группе g:

$V_g(b) := (2b^T \Sigma b_g - b^T \Sigma b)/n_g$.

$$\hat{V}_g(b) := \frac{1}{n_g}\left(2b^T X_g^T Y_g - b^T X_g^T X_g(b)\right) \qquad (2.4)$$

Как показано в работе [2.12], использование этой оценки может привести к более надежным оценкам для разнородных данных по сравнению с оценкой, которая не учитывает группировку, то есть объединенной оценкой. Предположение заключается в том, что максимальная оценка извлекает только те объекты, которые активны с одинаковым знаком в разных группах, при этом для объектов, специфичных для группы, устанавливается нулевое значение. Это делает ее более грубой оценкой по сравнению с оценкой, полученной с использованием методологии полной смешанной модели; однако в принципе она также более надежна и потенциально более привлекательна с точки зрения вычислений. В крупномасштабных системах обработки данных, где обычно встречается неоднородность данных, вычислительный аспект оценки имеет решающее значение.

Мы устраняем это вычислительное препятствие, заменяя функцию максимума следующей гладкой функцией. Для $G \in N$ и $z \neq 0$ рассмотрим масштабированную экспоненциальную функцию lse логарифмической суммы

$$\mathrm{lse}_z(x) = \frac{\log\left(\sum_j e^{z x_j}\right)}{z}, \quad x \in R^G \qquad (2.5)$$

Очевидно, что lse дифференцируема, и, как мы покажем далее, она обладает дополнительными свойствами, которые делают ее хорошо подходящим для целей оптимизации. Во-первых, основные свойства, изложенные далее, легко проверяются и подчеркивают, почему (2.5) является разумным выбором в качестве аппроксимации функции максимума.

Предположение 2.1. Пусть $G \in N$ и $x \in R^G$.

(2.1) При $z>0$

$$\max(x_1,...,x_G) \le lse_z(x) \le \frac{\log(G)}{z} + \max(x_1,...,x_G) \qquad (2.6)$$

и в частности $lse(x) \searrow \max_g\{x_g\}$ при $z \to \infty$.

(2.2) При $z \to 0$

$$lse_z(x) = \frac{1}{G}\sum_{j=1}^{G} x_j + \frac{\log(G)}{z} + o(1)$$

Определим мягкую функцию максимальных потерь как

$$s_z(b) = lse_z\left(-\hat{V}(b)\right), b \in R^p, z > 0$$

где $\hat{V}(b) = \left(\hat{V}_1(b),...,\hat{V}_G(b)\right)^T$. Для $k>0$ и $z>0$, оценка мягкого максимума теперь может быть определена как

$$\hat{b}^k_{smm} = \arg\min_b lse_z\left(-\hat{V}(b)\right) \quad \text{такое, что } \|b\|_1 \le k \qquad (2.7)$$

Используя предположение 2.1, можно количественно оценить влияние параметра на производительность оценки мягкого максимума (2.7). Следующий результат дает оценку максимальной отрицательной объясненной дисперсии оценки мягкого максимина.

Утверждение 2.1. Пусть $D = \max_g \left\|\hat{S}_g - S\right\|_\infty$ и $d = \max_g \left\|X_g^T e_g\right\|_\infty$. Тогда

$$\max_g\left\{-V_{b_g}\left(\hat{b}^k_{smm}\right)\right\} \le \max_g\left\{-V_{b_g}\left(b^*\right)\right\} + 6Dk^2 + 4kd + \frac{\log(G)}{z}$$

где $z>0$, $k>0$, $k \geq \max_g \|b_g\|_1$, b^* - максимальный эффект. В частности

$$\left\| \hat{b}^k_{smm} - b^* \right\|_S \leq 6Dk^2 + 4kd + \frac{\log(G)}{z}$$

Утверждение 2.1 доказывается путем объединения Предположения 2.1 и результатов [2.12]. В частности, потеря производительности, возникающая при использовании оценки мягкого максимума, ограничена той же величиной, что и у максиминной оценки, плюс логарифм смещения аппроксимации мягкого максимума $\log(G)/z$ из Предположения 2.1. Таким образом, при управлении параметром z мягкая оценка максимина обладает теоретическими свойствами, аналогичными свойствам (жесткой) оценки максимина. В частности, для D=0 (например, для фиксированного проекта) и фиксированного числа групп, если $n_g \to \infty$ для всех g, оценка мягкого максимина сохраняет только смещение аппроксимации.

Утверждение 2.1 устанавливает связь между производительностью мягкого максимина и эффектом максимина и показывает, что для $z \uparrow \infty$ мы действительно получаем максиминную оценку производительности. Однако это также подчеркивает, что при $z \downarrow 0$ производительность мягкой оценки максимина может сколь угодно сильно отклоняться от производительности оценки максимина. Действительно, согласно Предположению 2.1, для малых $z>0$,

$$s_z(b) \approx -\frac{1}{G}\sum_{i=1}^{G} \hat{V}_j(b) + \frac{\log(G)}{z} \approx \frac{1}{n}\sum_{j=1}^{G}\sum_{i=1}^{n_j} \frac{n}{Gn_j}\left(\left(X_j b\right)_i - Y_{j,i}\right)^2$$

и (2.7) фактически становится задачей взвешенных наименьших квадратов со штрафом (PWLS) по всем n наблюдениям. Таким образом, решение (2.7) для малого $z=0$ приблизительно дает объединенную оценку WLS с весами, усиливающими наблюдения из групп меньшего размера, чем в среднем. При одинаковом количестве наблюдений в каждой группе программа оценки мягкого максимума, в свою очередь, интерполирует объединенную

оценку PLS и максимальную оценку.

В этом смысле z отражает неоднородность данных. При небольшой неоднородности может хорошо работать низкий или даже нулевой показатель, соответствующий тому, что группировка не является релевантной.

Пример. Параметры генерации: G=20, n_g=400, 2D-пространство параметров. Для фиксированных эффектов $\{b_1,..., b_{20}\} \subset R^2$ мы производим выборку X_g и g, M=10 раз для каждого g, в результате чего получаем 10 различных наборов данных. Для каждого из этих десяти небольших наборов данных мы можем вычислить мягкую максиминную оценку (т.е. k=∞ в (7)) для последовательности значений z.

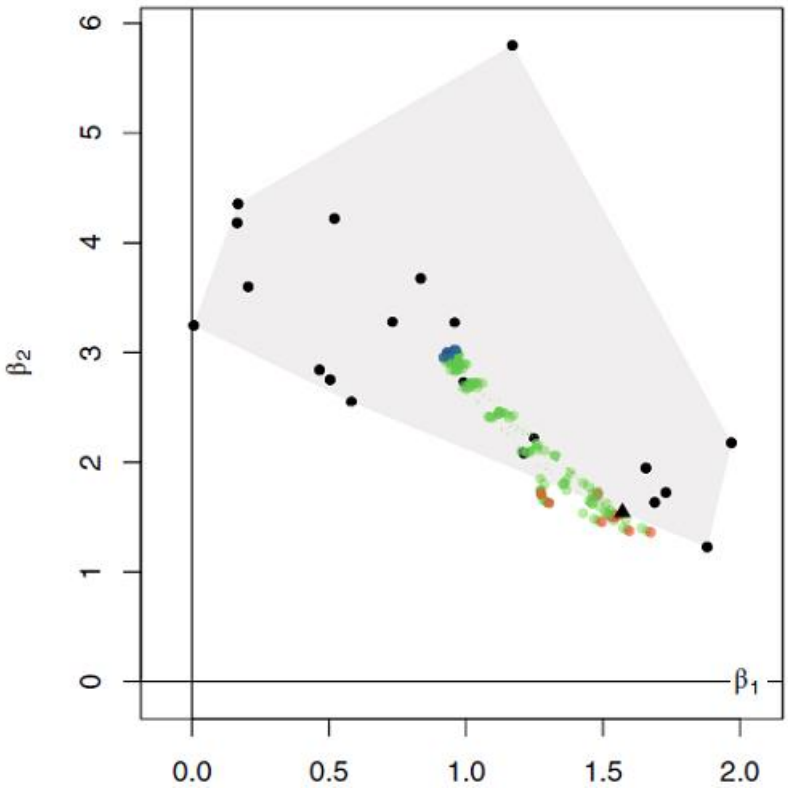
На рис. 2.2 показаны пути интерполяции мягкой максиминной оценки, соединяющие оценки LS населения и оценки maximin. Обратите внимание, что все оценки типа maximin сгруппированы вокруг теоретического максиминного эффекта, обозначенного символом ▲ на краю выпуклой оболочки $\{b_1,..., b_{20}\}$, в то время как объединенные оценки находятся далеко внутри выпуклой оболочки.

Важно, что совсем другой метод регрессии, якорная регрессия, был предложен в [2.15] для обработки неоднородности данных в ситуациях, когда распределение ответов может смещаться, что приводит к разнице между распределением обучающих данных и распределением тестовых данных. Если эта неоднородность может быть закодирована или сгенерирована известной опорной переменной, их метод может привести к улучшенной и, в частности, более стабильной производительности прогнозирования. В частности, контролируя параметр привязки g>0, этот метод интерполирует три различных метода регрессии, где g=1 дает регрессию OLS, а g=∞ - регрессию инструментальной переменной.

В некотором смысле точка привязки играет роль группирующей

структуры I_1,..., I_G, используемой в определении оценки мягкого максимума. Однако в общей настройке с неизвестными группами, учитывая определенные структурные допущения, можно построить наборы индексов I_1,..., I_G, как в примере выше, или путем случайной выборки. Теоретические гарантии в этом случае даны в [2.12] для максимальной оценки и аналогично Утверждению 2.1 должны распространяться на оценку мягкого максимума.

В дополнение к математической сложности, это добавляет процедуре вывода существенный уровень вычислительной сложности, поскольку число групп G в таком случае является гиперпараметром, который необходимо вывести, например, путем перекрестной проверки (CV). Таким образом, этот уровень кластеризации только усиливает важность эффективно вычисляемой базовой оценки.



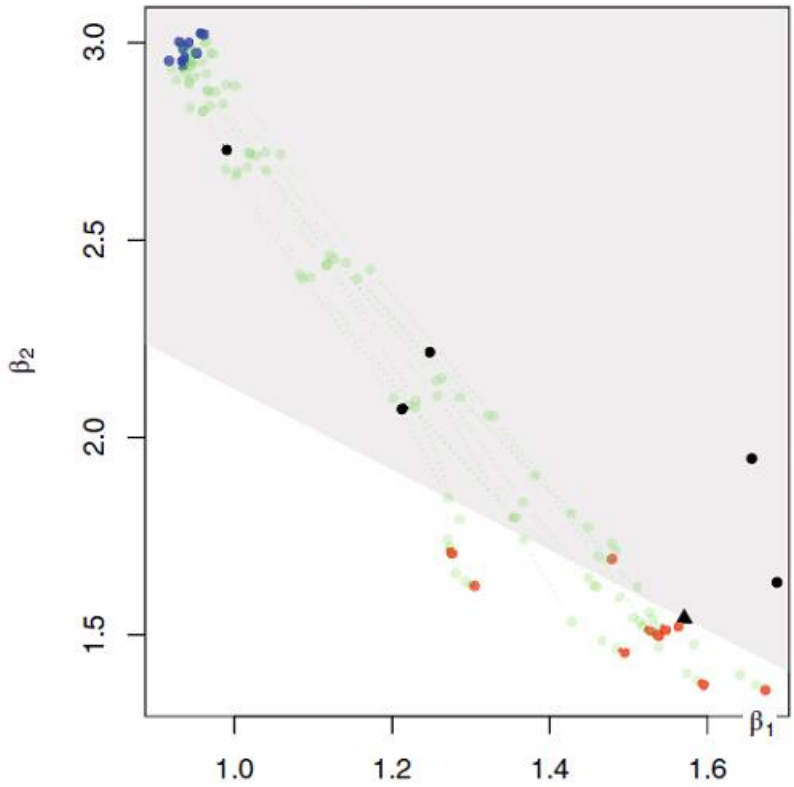


Рис. 2.2. Вверху: Выпуклая оболочка (серая заштрихованная область) supp (F_B) (черные точки). Теоретический максимальный эффект $b^*=b_*$, максимальные оценки (красные точки), мягкие максимальные оценки (зеленые точки) для различных z и популяционные оценки OLS (синие точки). Внизу: Крупный план

2.3. Вычислительные свойства

Здесь мы рассмотрим общую проблему оценки, в которой эмпирически объясненная дисперсия $\hat{V}_g$ из (2.4) [2.3] заменяется на $h_g : R^p \to R$. Пусть $\psi: R^n \to R$ − выпуклая функция,

$$D_y(x,y) = y(x) - \tilde{N}y(x)^T x,\ x, y \hat{I}\ R^n.$$

Пусть $h_g(b) = D_\psi(\eta_g(b), Y_g)$, $b \in R^p$, где $\eta_g(b) = \mathbf{X}_g$ – линейный предиктор в группе g.

Поскольку функция выпукла, D_ψ, подобно дивергенции Брегмана, выпукло в своем первом аргументе и, в частности, h_g выпукло. Однако, в отличие от дивергенции Брегмана, D_ψ не является неотрицательным.

Общая мягкая функция максимальных потерь l_ζ: $R^p \to R$ теперь задается формулой

$$l_x(b) = lse_x g h(b) = \frac{\log\left(\sum_{j=1}^{G} e^{x h_j(b)}\right)}{x}, x > 0$$

и наша цель состоит в том, чтобы решить общую задачу мягкого максимина, сформулированную следующим образом

$$\min_{b \in R^p} \left(l_x(b) + \lambda J(b) \right), \lambda \geq 0 \qquad (2.8)$$

Здесь J - правильная выпуклая штрафная функция, а λ - штрафной параметр.

Выбор ψ в качестве квадратичной нормы приводит к тому, что $-\hat{V}_g$ есть отрицательная эмпирически объясненная дисперсия как групповая дивергенция, то есть $h = -\hat{V}$. Если $J = \| \cdot \|_1$, то и потери, и штраф в этом случае выпуклы, и (2.8) эквивалентно (ограниченной) задаче мягкого максимума (2.7) [2.3] сильной лагранжевой двойственностью. Следовательно, в данном случае решение (2.8) является в точности мягкой оценкой максимума.

Заметим, что при другом выборе ψ мы получили бы совершенно новую оценку, потенциально со свойствами, сильно отличающимися от свойств оценки мягкого максимина. Немедленное обобщение типа Махаланобиса возникло бы, если бы мы позволили ψ быть заданным как взвешенная квадратичная норма.

Решение (2.8) в крупномасштабных условиях требует эффективного алгоритма оптимизации для недифференцируемых задач. В отличие от задачи жесткого максимина, (2.8) является, в дополнение к выпуклой и недифференцируемой, (частично) дифференцируемой и разделимой задачей [2.4]. Это означает, что ряд эффективных алгоритмов решит задачу

(2.8), например, алгоритмы разделения операторов первого порядка, такие как ADMM, или алгоритм второго порядка, такой как координатный спуск. Здесь рассмотрим модифицированные версии алгоритма проксимального градиента.

2.4. Алгоритм решения

Алгоритм проксимального градиента принципиально работает путем итеративного применения проксимального оператора

$$\text{prox}_{\Delta J}(b) = \arg\min_{g \in R^p} \left\{ \frac{1}{2\Delta} \|g - b\|_2^2 + J(g) \right\}, \Delta > 0$$

к шагам на основе градиента. Для функции потерь с непрерывным градиентом Липшица с постоянной L такой алгоритм гарантированно сходится к решению до тех пор, пока $\Delta \in (0,2/L)$, что делает привлекательным получение наименьшей возможной постоянной Липшица L.

С известным L и фиксированным $\Delta \in (0,2/L)$ проксимальный градиентный алгоритм содержит следующие шаги:

1. оценить градиент потерь;

2. оценить проксимальный оператор $\text{prox}_{\Delta J}$;

3. оценить целевые функции.

Вообще говоря не гарантируется, что шаги 1–3 приведут к решению (2.8) для любого фиксированного Δ. Однако мы проверяем, что следующий алгоритм немонотонного проксимального градиента (NPG) [2.6, 2.7] будет сходиться к решению (2.8) при некоторых условиях регулярности (алгоритм 2.1).

В частности, показано, что, хотя l_ζ в целом не имеет непрерывного градиента Липшица, сходимость алгоритма NPG по-прежнему гарантируется при общих условиях для групповых функций $h_1,\ldots, h_G$.

Кроме того, в частном случае, когда h_g - $\hat{V}_g$ и все группы имеют одинаковую конструкцию, мы показываем, что l_ζ имеет глобальный непрерывный градиент Липшица, и мы выводим постоянную Липшица.

Первый результат гласит, что l_ζ наследует сильную выпуклость от любой отдельной групповой функции расхождения h_g, учитывая, что все $h_1,..., h_G$ выпуклы и дважды непрерывно дифференцируемы.

Алгоритм 1. NPG-минимизация $F = f + \lambda J$

Требуется: β^0, $L_{max} \geq L_{min} > 0$, $\tau > 1$, $c > 0$, $M \in N$.

1: **for** $k = 0$ to $K \in N$ **do**

2: выбрать $L_k \in [L_{min}, L_{max}]$

3: расчет $\beta = prox_{\lambda J / L_k}\left(\beta^{(k)} - \frac{1}{L_k}\nabla f(\beta^{(k)})\right)$

4: **if** $F(\beta) \leq \max_{[k-M]_+ \leq i \leq k} F(\beta^i) - \frac{c}{2}\left\|\beta - \beta^{(k)}\right\|^2$ **then**

5: $\beta^{(k+1)} = \beta$

6: **else**

7: $L_k = \tau L_k$ и **go to** 3

8: **end if**

9: **end for**

Предложение 2.2. Для $g \in \{1,\ldots,G\}$ предположим, что hg дважды непрерывно дифференцируема и пусть $w_{g,z}(b) = e^{zh_g(b) - zl_z(b)}$, $b \in R^p$. Тогда $(w_{j,\zeta}(b))_j$ - выпуклые веса и

$$\tilde{N}l_z(b) = \sum_{j=1}^{G} w_{j,z}(b)\tilde{N}h_j(b) \qquad (2.10)$$

$$\nabla^2 l_z(b) = \sum_{i=1}^{G}\sum_{j=i+1}^{G} w_{i,z}(b)w_{j,z}(b)\left(\nabla h_i(b) - \nabla h_j(b)\right)\cdot$$
$$\cdot\left(\nabla h_i(b) - \nabla h_j(b)\right)^T + \sum_{j=1}^{G} w_{j,z}(b)\nabla^2 h_j(b) \qquad (2.11)$$

Кроме того, если $h_1,\ldots, h_G$, причем по крайней мере один h_g сильно выпуклый, то l_ζ и e^{zl_z} сильно выпуклы.

Предложение 2.2 применимо к мягким максимальным потерям с $h_g = -\hat{V}_g$. В этом случае $\nabla^2 h_g = 2X_g^T X_g / n_g$ и h_g сильно выпукло тогда и только тогда, когда $\mathbf{X}_g$ имеет ранг p. Рассмотрим мягкие максимальные потери при G=2, $p=n_1=n_2=2$ и

$$X_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 0 \\ \sqrt{2} & 0 \end{pmatrix} \qquad (2.12)$$

Примем также $y_1=y_2=0$. Тогда $b_1=b_2=r\in R$, отсюда $h_1(b)=h_2(b)= r^2$ и следовательно $w_{1,\zeta}=w_{2,\zeta}=1/2$ для любого ζ, поскольку выражение

$$\left(\nabla h_1(b) - \nabla h_2(b)\right)\cdot\left(\nabla h_1(b) - \nabla h_2(b)\right)^T = \begin{pmatrix} b_1^2 & -b_1b_2 \\ -b_1b_2 & b_2^2 \end{pmatrix} = \begin{pmatrix} r^2 & -r^2 \\ -r^2 & r^2 \end{pmatrix}$$

является неограниченным. В свою очередь, градиент не является липшицевым.

Следствие 2.1. Пусть

$$\forall\, g = \{1,\ldots,G\}: \begin{cases} X_g = X - \text{ матрица } m \times p \\ Y_g - \text{ вектор } m \times 1 \end{cases}$$

и групповая дивергенция задана формулой

$h_g(b) = (2b^T\mathbf{X}^T\mathbf{Y}_g - b^T\mathbf{X}^T\mathbf{X}\, b)/m.$

Тогда ∇l_ζ имеет липшицеву постоянную L, ограниченную величиной

$$\frac{4}{m^2}\max_{i,j}\left\|X^T(Y_i - Y_j)\right\|^2 + \frac{2\left\|X^TX\right\|}{m} \le \frac{4\left\|X^TX\right\|}{m^2}\left(\max_{i,j}\left\|(Y_i - Y_j)\right\|^2 + \frac{m}{2}\right) \qquad (2.13)$$

где |||·||| есть норма матрицы, порожденная обычной двумерной нормой ||·||.

Согласно следствию 2.1, если у нас одинаковые конструкции в разных группах, мы можем получить оценку мягкого максимума, применив алгоритм быстрого проксимального градиента из [2.5] к задаче оптимизации (2.8). Кроме того, в этой настройке следствие дает явное выражение для константы Липшица, которое даст эффективный размер шага Δ для алгоритма решения.

Наконец, в общей постановке следующее предложение показывает, что алгоритм 2.1, который не полагается на глобальное свойство Липшица [2.7], решает задачу (2.8) с учетом допущений в предложении 2.2. Доказательство предложения приведено в приложении.

Предложение 2.3. Предположим, что $h_1,\ldots, h_G$ удовлетворяют предположениям из предложения 2.2. Пусть $(b^{(k)})_k$ - последовательность итераций, полученная путем применения алгоритма NPG к (2.8). Тогда $b^{(k)} \to b^*$, где b^* - критическая точка $l_\zeta + l J$.

2.5. Численные эксперименты

2.5.1. Данные о запросах на обработку больших массивов данных в телекоммуникационной компании

Чтобы продемонстрировать свойства метода оценки мягкого максимума, мы приводим два примера данных.

Данные, использованные для получения результатов на рис. 2.1, описаны в [2.7]. Набор данных содержит данные за 2 года (2021 и 2022) (переменная cnt, см. рис. 2.3) из схемы обработки больших массивов данных в телекоммуникационной компании, а также вспомогательные данные, предположительно относящиеся к использованию велосипедов, такие как погода. Моделируем почасовое количество запросов на

обработку больших массивов данных, показанного на рисунке 3. Модельное уравнение для наблюдения i может быть записано в виде:

$$\sqrt{cnt_i} = \sum_{j=1}^{23} a_j j_j(hr_i) + \sum_{j=1}^{5} b_j j_j(wd_i) + \sum_{j=1}^{3} g_j l_j(ws_i) + e_i \qquad (2.14)$$

где использованы кубические базисные сплайновые функции.

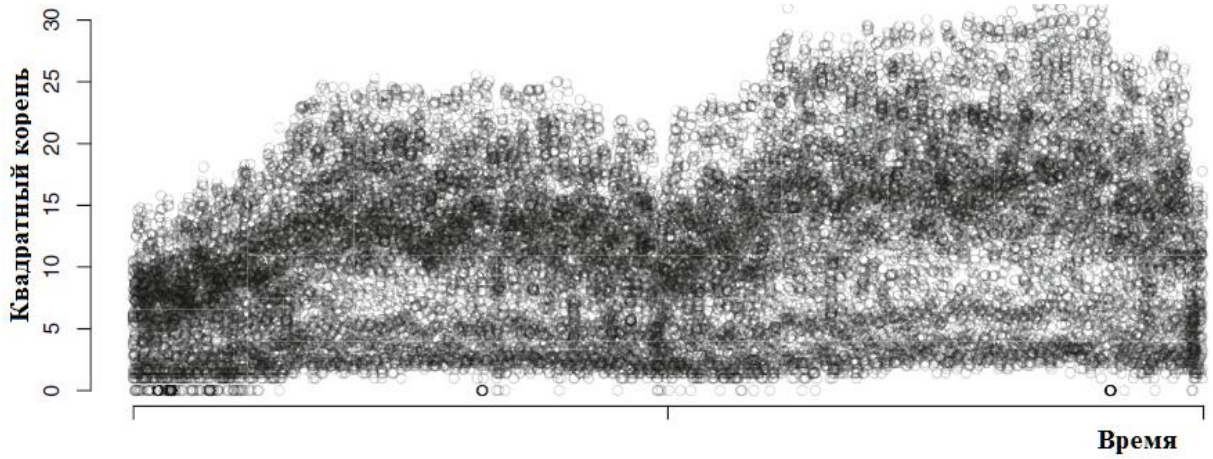
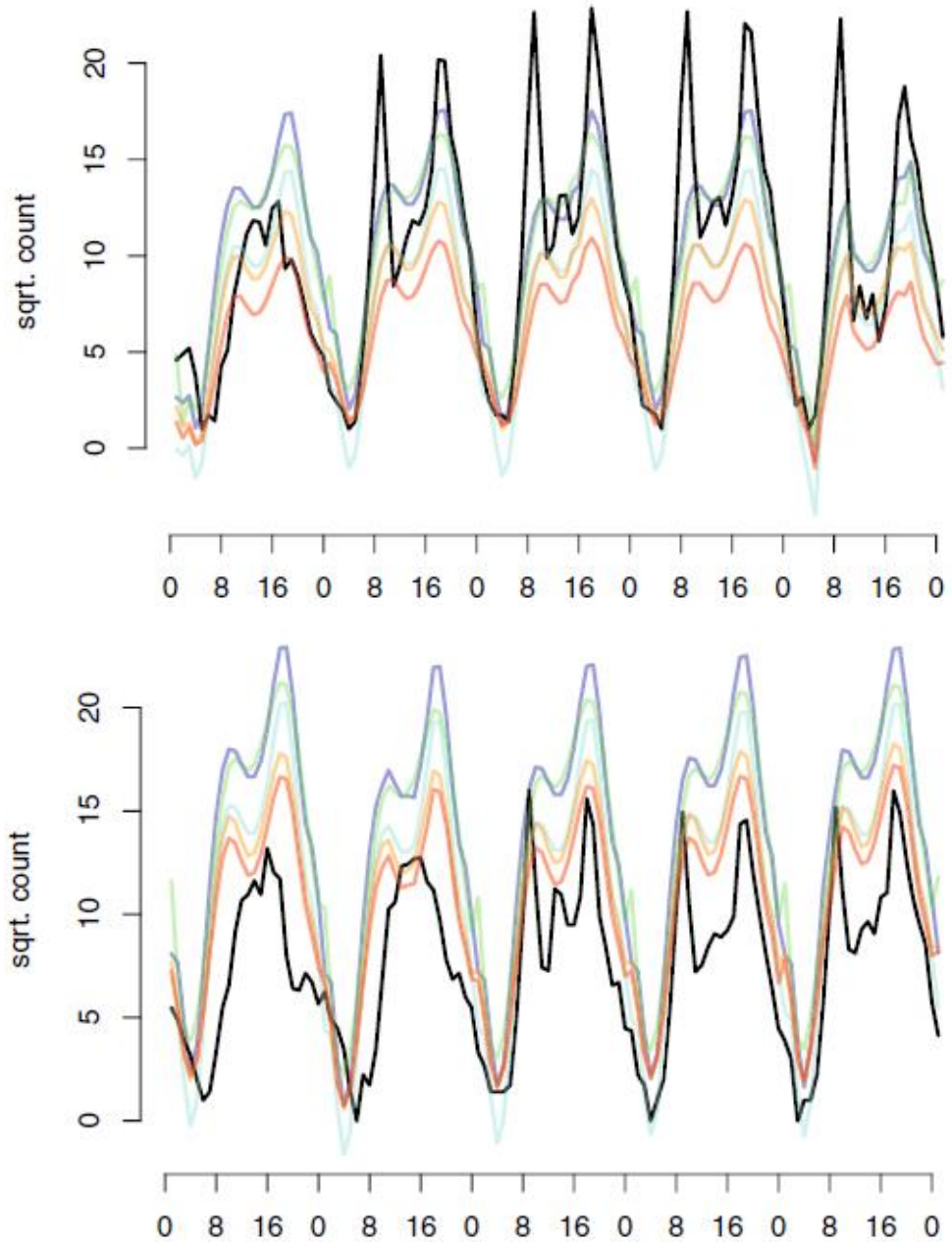


Рис. 2.3. Квадратный корень из почасового числа запросов на обработку больших массивов данных в телекоммуникационной компании, в 2021 и 2022 годах

Результаты, приведенные на рис. 2.1, соответствуют рис. 2.4. Слева видно, софтмаксимум (красный) не соответствует данным за 2022 год, в то время как низкий мягкий максимум, подобный объединению, дает хорошие прогнозы. И наоборот, на правой панели низкие значения и объединение в пул превышают высокие уровни 2022 г.

Обучение (рис. 2.4(на данных за 2021 год и тестирование на данных за 2022 год дают усредненные за неделю ошибки прогнозирования для метода обобщения оценок с более низкой медианой и более слабой поддержкой по сравнению с методами мягкой максимизации. И наоборот, обучение на данных за 2022 год и тестирование на строго положительных z с данными за 2021 год дают более стабильные прогнозы, а также более низкую медиану.


sqrt. count
0
5
10
15
20
0 8 16 0 8 16 0 8 16 0 8 16 0 8 16 0
sqrt. count
0
5
10
15
20
0 8 16 0 8 16 0 8 16 0 8 16 0 8 16 0

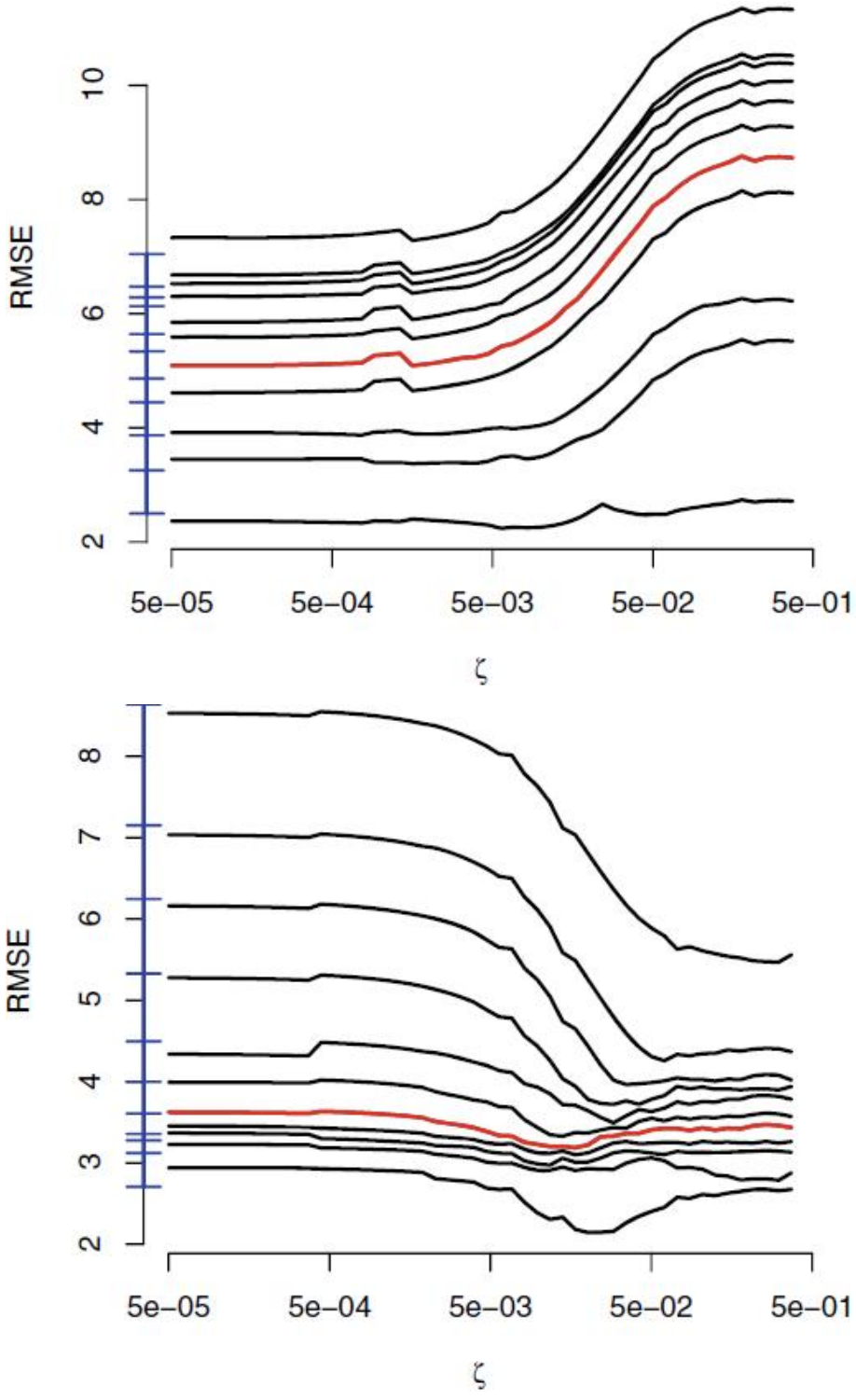


Рис. 2.4. Прогнозы на 2022 год при обучении в 2021 году и на 2021 год при обучении в 2022 году

2.6.2. Синтетический тест

Обучаем модель для $z \in \{2{,}100,\ 200\}$. Мы также вычисляем

объединенную оценку, соответствующую $z=0$, и оценку агрегирования максимумов (magging), предложенную в [2.2] . В целом, magging - это приблизительная (жесткая) оценка максимума, и поэтому она должна соответствовать $z=\infty$.

Это, в свою очередь, влечет за собой решение пяти различных задач оценки с l_1-штрафом:

• Чтобы получить три мягкие максимальные оценки, нужно решить задачу (8) с l_1-штрафом для каждого $z \in \{2, 100, 200\}$. Для этого используем R-пакет SMME [2.10].

• При одинаковом (фиксированном) распределении по группам мы получаем (скорректированную) объединенную оценку в виде (скорректированной) регрессии эмпирического среднего значения по группам при фиксированном распределении. Используем R-пакет glamlasso [2.9], чтобы решить результирующую задачу.

• Для оценки привязки мы должны решить задачу для каждой группы, учитывая ее дизайн. Используем пакет R glamlasso [2.9] чтобы подобрать индивидуальную подгонку для группы. Эти данные максимально обобщаются по группам путем решения задачи квадратичной оптимизации, предложенной в [2.2].

Имитация массива данных

Моделируем данные с тремя компонентами: (1) общий гауссовский сигнал, представляющий интерес

$$\varphi(x,y,t)=200\varphi_{12.5,4}(x)\varphi_{12.5,4}(y)\varphi_{50,25}(t)$$

где j_{m,s^2} - плотность распределения $N(m, s^2)$.

На него накладывались (2) периодические групповые сигналы со случайным изменением частоты и фазы и (3) аддитивный белый шум. Конкретно для каждого $g \in \{1,..., G\}$ массив трехмерных данных

моделировался в соответствии с

$$Y_{g,i,j,k} = j\,(x_i, y_j, t_k) + 5\sum_{j \in J_g} j_j(x_i + p_g)j_j(y_i + p_g)j_j(t_k + p_g) + e_{g,i,j,k} \qquad (2.22)$$

где x_i = 1, 2,..., 25, y_i = 1, 2,..., 25 и t_k = 1, 2,..., 101. Здесь J_g - набор из целых чисел, равномерно выбранных из {1,..., 101}, j − j-я базисная функция Фурье, $p_g \in (-p, p)$ и $e_{g,i,j,k} \in N(0, 10)$.

Общий трехмерный гауссовский сигнал, благодаря своим шлейфам, локализован как в пространстве, так и во времени.

На рис. 2.5 показаны смоделированные сигналы.

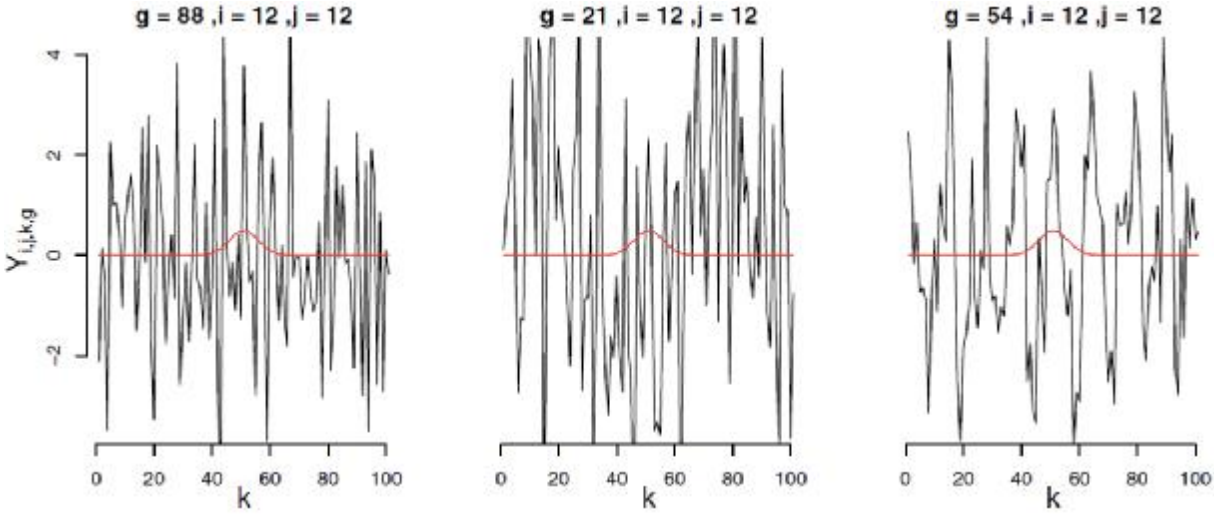


Рис. 2.5. Временной график моделируемых данных (черный) для трех различных групп и лежащий в их основе общий гауссовский сигнал (красный)

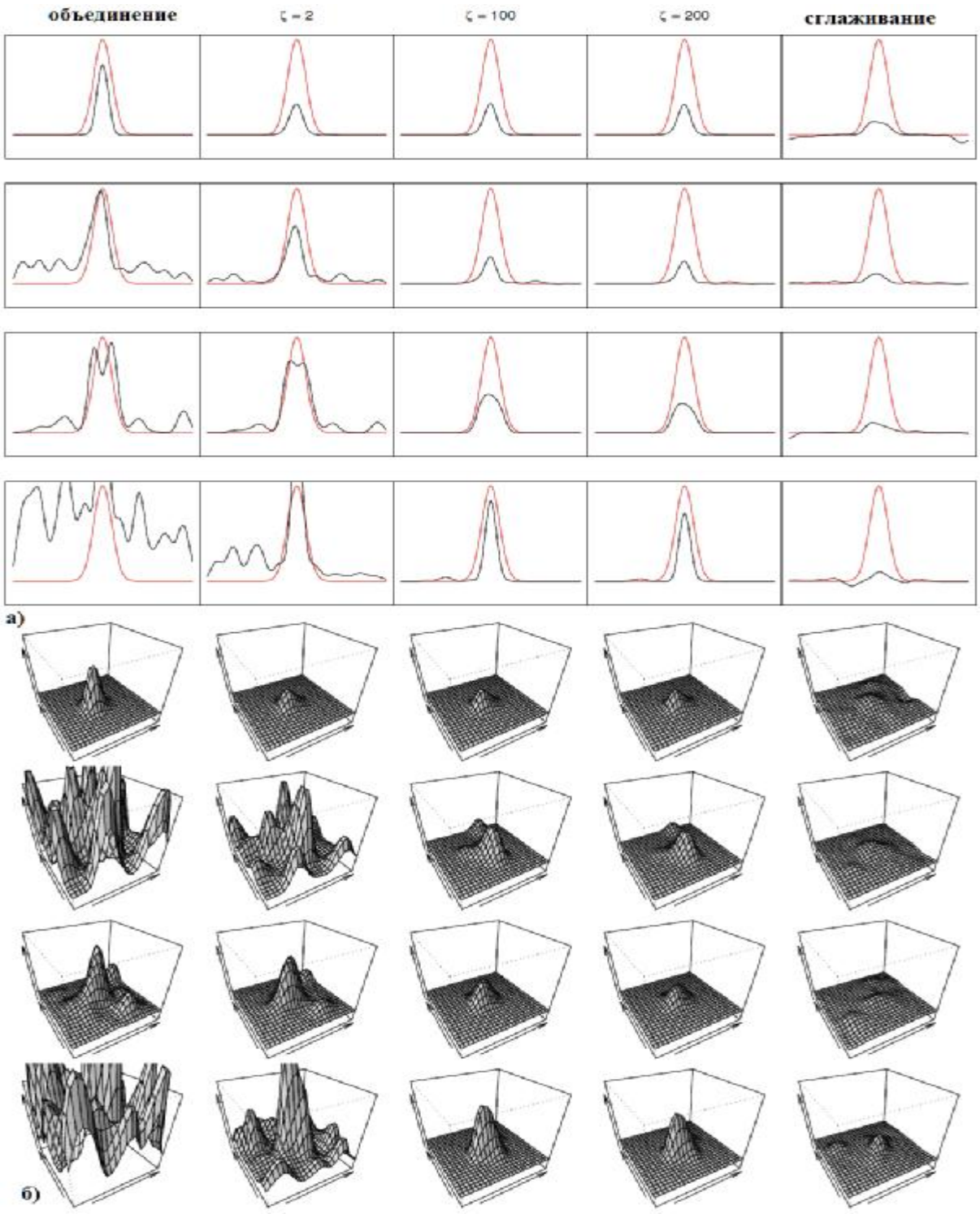


Рис. 2.6. а) Временные графики для (x,y)=(12,122). Истинный сигнал φ(x,y,t) (красный) и оценка $\hat{Y}_{x,y,t}$ (черный). Столбцы слева направо; объединение в пул, мягкий максимум $z\hat{I}$ {2, 100, 200}, сглаживание. Каждая строка соответствует обучающему набору из 14 групп; б) Пространственные графики для t = 50 оценок из наборов а)

Результаты эксперимента

Прогнозы для эксперимента для четырех различных наборов данных для проверки показаны на рис. 2.6. Для одного набора данных объединенный оценщик успешно извлекает четкий гауссовский сигнал, но для трех других он терпит неудачу. Напротив, мягкие максиминные оценки ($z \approx \{100,200\}$) хорошо работают на всех множествах и демонстрируют меньшую вариабельность.

На рис. 2.7 показано среднее значение среднеквадратичной ошибки прогнозирования (RMSPE) в зависимости от сложности модели (l) для каждого метода. RMSPE определяется как $\left\| \hat{Y}_{x,s} - Y_{-s} \right\|_2 / \sqrt{m}$, где $\hat{Y}_{x,s}$ – это прогноз, полученный в результате обучения на множестве s с использованием метода x, а Y_{-s} - это наблюдения в дополнение к s.

Пунктирная линия представляет собой среднюю ошибку, полученную при использовании нулевого сигнала, что является наиболее консервативной оценкой. Черная линия есть оптимальный прогноз.

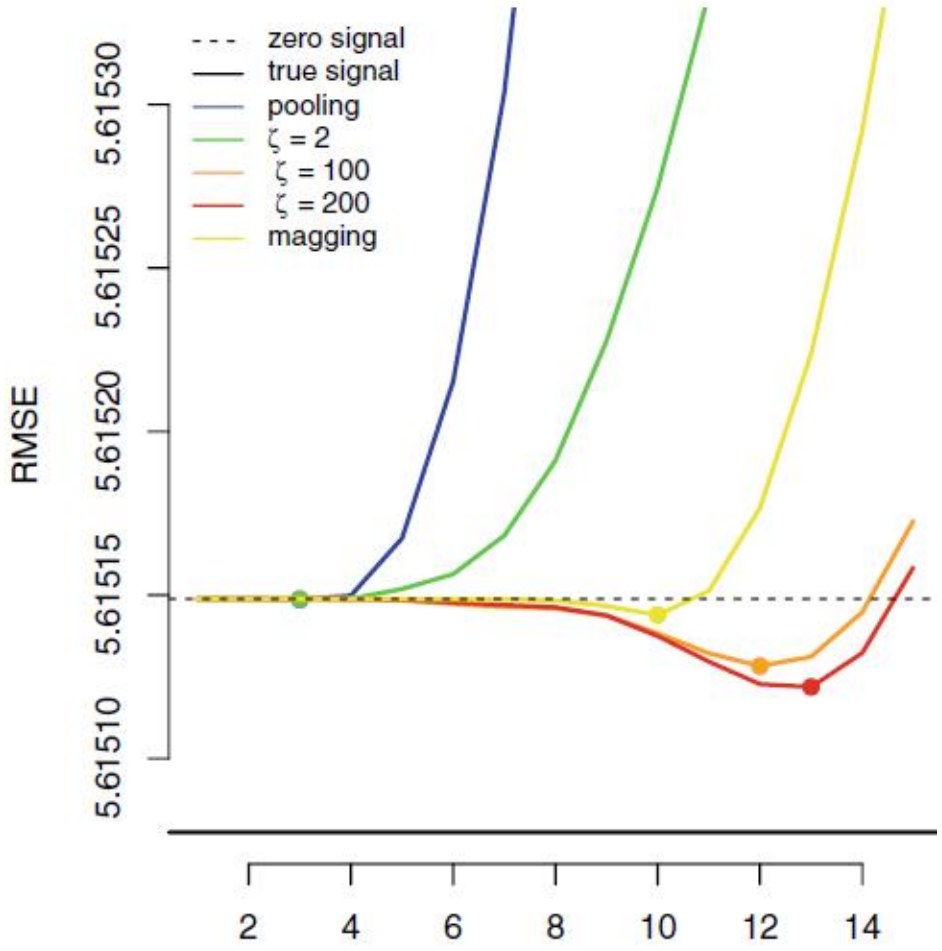
Изображение вверху на рис. 2.7 иллюстрирует различия в точности прогнозирования для различных методов с использованием их относительного отклонения в RMSE от значения нулевого прогноза. Низкие значения оценок z (объединенные и $z=2$) демонстрируют высокую вариабельность, отражая тенденцию к переопределению сигналов, характерных для конкретной группы, в данных. Оценки с высоким мягким максимумом и сглаживанием демонстрируют гораздо меньшую вариабельность. Это подчеркивает надежность методологии оценки.

Только высокоточным методам удается извлечь общий сигнал, который значительно точнее, чем нулевой прогноз.

Несмотря на то, что выигрыш в производительности прогнозирования по сравнению с нулевым прогнозированием невелик из-за низкого отношения сигнал/шум, он не является незначительным с точки

зрения качества извлекаемого сигнала, как показано на рис. 2.6. Здесь показано соответствие четырех различных тренировочных наборов, и мы непосредственно наблюдаем, как низкие методы, как правило, соответствуют групповым колебаниям по сравнению с высокими методами.

Количественно оцениваем это, вычисляя среднее значение среднеквадратичной ошибки сигнала $\left\|\hat{Y}_{x,s} - Y_{-s}\right\|_2 / \sqrt{m}$, вызванной прогнозированием, полученным путем обучения на множестве s с использованием метода x. Вверху рис. 2.8 показан результат для каждого метода x и набора s в зависимости от сложности модели, а также среднее значение по s (срезам), что подтверждает оценки, полученные на рис. 2.6 и 2.7.



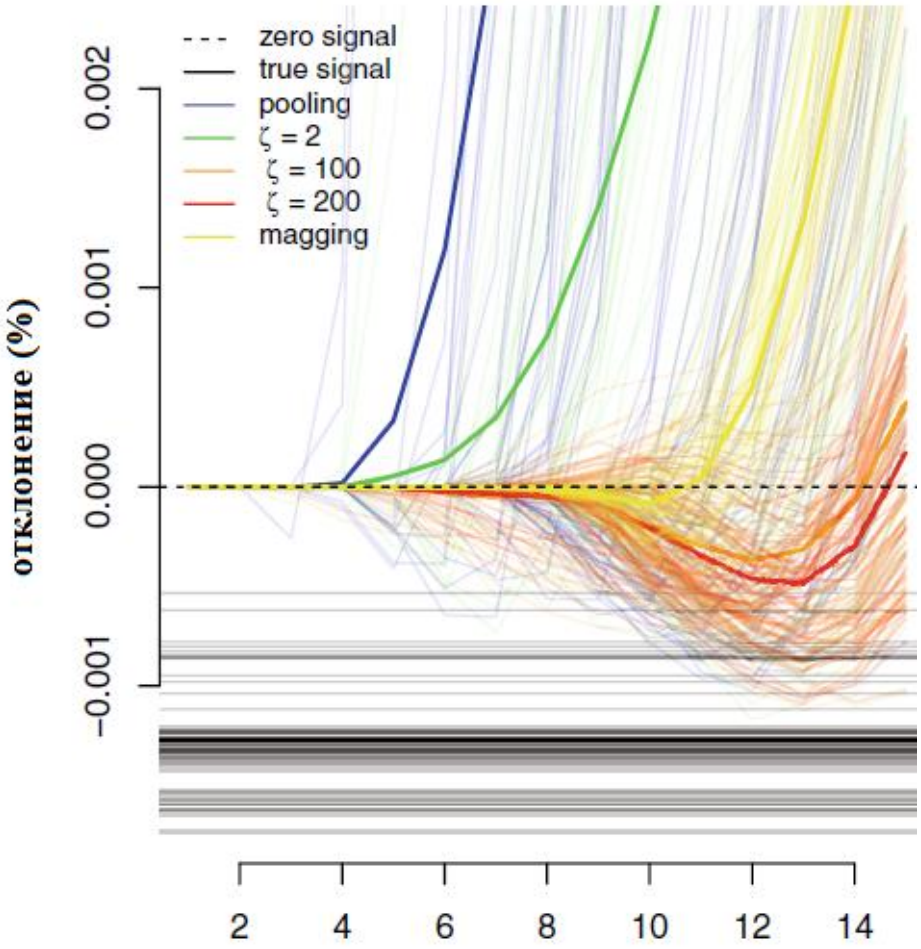
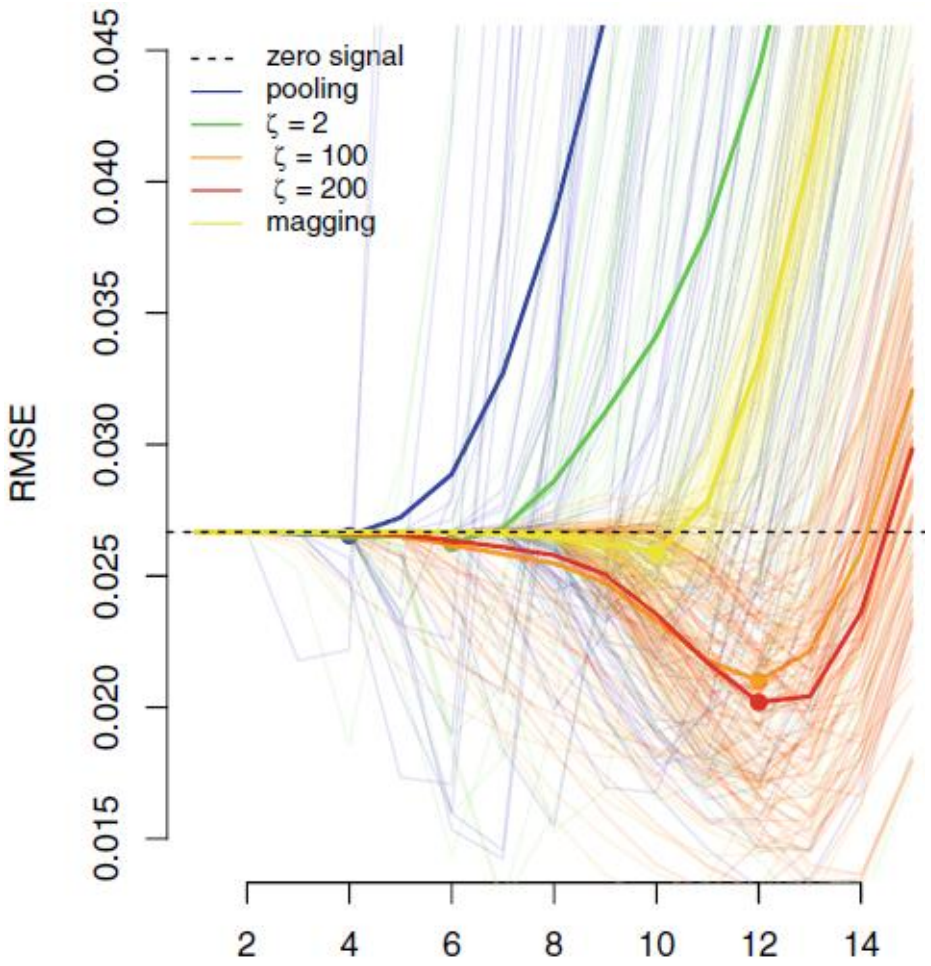


Рис. 2.7. Вверху: Среднее значение RMSE по 70 наборам тестов в зависимости от сложности модели (номер модели). Нулевой сигнал (пунктирный), истинный сигнал (черный), суммарная оценка (синий), мягкое значение максимума z=2 (зеленый), z=100 (оранжевый), z=200 (красный). Внизу: Отклонение RMSE от нулевого прогнозируемого RMSE (разница в процентах) для каждого метода и каждого набора тестов (тонкие линии) в зависимости от сложности модели. Соответствующие средние значения обозначены толстой линией того же цвета

Наконец, внизу рис. 2.8 показано, как работает каждый метод с точки зрения времени выполнения. В этой настройке, учитывая средний отклик по 14 группам одновременно, вычисление объединенной оценки имеет ту же сложность, что и вычисление соответствия одной группы в процедуре удаления. Неудивительно, что объединенный метод оценки (0,8 с) примерно в 12 раз быстрее, чем метод сглаживания (9,7 с), а также

быстрее, чем методы с высоким z (8,1 с в сравнении с 16,6 с). Однако мягкая оценка максимума с z=2 (0,5 с) выполняется быстрее, чем объединенная оценка по методу наименьших квадратов.



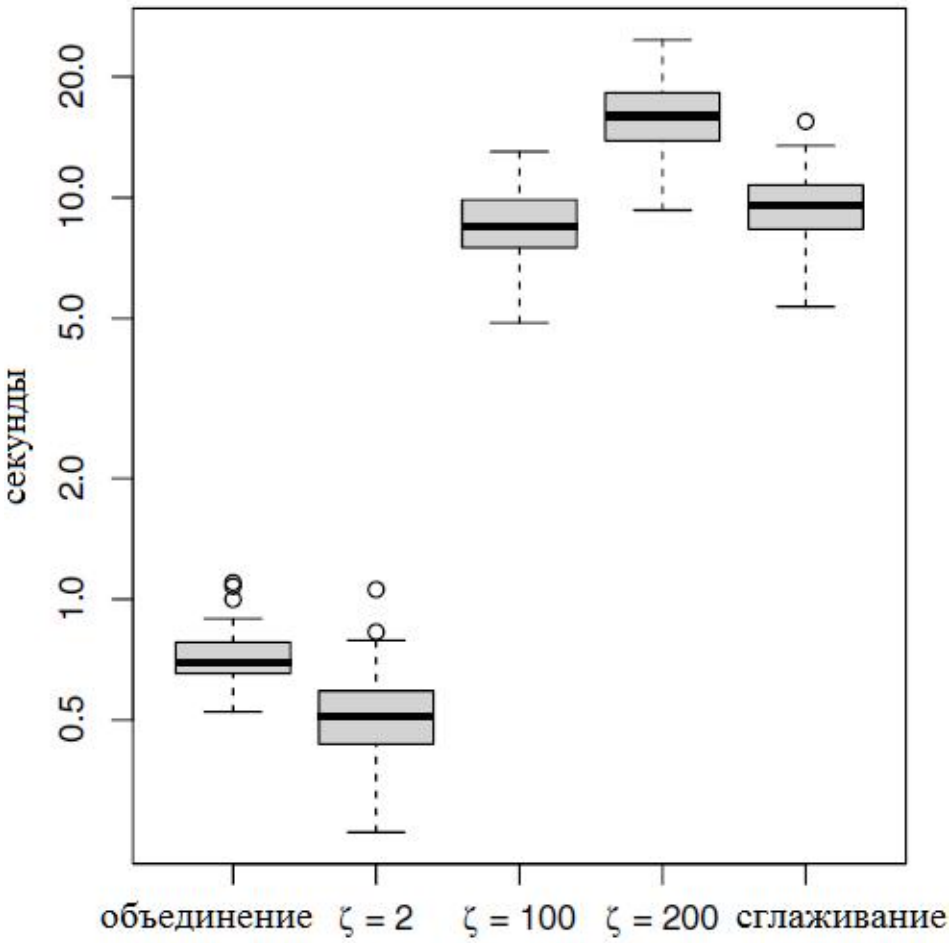


Рис. 2.8. Вверху: Среднеквадратичная ошибка сигнала, полученная на основе каждого из 70 обучающих наборов. Суммарная оценка (синий), мягкое значение максимума z=2 (зеленый), z=100 (оранжевый), z=200 (красный). Внизу: Сводные данные о времени выполнения (логарифмическая шкала) для 70 тренировочных наборов для каждого метода

Выводы к главе 2

Разработана мягкая максиминная оценка в качестве альтернативы максиминной оценке, которая сохраняет желаемые статистические свойства и является более эффективной с точки зрения вычислений.

Кроме того, параметр soft maximin определяет соотношение между группами с большой объясненной дисперсией и группами с малой объясненной дисперсией, что приводит к интерполяции объединенной

оценки и оценки максимального значения.

Градиентное представление (2.10) наглядно показывает, как работает этот компромисс: градиент мягких максимальных потерь представляет собой выпуклую комбинацию градиентов групповых функций потерь с весами, управляемыми x. Наибольшие веса имеют те группы, у которых наименьшие объясненные отклонения, а при $x \to \infty$ веса концентрируются на группах с минимальными объясненными отклонениями.

Источники к главе 2

2.1. Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1), 183–202.

2.2. Bühlmann, P., & Meinshausen, N. (2016). Magging: Maximin aggregation for inhomogeneous large-scale data. Proceedings of the IEEE, 104(1), 126–135.

2.3. Buis, P. E., & Dyksen, W. R. (1996). Efficient vector and parallel manipulation of tensor products. ACM Transactions on Mathematical Software (TOMS), 22(1), 18–23.

2.4. Chen, X., Lu, Z., & Pong, T. K. (2016). Penalty methods for a class of non-lipschitz optimization problems. SIAM Journal on Optimization, 26(3), 1465–1492.

2.5. Currie, I. D., Durban, M., & Eilers, P. H. (2006). Generalized linear array models with applications to mul-tidimensional smoothing. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(2), 259–280.

2.6. De Boor, C. (1979). Efficient computer manipulation of tensor products. ACM Transactions on Mathematical Software (TOMS), 5(2), 173–182.

2.7. Fanaee T. H., & Gama, J. (2023). Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence, 2(2), 113–127.

2.8. Grinvald, A., & Bonhoeffer, T. (2002). Optical imaging of electrical activity based on intrinsic signals and on voltage sensitive dyes: The methodology.

2.9. Lund, A. (2018). glamlasso: Penalization in large scale generalized linear array models. R package version 3.0.

2.10. Lund, A. (2021). SMME: Soft maximin estimation for large scale heterogeneous data. R package version 1.0.1.

2.11. Lund A., Mogensen S.W., Hansen N.R. Soft Maximin Estimation for Heterogeneous Data, 2022. - https://arxiv.org/pdf/1805.02407.

2.12. Meinshausen, N., & Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. The Annals of Statistics, 43(4), 1801–1830.

2.13. Roland, P. E., Hanazawa, A., Undeman, C., Eriksson, D., Tompa, T., Nakamura, H., Valentiniene, S., & Ahmed, B. (2006). Cortical feedback depolarization waves: A mechanism of top-down influence on early visual areas. Proceedings of the National Academy of Sciences, 103(33), 12586–12591.

2.14. Roll, J. (2008). Piecewise linear solution paths with application to direct weight optimization. Automatica, 44(11), 2732–2737.

2.15. Rothenhäusler, D., Meinshausen, N., Bühlmann, P., & Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. Journal of the

Royal Statistical Society: Series B (Statistical Methodology), 83(2), 215–246.

2.16. Tseng, P., & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming, 117(1-2), 387–423.

2.17. Wright, S.J., Nowak, R.D., & Figueiredo, M.A. (2009). Sparse reconstruction by separable approximation. IEEE Transactionson Signal Processing, 57(7), 2479–2493.

2.18. Tseng, P., & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization.

2.19. Atlasov D.I., Kravets O.Ja. To the formulation of the problem of extracting a common signal from heterogeneous data of heterogeneous information systems// Modern informatization problems in simulation and social technologies (MIP-2023'SCT): Proceedings of the XXVIII-th International Open Science Conference (Yelm, WA, USA, January 2023). - Yelm, WA, USA: Science Book Publishing House, 2023. – Pp. 8-13.

2.20. Атласов Д.И., Кравец О.Я., Пашкевич А.С. Управление специальной информацией в беспроводной сети на основе гетерогенной последовательности данных// Системы управления и информационные технологии, №1(91), 2023. – С. 49-55.

2.21. Атласов Д.И., Красновский Е.Е., Сараев П.В. Пути интерполяции мягкой максиминной оценки для гетерогенных данных// Системы управления и информационные технологии, №4(94), 2023.

3. Измерение неопределенности гетерогенных данных и редукция атрибутов в гетерогенных информационных системах

В эпоху больших данных появляются мультимедиа, гипермедиа и социальные сети, а объем информации стремительно растет. Когда люди участвуют в процессе массовой обработки данных, они сталкиваются с данными, имеющими различную структуру, поэтому данные неоднородны. Важной проблемой в области искусственного интеллекта является то, как извлекать скрытые и ценные знания из разнородных данных и измерять их неопределенность. В данной работе исследуется измерение неопределенности для разнородных данных и приводится его применение для сокращения атрибутов. Сначала предлагается концепция гетерогенной информационной системы (HIS).

Затем строится отношение эквивалентности на множестве объектов. Затем исследуется измерение неопределенности для HIS, проводится численный эксперимент и проводится дисперсионный анализ, корреляционный анализ, а также тест Фридмана и тест Бонферрони–Данна в статистике. Наконец, в качестве применения предложенных мер изучается уменьшение атрибутов в HIS, и предлагаются соответствующие алгоритмы и их анализ.

3.1. Проблема измерения неопределенности гетерогенных данных и редукции атрибутов в гетерогенных информационных системах

3.1.1. Связанные с проблемой работы

Неопределенность широко распространена в реальном мире, и ее формы разнообразны. Случайность и нечеткость - это два основных вида неопределенности. Случайность означает отсутствие четкой причинно-

следственной связи между условиями и результатами, поскольку условия для событий недостаточны. Нечеткость относится к тому, соответствует ли объект этой концепции, которую трудно определить, и не имеет четкого значения в природе, поскольку нет определенного ограничения по количеству. В отличие от случайности и нечеткости, шероховатость - это еще один вид неопределенности, которому в последние годы исследователи уделяют все больше внимания. Это относится к различению объектов из-за неполноты или неточности знаний.

Что касается нечеткости и шероховатости, то для их изучения мы можем использовать теорию нечетких множеств [3.27] и теорию грубых множеств [3.15] соответственно. Теория грубых множеств - это математическая теория, которая имеет дело с различной неполной информацией, такой как неточная, противоречивая и неокончательная информация. У этой теории есть несколько уникальных точек зрения, которые делают грубые множества особенно подходящим для анализа данных. Например, детализация знаний. Согласно теории грубых множеств, детализация знаний является причиной того, что некоторые понятия не могут быть точно выражены с помощью существующих знаний. Эта теория определяет знания как совокупность взаимосвязанных элементов, которые придают знаниям четкий математический смысл и могут быть обработаны математическими методами. Она предоставляет математический метод обнаружения знаний, который имеет много преимуществ. Эта теория также позволяет анализировать факты, скрытые в данных, без какой-либо дополнительной информации о них. В настоящее время эта теория успешно применяется в области искусственного интеллекта, поиска знаний и данных, распознавания образов и классификации, обнаружения неисправностей [3.16-3.18].

Из-за широкого применения этой теории она привлекает все большее внимание международного академического сообщества [3.11, 3.13, 3.26, 3.28].

В последние годы все больше внимания уделяется исследованию неопределенности в теории приблизительных множеств. Размер зерна знаний в приближенном пространстве напрямую влияет на неопределенность приблизительного множества. С точки зрения интуитивного понимания, чем больше объем знаний, тем меньше информации, тем больше будет неопределенность; чем меньше объем знаний, тем больше информации, тем меньше будет неопределенность. Это центр исследований в области искусственного интеллекта, и это также важная передовая тема в области искусственного интеллекта. В качестве математического инструмента для работы с неточными и неполными данными информационная энтропия [3.19] является эффективным методом борьбы с неопределенностью. В настоящее время измерение неопределенности является важным направлением исследований. Например, в [3.4] применена энтропия Шеннона к измерению решающих правил в теории грубых множеств; в [3.2] исследован основанный на энтропии подход к измерению неопределенности в системах окрестности; в [3.10] исследовано измерение неопределенности для информационной системы с нечеткими отношениями; в [3.5] изучено измерение неопределенности для систем принятия решений с интервальными значениями, основанных на расширенной условной энтропии; в [3.24] представлены новые показатели неопределенности для информационной системы с интервальными значениями; в [3.20] исследованы нечеткие информационные гранулы высшего типа, полученные в результате измерения неопределенности; в [3.23] исследована мера неопределенности в теории доказательств и ее приложения; в [3.21] рассмотрена грубая аппроксимаця нечеткой концепции в гибридной атрибутивной

информационной системе и ее меру неопределенности; в [3.12] измерена неопределенность полностью нечеткой информационной системы с помощью гауссова ядра.

3.1.2. Важность исследования неопределенности для гетерогенной информационной системы

Неопределенность присутствует повсюду, и информационная система не является исключением. Гетерогенная информационная система (HIS) или информационная система с гетерогенными данными означает информационную систему, наборы данных которой содержат данные трех типов (т.е. масштабированные типы, упорядоченные типы и обычные типы). Кроме того, гетерогенная информационная система обладает неопределенностью. Измерение неопределенности для гетерогенной информационной системы отражает способность этой системы к классификации, которая оказывает существенное влияние на точность классификации данных.

Следовательно, поиск значений неопределенности для гетерогенной информационной системы является важной темой исследования. В [3.7] указано, что приблизительные множества окрестностей являются гибкими для обработки гетерогенных данных. Кроме того, нечеткие приблизительные множества могут также использоваться для обработки гетерогенных данных. Поэтому уже были предложены методы измерения информационной энтропии или неопределенности, приведенные в [3.8, 3.22, 3.29, 3.31], которые применимы для управления разнородными данными.

Для обработки разнородных данных в [3.22] введено нечеткое отношение в наборе объектов HIS для каждого атрибута и определили отношение эквивалентности в наборе объектов, основанное на равенстве гранул нечеткой информации о двух точках. Трудно добиться равенства

гранул нечеткой информации о двух точках. Рассматриваем нечеткое отношение, вводя приблизительное равенство между нечеткими множествами, определяем отношение эквивалентности на множестве объектов HIS для каждого атрибута, предлагаем измерения неопределенности для HIS и приводим применение предложенных измерений для уменьшения атрибута в HIS.

Оставшаяся часть материала организована следующим образом. В разделе 3.2 кратко представлены некоторые связанные понятия о нечетких отношениях, нечеткой энтропии и гетерогенных информационных системах. В разделе 3.3 предлагаются некоторые инструменты для измерения неопределенности гетерогенной информационной системы. В разделе 3.4 приводится численный эксперимент и проводится анализ эффективности с точки зрения двух аспектов - дисперсии и корреляции в статистике. В разделе 3.5 приводится применение предложенных мер для уменьшения атрибутов в гетерогенной информационной системе. В разделе 3.6 приводятся выводы.

3.2. Связанные понятия о нечетких отношениях, нечеткой энтропии и гетерогенных информационных системах

В этом разделе мы кратко представим некоторые связанные понятия о нечетких отношениях, нечеткой энтропии и гетерогенных информационных системах.

U обозначает непустое конечное множество, а I есть [0, 1]. Обозначим

$U=\{x_1, x_2, \ldots, x_n\}$.

3.2.1. Нечеткие отношения

Напомним, что F является нечетким множеством всякий раз, когда F является функцией, определяемой формулой $F \in U \rightarrow I$.

Для $a \in I$, ā указывает на постоянное нечеткое множество в U, т.е. $\forall u \in U$, ā(u)=a.

В этой статье I^U показывает набор нечетких множеств в U.

Учитывая , что $F \in I^U$, $|F| = \sum_{u \in U} F(u)$, F(u) представляет мощность F.

Если R - нечеткое множество в U×U, то R называется нечетким отношением в U.

В этой статье $I^{U\times U}$ обозначает совокупность всех нечетких отношений на U.

Пусть $R \in I^{U\times U}$. Тогда R может быть выражено следующей матрицей

$M(R)=(R(x_i, x_j))_{n\times n}$

Если M(R)=E (здесь E - единичная матрица), то R называется нечетким тождественным отношением на U, и мы записываем как R=△; если M(R)=0, то R называется нечетким нулевым отношением на U, и мы записываем как R=o; если $R(x_i, x_j)=1$ для любого i, j, то R называется нечетким универсальным соотношением на U, и мы записываем как R=ω.

3.2.2. Нечеткая энтропия для нечеткого отношения

Определение 3.2.1 [3.10] Пусть $R \in I^{U\times U}$. Для любого $x \in U$, определим

$G_R(x)(y)=R(x,y), \ " \ y \in U.$

Тогда $G_R(x)$ называется гранулой нечеткой информации точки x относительно R.

В [22] $G_R(x)$ обозначено как $[x]_R$.

Определение 3.2.2 [3.22] Пусть $R \in I^{U\times U}$. Определим

$ER=\{(x,y) \in U\times U: G_R(x)=G_R(y)\}$

Легко видеть, что ER - это отношение эквивалентности на U. Тогда ER может индуцировать разбиение U, которое обозначается через U\ER.

Определение 3.2.3 [3.22] Пусть $R \in I^{U\times U}$. Обозначим

$U\backslash ER=\{X_1, X_2, \ldots, X_m\}$.

Тогда нечеткую энтропию R определим как

$$H(R)=-\sum_{i=1}^{m} \frac{|X_i|}{n} \log_2 \frac{|G_R(x)|}{n}, x \in X_i$$

Нечеткую энтропию можно рассматривать как расширение энтропии Шеннона [3.19].

Утверждение 3.2.4 [3.22] Выполняются следующие свойства.

(1) $H(\Delta)=\log_2 n$.

(2) $H(\omega) = 0$.

(3) Пусть $R \in I^{U\times U}$. Если R рефлексивно, то $0 \le H(R) \le \log_2 n$.

(4) Если R - четкое отношение эквивалентности на U, то

$$H(R)=\sum_{i=1}^{m} \frac{|X_i|}{n} \log_2 \frac{|X_i|}{n},$$

где $U\backslash R = \{X_1, X_2,\ldots, X_m\}$.

(5) Пусть $R_1, R_2 \in I^{U\times U}$. Если $R_2 \subseteq R_1$, то $H(R_1) \le H(R_2)$.

Определение 3.2.5 [3.22] Пусть $R_1, R_2 \in I^{U\times U}$. Обозначим

$U\backslash ER_1=\{X_1, X_2,\ldots, X_m\}$, $U\backslash ER_2=\{Y_1, Y_2, \ldots, Y_l\}$.

Тогда условная нечеткая энтропия R_1 от R_2 определяется как

$$H(R_1 / R_2) = -\sum_{i=1}^{m} \sum_{j=1}^{l} \frac{|X_i \cap Y_j|}{n} \log_2 \frac{|G_{R_1}(x) \cap G_{R_2}(y)|}{|G_{R_2}(y)|}, x \in X_i, y \in Y_j$$

Очевидно, что из $R_2 \subseteq R_1$ следует $H(R_1/R_2)=0$.

Основанный на определенной нечеткой энтропии для нечеткого отношения, метод редукции в информационной системе был предложен в работе [3.22].

3.2.3. Гетерогенная информационная система

Определение 3.2.6 [3.15] Пусть U - конечный набор объектов. Предположим, что A выражает конечный набор атрибутов. Тогда упорядоченная пара (U, A) называется информационной системой, если $a \in A$ может определять функцию $a: U \to V_a$, где $V_a=\{a(u): u \in U\}$.

Определение 3.2.7 Пусть (U, A) - информационная система. Тогда (U, A) называется гетерогенной информационной системой (HIS) или информационной системой с гетерогенными данными, если ее набор данных содержит три типа данных (т.е. масштабируемые типы, упорядоченные типы и обычные типы).

Пусть (U, A) - это HIS. Задано $a \in A$. Если a масштабируется или является порядковым номером, то для $x \in U$ запишем $a(x)' = \frac{a(x) - a_{min}}{a_{max} - a_{min}}$

где $a_{max}=\max\{a(u): u \in U\}$, $a_{min}=\min\{a(u): u \in U\}$.

Для любого $a \in A$, определим

$$R_a(x,y) = \begin{cases} 1, a(x) = a(y) \\ 1 - |a(x)' - a(y)'|, a(x) \neq a(y), a \text{ масштабируется} \\ \max\{a(x)' - a(y)', 0\}, a(x) \neq a(y), a - \text{порядковое} \\ 0, a(x) \neq a(y), a - \text{номинальное} \end{cases}$$

Затем мы получаем нечеткое отношение $R_a=(R_a(x_i, x_j))_{n \times n}$.

Пусть (U, A) - это HIS. Задано $a \in A$ и $x, y \in U$. $G_{R_a}(x), G_{R_a}(y) \in I^U$. Тогда $G_{R_a}(x) = G_{R_a}(y)$ тогда и только тогда, когда для любого $u \in U$, $G_{R_a}(x)(u) = G_{R_a}(y)(u)$. Следует отметить, что $G_{R_a}(x)(u)$ и $G_{R_a}(y)(u)$ есть два числа в единичном интервале [0, 1]. Тогда равенство, "$G_{R_a}(x)(u) = G_{R_a}(y)(u)$" на самом деле очень трудно достижимо. Таким образом, равенство "$G_{R_a}(x) = G_{R_a}(y)$" также очень трудно достижимо.

Описанная выше ситуация указывает на то, что определение 3.2.2 нуждается в улучшении.

Ниже мы попытаемся разобраться с нечетким отношением R_a, введя приблизительное равенство между нечеткими множествами.

Определение 3.2.8 [3.30] Пусть $k \in N$. Возьмем a, $b \in [0, 1]$. Если $a=b=0$ или $a, b \in \frac{1}{10^k}$ или …, …, или $a = b = \frac{10^k - 1}{10^k}$ или $a, b \in \left[\frac{10^k - 1}{10^k}, 1\right]$ илиr $a=b=1$, тогда a и b считаются классово совместимыми, а k - пороговым значением. Обозначим это как $a \approx_k b$.

В раоте выбираем k=1.

Определение 3.2.9 [3.30] Предположим, что $A, B \in I^U$. Тогда

$$A \approx_1 B \Leftrightarrow \forall x \in U, A(x) \approx_1 B(x)$$

Определение 3.2.10 Пусть (U, A) - HIS. Дано $B \subseteq A$. Определим

$$R_a^* \text{ или } R_{\{a\}}^* = \left\{(x,y) \in U \times U : G_{R_a}(x) \approx_1 G_{R_a}(y)\right\},$$
$$R_B^* = \bigcap_{a \in B} R_a^*$$

Легко видеть, что R_B^* - отношение эквивалентности на U. Тогда R_B^* называется отношением эквивалентности, индуцированным B в (U, A). А разбиение на U, индуцированное с помощью R_B^*, обозначается как $U \setminus R_B^*$

Для любого $x \in U$, обозначим

$$[x]_{R_B^*} = \left\{y \in U : (x,y) \in R_B^*\right\}$$

3.3. Измерение неопределенности HIS

В [22] представлены два типа показателей неопределенности (нечеткая энтропия и энтропия нечеткой окрестности) для общих нечетких отношений. В этом разделе мы используем четыре инструмента (показатели детализации, информационная энтропия, грубая энтропия и объем информации) для измерения неопределенности HIS.

3.3.1. Мера детализации для HIS

Определение 3.3.1 Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда детализация информации в подсистеме (U, B) определяется следующим образом

$$G(B) = -\frac{1}{n^2}\sum_{i=1}^{m}|X_i|^2$$

где $U \setminus R_B^* = \{X_1, X_2, ..., X_m\}$.

Утверждение 3.3.2 Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда

$$G(B) = \frac{1}{n^2}\sum_{i=1}^{n}\left|[x_i]_{R_b^*}\right|$$

Доказательство. Обозначим $U \setminus R_B^* = \{X_1, X_2, ..., X_m\}$. Предположим, что $X_i=\{x_{i1}, x_{i2},\ldots, x_{isi}\}$. Тогда $|X_i|=s_i$. Таким образом,

$$\sum_{i=1}^{m} s_i = n, X_i = [x_{i1}]_{R_B^*} = [x_{i2}]_{R_B^*} = ... = \left[x_{i,s_i}\right]_{R_B^*}.$$

Отсюда следует, что

$$X_i = \left|[x_{i1}]_{R_B^*}\right| = \left|[x_{i2}]_{R_B^*}\right| = ... = \left|\left[x_{i,s_i}\right]_{R_B^*}\right| = s_i$$

Следовательно, $\forall i$,

$$|X_i|^2 = s_i|X_i| = \sum_{k=1}^{s_i}\left|[x_{ik}]_{R_B^*}\right|$$

Поэтому

$$G(B) = \frac{1}{n^2}\sum_{i=1}^{m}|X_i|^2 = \frac{1}{n^2}\sum_{i=1}^{m}\sum_{k=1}^{s_i}\left|[x_{ik}]_{R_B^*}\right| = \frac{1}{n^2}\sum_{i=1}^{n}\left|[x_i]_{R_B^*}\right|$$

Утверждение доказано.

Утверждение иллюстрирует, что детализация информации в HIS полностью основана на информационной структуре этого нечеткого отношения.

Утверждение 3.3.3 (Ограниченность) Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда

$$\frac{1}{n} \leq G(B) \leq 1.$$

Более того, если $R_B^* = \Delta$, то G(B) достигает минимального значения $\frac{1}{n}$; если $R_B^* = \omega$, то G(B) достигает максимального значения 1.

Доказательство.

(1) Необходимо отметить, что $\forall i$, $1 \leq \left|[x_i]_{R_B^*}\right| \leq n$.

Тогда $\forall i$,

$$n \leq \sum_{i=1}^{n} \left|[x_i]_{R_B^*}\right| \leq n^2$$

Из Утверждения, $\frac{1}{n} \leq G(B) \leq 1$.

(2) Предположим, что $R_B^* = \Delta$. Тогда

$$R_B^* = \{(x_1, x_1), (x_2, x_2), ..., (x_n, x_n)\}$$

Это подразумевает $U \setminus R_B^* = \{\{x_1\}, \{x_2\}, ..., \{x_n\}\}$. Пусть $\forall i$, $\left|[x_i]_{R_B^*}\right| = 1$. Следовательно, G(B)= $\frac{1}{n}$.

(3) Предположим, что $R_B^* = \omega$. Тогда $R_B^* = U \times U$. Это подразумевает $U \setminus R_B^* = \{\{x_1\}, \{x_2\}, ..., \{x_n\}\}$. Пусть $\forall i$, $\left|[x_i]_{R_B^*}\right| = n$. Следовательно, G(B)=1.

Утверждение доказано.

Утверждение 3.3.3 дает диапазон значений грануляции информации для неопределенности HIS.

Утверждение 3.3.4 (Монотонность) Пусть (U, A) - это HIS. Если $P \subseteq Q \subseteq A$, то $G(Q) \leq G(P)$.

Доказательство Из Утверждения 3.3.2,

$$G(P)=\frac{1}{n^2}\sum_{i=1}^{n}\left|[x_i]_{R_P^*}\right|,\quad G(Q)=\frac{1}{n^2}\sum_{i=1}^{n}\left|[x_i]_{R_Q^*}\right|$$

Поскольку $P\subseteq Q\subseteq A$, имеем R_Q^* Í R_P^* Then $\forall i$, $[x_i]_{R_Q^*}$ Í $[x_i]_{R_P^*}$

Тогда $\forall i$,

$$\left|[x_i]_{R_Q^*}\right| £ \left|[x_i]_{R_P^*}\right|$$

Следовательно, $G(Q)\leq G(P)$.

Утверждение доказано.

Детализация информации соответствует характеристике грануляционных вычислений и уточняет информацию с разных уровней. Этот показатель монотонно уменьшается по мере уточнения эквивалентного класса. Недостаток заключается в том, что этот показатель не позволяет отличить два обращения с одинаковой детализацией информации, но разными информационными структурами.

3.3.2. Мера энтропии для HIS

В физике энтропия часто используется для измерения степени нарушения порядка в системе. Чем больше значение энтропии, тем выше степень нарушения порядка в системе. В [3.19] применено понятие энтропии в физике к теории информации для измерения неопределенности системы.

Определение 3.3.5 Пусть (U, A) - это HIS. Задано, что $B\subseteq A$. Тогда информационная энтропия подсистемы (U, B) определяется следующим образом:

$$H(R)=-\sum_{i=1}^{m}\frac{|X_i|}{n}\log_2\frac{|X_i|}{n}$$

где

$$U \setminus R_B^* = \{X_1, X_2, ..., X_m\}.$$

Утверждение 3.3.6 Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда

$$H(B) = -\sum_{i=1}^{m} \frac{1}{n} \log_2 \frac{\left|[x_i]_{R_B^*}\right|}{n}$$

Доказательство. Обозначим

$$U \setminus R_B^* = \{X_1, X_2, ..., X_m\}$$

Пусть $X_i = \{x_{i1}, x_{i2}, \ldots, x_{i,si}\}$. Тогда $|X_i| = s_i$. Отсюда

$$\sum_{i=1}^{m} s_i = n, X_i = [x_{i1}]_{R_B^*} = [x_{i2}]_{R_B^*} = ... = \{x_{is_i}\}_{R_B^*}$$

Отсюда следует, что

$$|X_i| = \left|[x_{i1}]_{R_B^*}\right| = \left|[x_{i2}]_{R_B^*}\right| = ... = \left|\{x_{is_i}\}_{R_B^*}\right| = s_i$$

Поэтому $\forall i$,

$$\frac{|X_i|}{n} \log_2 \frac{|X_i|}{n} = s_i \frac{1}{n} \log_2 \frac{|X_i|}{n} = \sum_{k=1}^{s_i} \frac{1}{n} \log_2 \frac{\left|[x_{ik}]_{R_B^*}\right|}{n}$$

Следовательно

$$H(B) = -\sum_{i=1}^{m} \frac{|X_i|}{n} \log_2 \frac{|X_i|}{n} = \sum_{i=1}^{m} \sum_{k=1}^{s_i} \frac{1}{n} \log_2 \frac{\left|[x_{ik}]_{R_B^*}\right|}{n} = -\sum_{k=1}^{s_i} \frac{1}{n} \log_2 \frac{\left|[x_i]_{R_B^*}\right|}{n}$$

Утверждение доказано.

Утверждение 3.3.7 Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. тогда $0 \le H(B) \le \log_2 n$.

Более того, если $R_B^* = \Delta$, то H(B) достигает максимального значения $\log_2 n$; если $R_B^* = \omega$, то H(B) достигает минимального значения 0.

Доказательство

(1) Из Утверждения 3.3.6,

$$H(B)=-\sum_{i=1}^{m}\frac{1}{n}\log_2\frac{\left|[x_i]_{R_B^*}\right|}{n}$$

$\forall i$, $1\le\left|[x_i]_{R_B^*}\right|\le n$. Тогда

$$0\le -\log_2\frac{\left|[x_i]_{R_B^*}\right|}{n}\le\log_2 n$$

Следовательно $0\le H(B)\le\log_2 n$.

(2) Пусть $R_B^*=\Delta$. Тогда

$$R_B^*=\{(x_1,x_1),(x_2,x_2),...,(x_n,x_n)\}$$

Отсюда $U\setminus R_B^*=\{\{x_1\},\{x_2\},...,\{x_n\}\}$. Поэтому $\forall$ i, $\left|[x_i]_{R_B^*}\right|=1$. Следовательно, $0\le H(B)\le\log_2 n$.

(3) Пусть $R_B^*=\omega$. Тогда $R_B^*=U\times U$. Отсюда $U\setminus R_B^*=\{\{x_1\},\{x_2\},...,\{x_n\}\}$. Поэтому $\forall i$, $\left|[x_i]_{R_B^*}\right|=n$. Следовательно, $H(B)=0$.

Утверждение доказано.

Утверждение 3.3.8 (Монотонность) Пусть (U, A) - это HIS. Если $P\subseteq Q\subseteq A$, то $H(P)\le H(Q)$.

Доказательство. Из Утверждения 3.3.6,

$$H(P)=-\sum_{i=1}^{m}\frac{1}{n}\log_2\frac{\left|[x_i]_{R_P^*}\right|}{n},\ H(Q)=-\sum_{i=1}^{m}\frac{1}{n}\log_2\frac{\left|[x_i]_{R_Q^*}\right|}{n}$$

Поскольку $P\subseteq Q\subseteq A$, имеем $R_Q^*\ \text{Í}\ R_P^*$. Тогда $\forall i$, $[x_i]_{R_Q^*}\ \text{Í}\ [x_i]_{R_P^*}$.

Поэтому $\forall i$,

$$\left|[x_i]_{R_Q^*}\right|\le\left|[x_i]_{R_P^*}\right|$$

Тогда, $\forall i$,

$$- \log_2 \frac{\left|[x_i]_{R_P^*}\right|}{n} = \log_2 \frac{n}{\left|[x_i]_{R_P^*}\right|} \le \log_2 \frac{n}{\left|[x_i]_{R_Q^*}\right|} = - \log_2 \frac{\left|[x_i]_{R_Q^*}\right|}{n}$$

Следовательно $H(P) \le H(Q)$.

Утверждение доказано.

Грубая энтропия, введенная в [3.25], используется для измерения степени детализации. Некоторые ученые также называют ее коэнтропией [3.1]. Аналогично определению 6 в работе [3.4], грубая энтропия HIS предлагается в следующем определении.

Определение 3.3.9 Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда грубая энтропия подсистемы (U, B) определяется следующим образом:

$$E_r(B) = - \sum_{i=1}^{m} \frac{|X_i|}{n} \log_2 \frac{1}{|X_i|}$$

где $U \setminus R_B^* = \{X_1, X_2, ..., X_m\}$

Пример 3.3.10 В соответствии с определением 3.2.1,

$$G_R(x_1) = \frac{0.47}{x_1} + \frac{0.22}{x_2} + \frac{0.45}{x_3} + \frac{0.66}{x_4}, G_R(x_2) = \frac{0.43}{x_1} + \frac{0.24}{x_2} + \frac{0.46}{x_3} + \frac{0.65}{x_4},$$

$$G_R(x_3) = \frac{0.55}{x_1} + \frac{0.77}{x_2} + \frac{0.58}{x_3} + \frac{0.33}{x_4}, G_R(x_4) = \frac{0.58}{x_1} + \frac{0.71}{x_2} + \frac{0.51}{x_3} + \frac{0.34}{x_4}$$

В соответствии с определением 3.2.10,

$$U \setminus R_B^* = \{X_1, X_2\},\ U \setminus R^1 = \{Y_1, Y_2, Y_3\},$$

где $X_1=\{x_1,x_2\}$, $X_2=\{x_3,x_4\}$, $Y_1=\{x_1,x_3\}$, $Y_2=\{x_2\}$, $Y_3=\{x_4\}$.

Отсюда,

$$E_r(R) = - \sum_{i=1}^{m} \frac{|X_i|}{n} \log_2 \frac{|X_i|}{n} = - \frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$$

Утверждение 3.3.11 Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда

$$E_r(B) = - \sum_{i=1}^{m} \frac{1}{n} \log_2 \frac{1}{\left|[x_i]_{R_B^*}\right|}$$

Доказательство. Обозначим

$$U \setminus R_B^* = \{X_1, X_2, ..., X_m\}$$

Предположим, что $X_i=\{x_{i1}, x_{i2},\ldots, x_{i,si}\}$, $|X_i|=s_i$. Тогда

$$\sum_{i=1}^{m} s_i = n, X_i = [x_{i1}]_{R_B^*} = [x_{i2}]_{R_B^*} = ... = [x_{i,s_i}]_{R_B^*}$$

Отсюда $\forall$ i,

$$|X_i| \log_2 \frac{1}{|X_i|} = s_i \log_2 \frac{1}{|X_i|} = \sum_{k=1}^{s_i} \frac{1}{|[x_{ik}]_{R_B^*}|}$$

Следовательно,

$$E_r(R_B^*) = - \sum_{i=1}^{m} \frac{|X_i|}{n} \log_2 \frac{1}{|X_i|} = - \frac{1}{n} \sum_{i=1}^{m} |X_i| \log_2 \frac{1}{|X_i|} =$$
$$= - \frac{1}{n} \sum_{i=1}^{m} \sum_{k=1}^{s_i} \log_2 \frac{1}{|[x_{ik}]_{R_B^*}|} = - \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{|[x_i]_{R_B^*}|} = - \sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{|[x_i]_{R_B^*}|}$$

Утверждение доказано.

Утверждение 3.3.12 Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда

$0 \le E_r(B) \le \log_2 n$

Боле того, если $R_B^* = \Delta$, то $E_r(B)$ достигает минимального значения 0; если $R_B^* = \omega$, то $E_r(B)$ достигает максимального значения $\log_2 n$.

Доказательство $\forall i$, $1 \le |[x_i]_{R_B^*}| \le n$.

(1) Тогда $\forall i$,

$$0 \le - \log_2 \frac{1}{|[x_i]_{R_B^*}|} = \log_2 |[x_i]_{R_B^*}| \le \log_2 n$$

Отсюда $\forall i$,

$$0 \le - \sum_{i=1}^{n} \log_2 \frac{1}{|[x_i]_{R_B^*}|} \le n \log_2 n$$

Из Утверждения 3.3.11,

$0 \le E_r(B) \le \log_2 n$

(2) Предположим, что $R_B^* = \Delta$. Тогда

$R_B^* = \{(x_1, x_1), (x_2, x_2), ..., (x_n, x_n)\}$. Отсюда $U \setminus R_B^* = \{\{x_1\}, \{x_2\}, ..., \{x_n\}\}$.

Поэтому $\forall i$, $\left|[x_i]_{R_B^*}\right| = 1$. Следовательно, $E_r(B) = 0$.

(3) Предположим, что $R_B^* = \omega$. Тогда $R_B^* = U \times U$. Отсюда $U \setminus R_B^* = \{x_1, x_2, ..., x_n\}$. Поэтому $\forall i$, $\left|[x_i]_{R_B^*}\right| = n$. Следовательно, $E_r(B) = \log_2 n$.

Утверждение доказано.

Утверждение 3.3.12 дает диапазон значений грубой энтропии для HIS.

3.3.3. Объем информации в HIS

Аналогично определению 10 в работе [3.14], объем информации в HIS предлагается в следующем определении.

Определение 3.3.13 Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда объем информации подсистемы (U, B) определяется следующим образом:

$$E(B) = \sum_{i=1}^{m} \frac{|X_i|}{n} \frac{|U - X_i|}{n}$$

где $U \setminus R_B^* = \{X_1, X_2, ..., X_n\}$

Утверждение 3.3.14. Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда

$$E(B) = \sum_{i=1}^{n} \frac{1}{n} \left(1 - \frac{\left|[x_i]_{R_B^*}\right|}{n}\right)$$

Доказательство. Обозначим

$U \setminus R_B^* = \{X_1, X_2, ..., X_n\}$

Предположим, что $X_i = \{x_{i1}, x_{i2}, \ldots, x_{i,si}\}$, тогда $|X_i| = s_i$. Отсюда

$$\sum_{i=1}^{m} s_i = n, X_i = [x_{i1}]_{R_B^*} = [x_{i2}]_{R_B^*} = ... = [x_{i,s_i}]_{R_B^*}$$

Отсюда следует, что $X_i = [x_{i1}]_{R_B^*} = [x_{i2}]_{R_B^*} = ... = [x_{is_i}]_{R_B^*} = s_i$

Поскольку $\{X_1, X_2, \ldots, X_m\}$ является разделом на U, $\forall i$, имеем

$$U - X_i = \left(\bigcup_{k=1}^{i-1} X_k\right) \cup \left(\bigcup_{k=i+1}^{m} X_k\right)$$

Тогда $|U - X_i| = \sum_{k=1}^{i-1} |X_k| + \sum_{k=i+1}^{m} |X_k| = |U| - |X_i| = n - |X_i|$

Отсюда $\forall i$, $|X_i||U - X_i| = s_i (n - |X_i|) \sum_{k=1}^{s_i} \left(n - \left|[x_{ik}]_{R_B^*}\right|\right)$

Поэтому

$$E(B) = \sum_{i=1}^{m} \frac{|X_i|}{n} \frac{|U - X_i|}{n} \log_2 \frac{1}{|X_i|} = \sum_{i=1}^{m} \sum_{k=1}^{s_i} \frac{n - \left|[x_{ik}]_{R_B^*}\right|}{n^2} =$$

$$= \sum_{i=1}^{n} \frac{n - \left|[x_{ik}]_{R_B^*}\right|}{n^2} = \sum_{i=1}^{n} \frac{1}{n} \frac{1 - \left|[x_{ik}]_{R_B^*}\right|}{n}$$

Утверждение доказано.

3.3.4. Особенности устранения неопределенности в HIS

Утверждение 3.3.15. Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда G(B) + E(B) = 1.

Доказательство. Обозначим

$$U \setminus R_B^* = \{X_1, X_2, ..., X_n\}$$

Пусть $X_i = \{x_{i1}, x_{i2}, \ldots, x_{i,si}\}$, тогда $|X_i| = s_i$. Отсюда $\sum_{i=1}^{m} s_i = n$.

Поскольку $\{X_1, X_2, \ldots, X_m\}$ является разделом на U, $\forall i$, имеем

$$U - X_i = \left(\bigcup_{k=1}^{i-1} X_k\right) \cup \left(\bigcup_{k=i+1}^{m} X_k\right)$$

Тогда $|U - X_i| = \sum_{k=1}^{i-1} |X_k| + \sum_{k=i+1}^{m} |X_k| = |U| - |X_i| = n - |X_i|$

Отсюда

$$E(B) = \sum_{i=1}^{m} \frac{|X_i|}{n} \frac{|U - X_i|}{n} = \sum_{i=1}^{m} \frac{s_i(n - s_i)}{n^2} =$$

$$= \sum_{i=1}^{m} \frac{s_i}{n} - \sum_{i=1}^{m} \frac{s_i^2}{n^2} = 1 - \sum_{i=1}^{n} \frac{|X^i|^2}{n^2} = 1 - G(B)$$

Следовательно, G(B) + E(B) = 1.

Утверждение доказано.

Следствие 3.3.16 (Ограниченность) Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда $0 \le E(B) \le 1 - \frac{1}{n}$

Доказательство. Из Утверждения 3.3.3, $\frac{1}{n} \le G(B) \le 1$

Из Утверждения 3.3.15, E(B)=1−G(B).

Следовательно, $0 \le E(B) \le 1 - \frac{1}{n}$.

Следствие доказано.

Следствие 3.3.16 указывает на диапазон значений количества информации для неопределенности HIS.

Следствие 3.3.17 (Монотонность) (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда $E(P) \le E(Q)$.

Доказательство. Доказательство может быть проведено с использованием утверждений 3.3.4 и 3.3.15.

Утверждение 3.3.18 (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда

$E_r(B)+H(B)=\log_2 n$.

Доказательство. Обозначим $U \setminus R_B^* = \{X_1, X_2, ..., X_n\}$

Пусть $X_i=\{x_{i1}, x_{i2},\ldots, x_{i,si}\}$, тогда $|X_i|=s_i$. So $\sum_{i=1}^{m} s_i = n$.

Отсюда

$$E_r(B) + H(B) = -\sum_{i=1}^{m} \frac{|X_i|}{n} \log_2 \frac{1}{|X_i|} - \sum_{i=1}^{m} \frac{|X_i|}{n} \log_2 \frac{|X_i|}{n} =$$

$$= -\sum_{i=1}^{m} \frac{s_i}{n} \log_2 \frac{1}{s_i} - \sum_{i=1}^{m} \frac{s_i}{n} \log_2 \frac{s_i}{n} = -\sum_{i=1}^{m} \frac{s_i}{n} (\log_2 1 - \log_2 s_i + \log_2 s_i - \log_2 n) =$$

$$-\sum_{i=1}^{m} \frac{s_i}{n} (\log_2 1 - \log_2 n) = -\sum_{i=1}^{m} \frac{s_i}{n} (-\log_2 n) = \log_2 n$$

Утверждение доказано.

Следствие 3.3.19 (Ограниченность) Пусть (U, A) - это HIS. Задано, что $B \subseteq A$. Тогда $0 \le H(B) \le \log_2 n$

Доказательство. Из следствия 3.3.16, $0 \le Er(B) \le \log_2 n$.

Из Утверждения 3.3.18, $H(B) = \log_2 n - E_r(B)$.

Отсюда $0 \le H(B) \le \log_2 n$.

Следствие доказано.

Следствие 3.3.19 демонстрирует диапазон значений информационной энтропии для неопределенности HIS.

Слдствие 3.3.20. (Монотонность) Пусть (U, A) - это HIS. Если $P \subseteq Q \subseteq A$, то $E_r(Q) \le E_r(P)$.

Доказательство. Доказательство может быть проведено из теорем 3.3.8 и 3.3.18.

3.4. Численные эксперименты и анализ эффективности

Чтобы оценить эффективность предлагаемых мер по устранению неопределенности в HIS, проведем численный эксперимент и анализ эффективности с трех точек зрения.

3.4.1. Численный эксперимент

Был проведен следующий численный эксперимент с тремя наборами данных, полученными из UCI (Хранилища баз данных машинного обучения), который показан в табл. 3.1, и проведено сравнение четырех инструментов для измерения неопределенности HIS.

Мы выбрали следующие наборы данных из базы данных UCI. В таблице 3.1 представлена информация для этих наборов данных.

Таблица 3.1

4 набора данных из UCI

N	Набор данных	Образцы	Масштаби-рованный	Порядковый	Номинальный	Классы
1	Набор 1	368	5	2	15	2
2	Набор 2	583	5	4	1	2
3	Набор 3	155	2	4	13	2
4	Набор 4	270	1	6	6	2

Набор 1 может выражать HIS (U, A) с помощью |U|=368, |A|=23. Обозначим $A_i=\{a_1,\ldots, a_i\}$ (i = 1,..., 23) . Тогда для каждого i существует отношение эквивалентности, индуцированное A_i в Наборе 1. Четыре набора мер в Наборе 1 определены следующим образом

$$X_G(He)=\left\{G\left(R^*_{A_1}\right),...,G\left(R^*_{A_{23}}\right)\right\}, X_E(He)=\left\{E\left(R^*_{A_1}\right),...,E\left(R^*_{A_{23}}\right)\right\},$$
$$X_{E_r}(He)=\left\{E_r\left(R^*_{A_1}\right),...,E_r\left(R^*_{A_{23}}\right)\right\}, X_H(He)=\left\{H\left(R^*_{A_1}\right),...,H\left(R^*_{A_{23}}\right)\right\}$$

Набор 2 может выражать HIS (V, B) с помощью |V|=583, |B|=1. Обозначим $B_i=\{b_1,\ldots,b_i\}$ (i=1,...,11) . Тогда для каждого i существует отношение эквивалентности, индуцированное B_i в Наборе 2. Четыре набора мер в Наборе 2 определены следующим образом

$$X_G(ID)=\left\{G\left(R^*_{B_1}\right),...,G\left(R^*_{B_{11}}\right)\right\}, X_E(ID)=\left\{E\left(R^*_{B_1}\right),...,E\left(R^*_{B_{11}}\right)\right\},$$
$$X_{E_r}(ID)=\left\{E_r\left(R^*_{B_1}\right),...,E_r\left(R^*_{B_{11}}\right)\right\}, X_H(ID)=\left\{H\left(R^*_{B_1}\right),...,H\left(R^*_{B_{11}}\right)\right\}$$

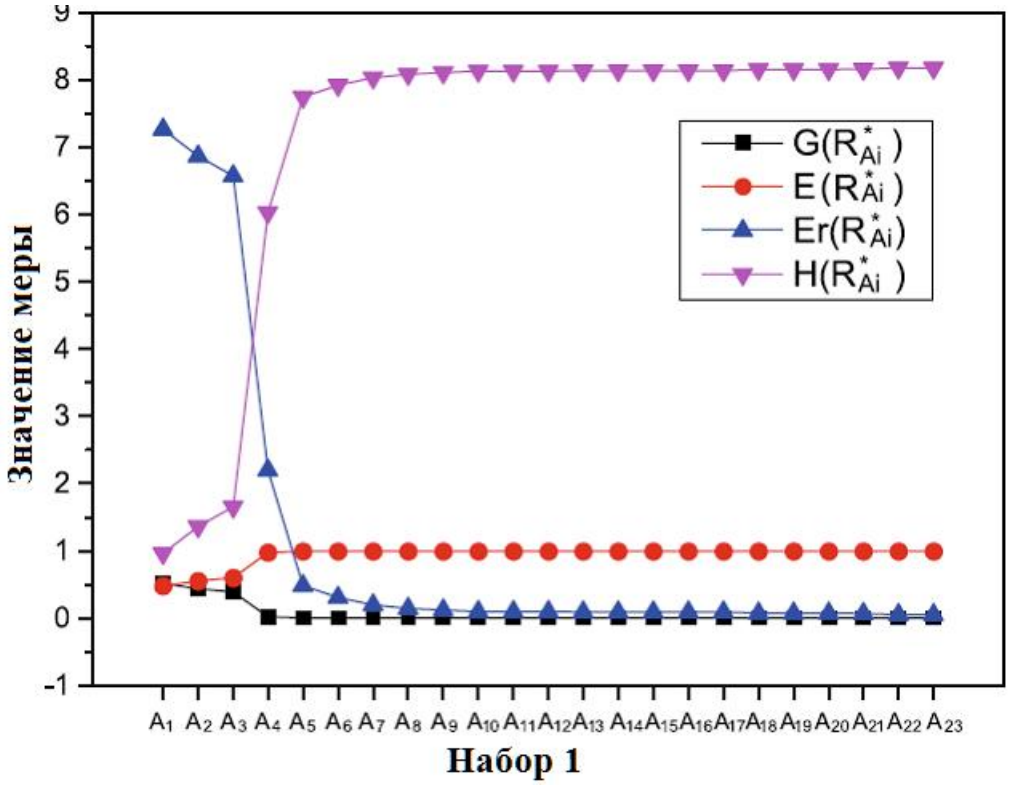
Набор 3 может выражать HIS (W, C) с помощью |W|=155, |C|= 20. Обозначим $C_i=\{c_1,..., c_i\}$ (i = 1,..., 20). Тогда для каждого i существует отношение эквивалентности, индуцированное C_i в Наборе 3. Четыре набора мер в Наборе 3 определены следующим образом:

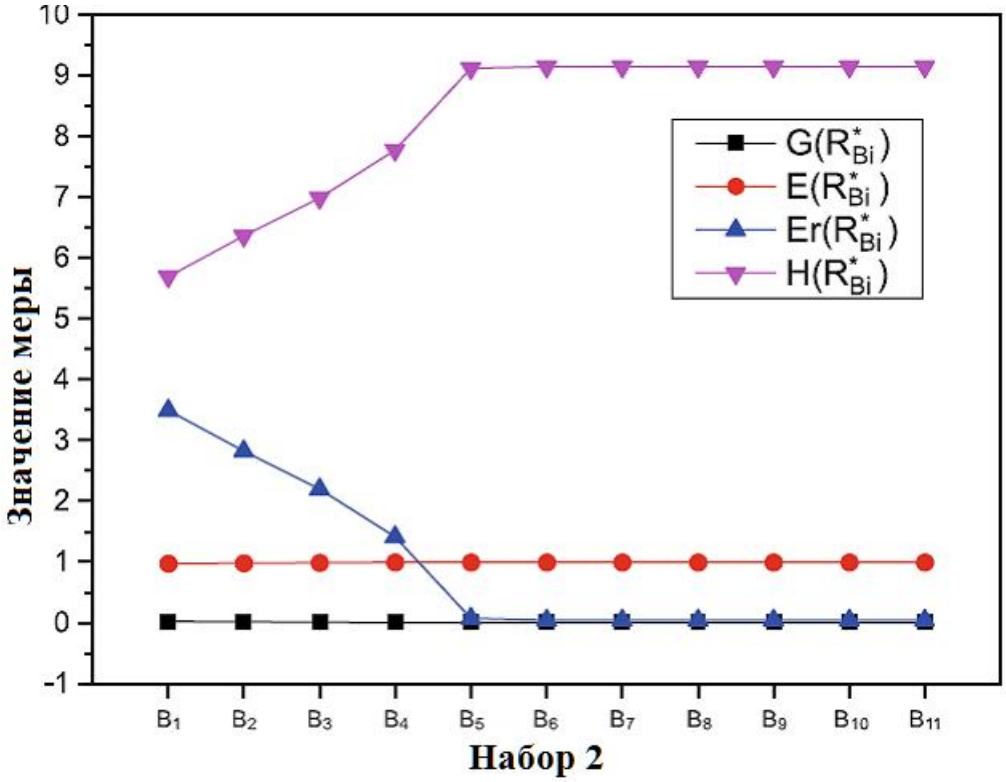
$$X_G(Hs)=\left\{G\left(R^*_{C_1}\right),...,G\left(R^*_{C_{20}}\right)\right\}, X_E(Hs)=\left\{E\left(R^*_{C_1}\right),...,E\left(R^*_{C_{20}}\right)\right\},$$
$$X_{E_r}(Hs)=\left\{E_r\left(R^*_{C_1}\right),...,E_r\left(R^*_{C_{20}}\right)\right\}, X_H(Hs)=\left\{H\left(R^*_{C_1}\right),...,H\left(R^*_{C_{20}}\right)\right\}$$

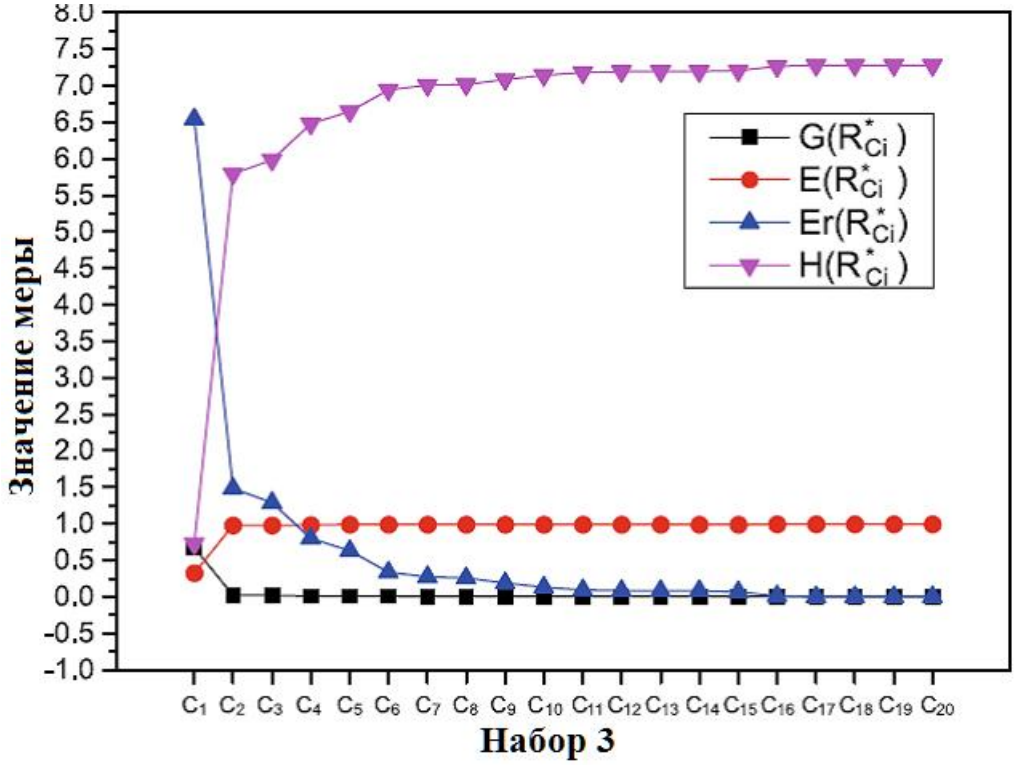
Набор 4 может выражать HIS (X, D) с |X|=270, |D|=14. Тогда для каждого $d \in D$, R_d - нечеткое соотношение, вызванное Набором 4. Обозначим $D_i=\{d_1,\ldots, d_i\}$ (i=1,...,14) . Тогда для каждого i, $R^*_{D_i}$ будет отношением эквивалентности, индуцированное D_i в Наборе 4. Четыре набора мер на наборе 3 определены следующим образом:

$$X_G(Ht)=\left\{G\left(R^*_{D_1}\right),\ldots,G\left(R^*_{D_{14}}\right)\right\}, X_E(Ht)=\left\{E\left(R^*_{D_1}\right),\ldots,E\left(R^*_{D_{14}}\right)\right\},$$
$$X_{E_r}(Ht)=\left\{E_r\left(R^*_{D_1}\right),\ldots,E_r\left(R^*_{D_{14}}\right)\right\}, X_H(Ht)=\left\{H\left(R^*_{D_1}\right),\ldots,H\left(R^*_{D_{14}}\right)\right\}$$

Результаты эксперимента показаны на рисунке 3.1. Можно предположить, что грануляция информации G и грубая энтропия E_r монотонно уменьшаются по мере увеличения отношений эквивалентности. В то же время количество информации E и информационная энтропия H монотонно увеличиваются с увеличением отношений эквивалентности. Это означает, что неопределенность нечеткого отношения уменьшается по мере увеличения отношений эквивалентности.




Значение меры
G(R*Bi)
E(R*Bi)
Er(R*Bi)
H(R*Bi)
B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11
Набор 2


Значение меры
G(R*Ci)
E(R*Ci)
Er(R*Ci)
H(R*Ci)
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14 C15 C16 C17 C18 C19 C20
Набор 3

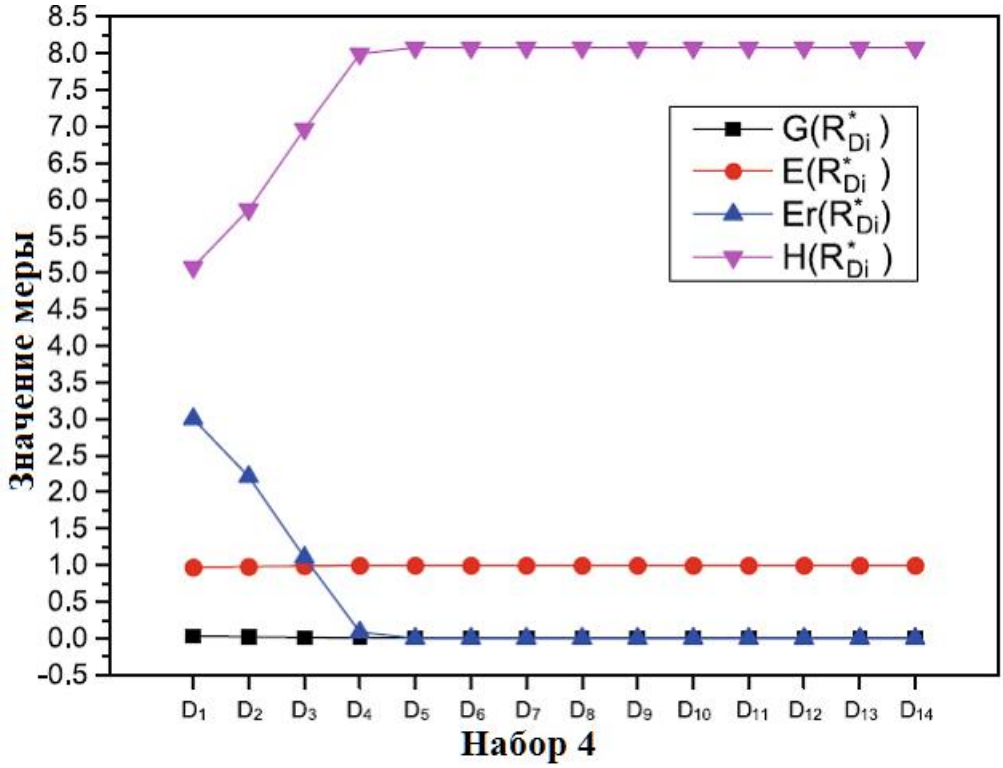


Рис. 3.1. CV-значения четырех наборов измерений

Таким образом, грануляция информации G, грубая энтропия E_r, объем информации E и информационная энтропия H могут быть применены для измерения неопределенности HIS.

3.4.2. Дисперсионный анализ

В реальном статистическом исследовании часто изучается степень дисперсии набора данных. Величина, используемая для измерения степени дисперсии набора данных, называется мерой разности. Общие показатели разности включают диапазон, четырехточечную разницу, среднюю разницу, стандартное отклонение, коэффициент стандартного отклонения и так далее. В данной работе для анализа эффективности предлагаемых мер применяется коэффициент стандартного отклонения.

Предположим, что $X=\{x_1,\ldots, x_n\}$ - это набор данных. Тогда среднее арифметическое значение, стандартное отклонение и коэффициент

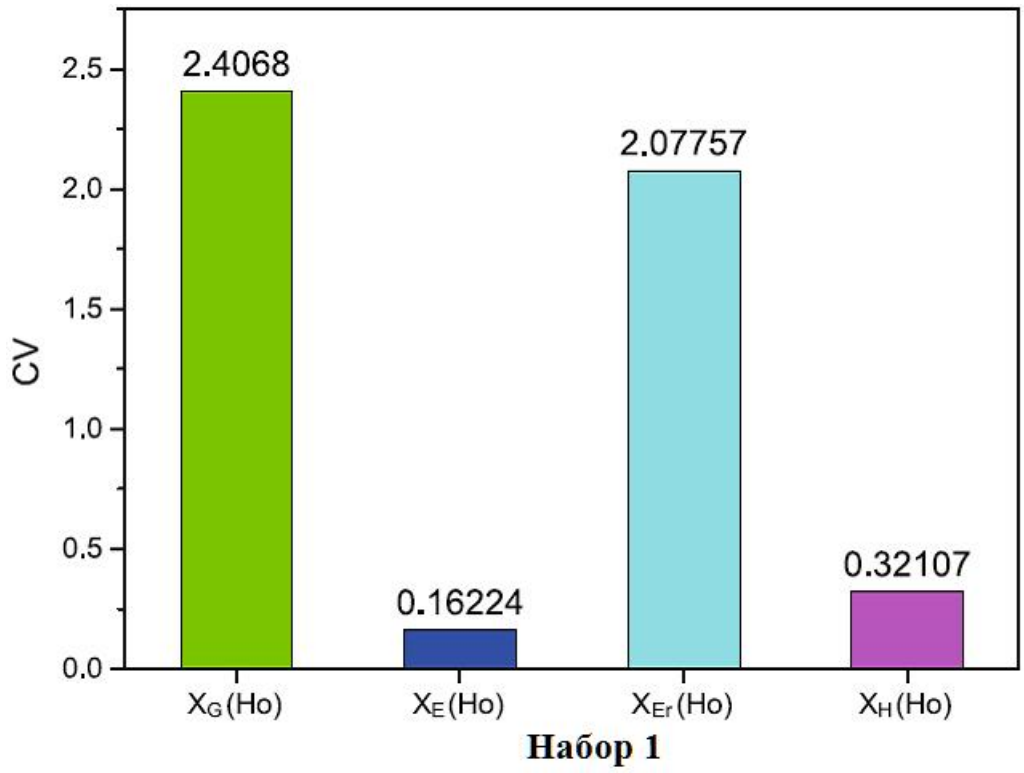
стандартного отклонения X, обозначаемые $\bar{x}$, σ(X) и CV(X), соответственно, определяются следующим образом:

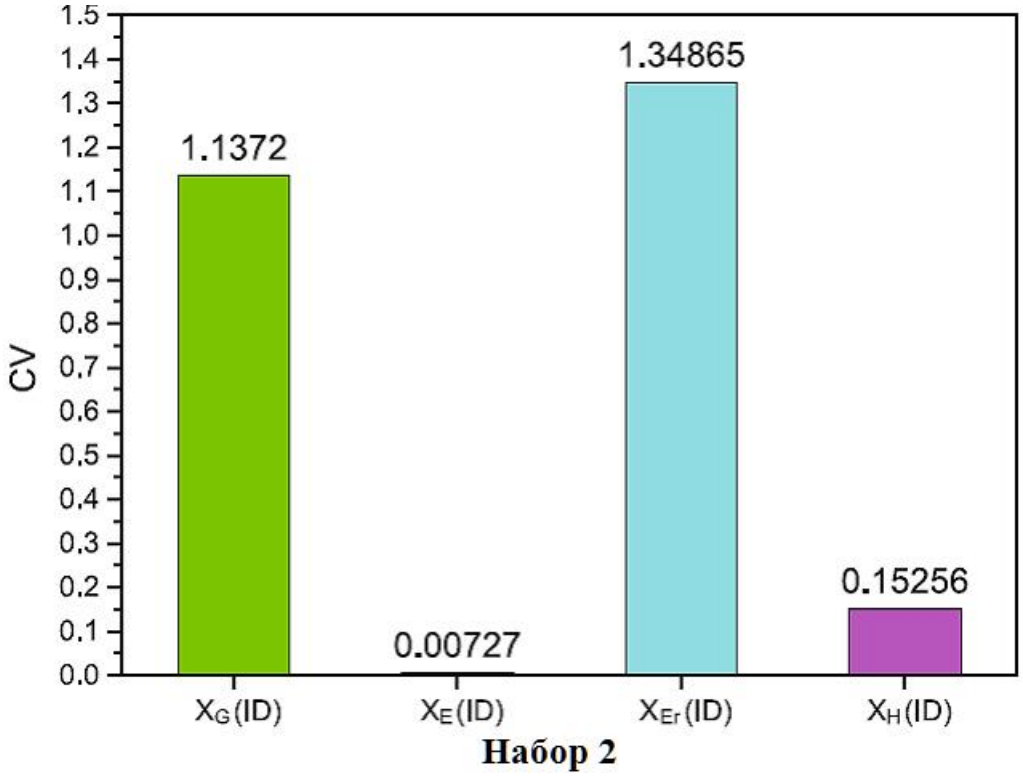
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, s(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \,, CV(X) = \frac{s(X)}{\bar{x}}.$$

Продолжая описанный выше эксперимент, сравниваются CV-значения четырех наборов измерений. Результаты показаны на рис. 3.2.

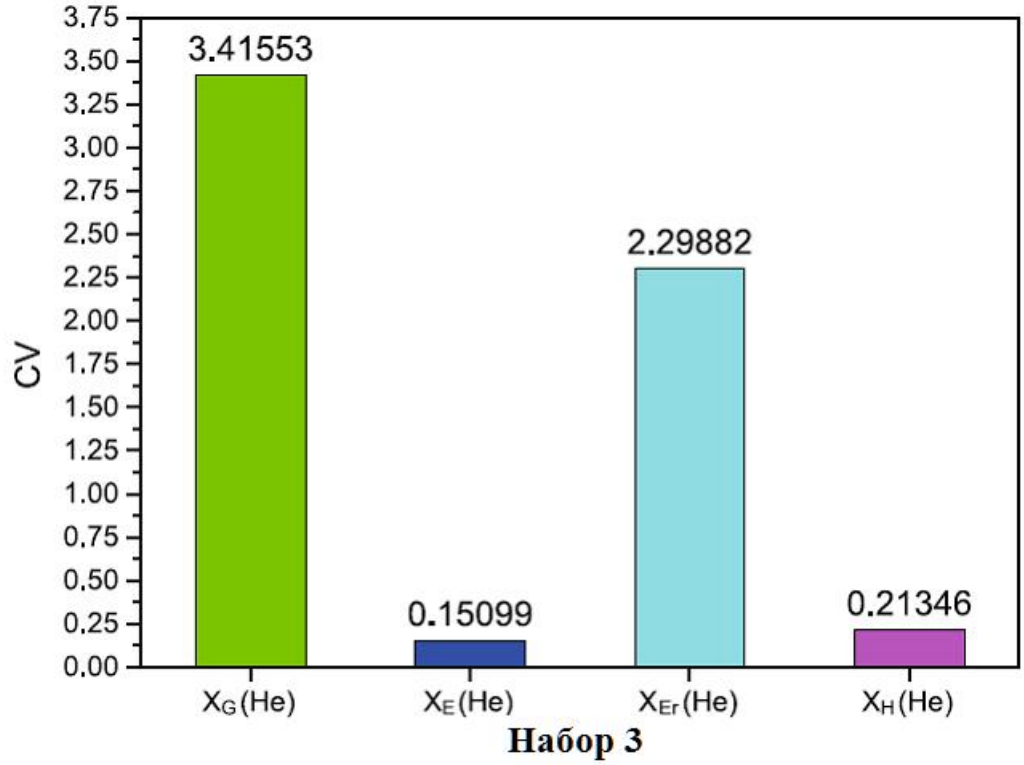
Степень дисперсии E минимальна, если в качестве тестового набора выбран Набор 1. Аналогично, степень дисперсии E минимальна, когда в качестве тестовых наборов выбраны другие наборы данных (например, Набор 2, Набор 3 и Набор 4).

Таким образом, объем информации E значительно лучше подходит для измерения неопределенности HIS.





Набор 2



Набор 3

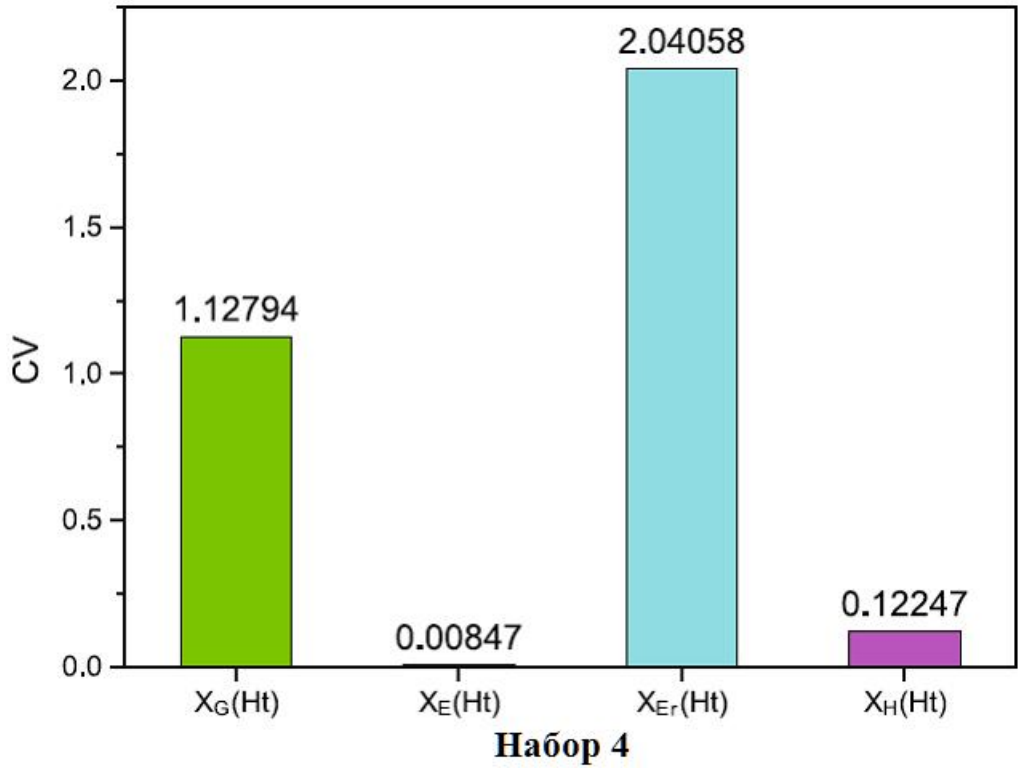


Рис. 3.2. CV-значения четырех наборов измерений

3.4.3. Корреляционный анализ

В статистике коэффициент корреляции Пирсона является мерой силы линейной корреляции между двумя наборами данных.

Предположим, что $X=\{x_1,\ldots, x_n\}$ и $Y=\{y_1,\ldots, y_n\}$ являются двумя наборами данных. Коэффициент корреляции Пирсона между X и Y, обозначаемый как r(X, Y), определяется как

$$r(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

где $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

Легко проверить, что

$$r(X,Y)=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

Очвидно, $-1\le r(X,Y)\le 1$.

Корреляция между X и Y может быть получена в соответствии с табл. 3.2.

Продолжая описанный выше эксперимент, сравниваются r-значения четырех наборов измерений. Результаты приведены в таблицах 3.3, 3.4, 3.5 и 3.6. Из таблиц 3.3, 3.4, 3.5 и 3.6 можно сделать следующие выводы (см. таблицы 3.7, 3.8, 3.9 и 3.10).

Таблица 3.2

Соответствующая корреляция между X и Y

r(X, Y)	Корреляция между X и Y	Сокращение
r(X, Y)=1	Полностью положительная корреляция	CPC
0.7≤ r(X, Y)<1	Положительная корреляция	HPC
0.4≤ r(X, Y)<0.7	Умеренная положительная корреляция	MPC
0<r(X, Y)<0.4	Слабая положительная корреляция	LPC
r(X, Y)=0	Отсутствие корреляции	NC
−0.4<r(X, Y)<0	Слабая отрицательная корреляция	LNC
−0.7≤r(X, Y)<−0.4	Умеренная отрицательная корреляция	MNC
−1≤r(X, Y)<−0.7	Отрицательная корреляция	HNC
r(X, Y)=−1	Полностью отрицательная корреляция	CNC

Таблица 3.3

r-значения для 16 пар из 4 наборов измерений на Наборе 1

r	X_G(He)	X_E(He)	X_{E_r} (He)	X_H(He)
X_G(He)	1			
X_E(He)	−1	1		
X_{E_r} (He)	0.9832	−0.9832	1	
X_H(He)	−0.9832	0.9832	−1	1

Таблица 3.4

r-значения для 16 пар из 4 наборов измерений на Наборе 2

r	X_G(ID)	X_E(ID)	X_{E_r} (ID)	X_H(ID)
X_G(ID)	1			
X_E(ID)	−1	1		
X_{E_r} (ID)	0.9762	−0.9762	1	
X_H(ID)	−0.9762	0.9762	−1	1

Таблица 3.5

r-значения для 16 пар из 4 наборов измерений на Наборе 3

r	X_G(HS)	X_E(HS)	X_{E_r} (HS)	X_H(HS)
X_G(HS)	1			
X_E(HS)	−1	1		
X_{E_r} (HS)	0.9643	−0.9643	1	
X_H(HS)	−0.9643	0.9643	−1	1

Таблица 3.6

r-значения для 16 пар из 4 наборов измерений на Наборе 4

r	X_G(HT)	X_E(HT)	X_{E_r} (HT)	X_H(HT)
X_G(HT)	1			
X_E(HT)	−1	1		
X_{E_r} (HT)	0.9843	−0.9843	1	
X_H(HT)	−0.9843	0.9843	−1	1

Таблица 3.7

Корреляция между двумя показателями на Наборе 1

	G	E	E_r	H
G	CPC			
E	CNC	CPC		
E_r	HPC	HNC	CPC	
H	HNC	HPC	CNC	CPC

Таблица 3.8

Корреляция между двумя показателями на Наборе 2

	G	E	E_r	H
G	CPC			
E	CNC	CPC		
E_r	HPC	HNC	CPC	
H	HNC	HPC	CNC	CPC

Таблица 3.9

Корреляция между двумя показателями на Наборе 3

	G	E	E_r	H
G	CPC			
E	CNC	CPC		
E_r	HPC	HNC	CPC	
H	HNC	HPC	CNC	CPC

Таблица 3.10

Корреляция между двумя показателями на Наборе 4

	G	E	E_r	H
G	CPC			
E	CNC	CPC		
E_r	HPC	HNC	CPC	
H	HNC	HPC	CNC	CPC

3.4.4. Тест Фридмана и тест Бонферрони–Данна

Для дальнейшей оценки эффективности предлагаемых мер для HIS в этом подразделе приведены тесты Фридмана [3.6] и Бонферрони–Дана [3.3]. Тест Фридмана - это статистический тест, который использует ранжирование алгоритмов. Статистика Фридмана определяется как

$$c_F^2 = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right)$$

где k - количество алгоритмов, N - количество наборов данных, r_i - средний рейтинг i-го алгоритма. Когда k и N достаточно велики, статистика Фридмана соответствует распределению χ^2 с k-1 степенями свободы. Однако такой критерий Фридмана слишком сложен и обычно заменяется следующей статистикой

$$F_F = \frac{(N-1)c_F^2}{N(k-1) - c_F^2}$$

Статистика FF соответствует распределению Фишера с k−1 и (k−1)(N−1) степенями свободы. Если статистический показатель FF превышает критическое значение $F_\alpha(k-1, N-1)$, это означает, что нулевая

гипотеза отклоняется в соответствии с критерием Фридмана. Тест Бонферрони–Данна может быть использован для дальнейшего изучения того, какой алгоритм лучше с точки зрения статистики. Если средний уровень расстояния превышает критическое значение CD_α, то производительность двух алгоритмов будет существенно отличаться. Критическое значение CD_α обозначается как

$$CD_a = q_a \sqrt{\frac{k(k+1)}{6N}}$$

где q_α – критическое табличное значение для теста, а α - уровень значимости теста Бонферрони-Данна.

Ниже мы рассматриваем четыре показателя неопределенности HIS как четыре алгоритма и демонстрируем статистическую значимость с помощью теста Фридмана и теста Бонферрони–Данна.

(1) На основе CV-значений четырех показателей на рис. 3.2 мы приводим ранжирование четырех показателей с четырьмя наборами данных соответственно (см. таблицу 3.11).

Таблица 3.11

Ранжирование CV-значений четырех показателей по четырем наборам данных

Датасеты	G	E	E_r	H
Набор 1	4	1	3	2
Набор 2	3	1	2	4
Набор 3	4	1	3	2
Набор 4	3	1	4	2
Среднее	3.5	1	3.0	2.5

(2) Мы проводим тест Фридмана, чтобы выяснить, существенно ли различаются значения CV для четырех показателей, соответственно. Для четырех показателей и четырех наборов данных FF соответствует распределению с 3 и 9 степенями свободы. Тогда можно легко вычислить статистику F_F = 7. Обратите внимание, что критическое значение

распределения Фишера $F_{0.05}(3, 9)$ равно 3,863. Очевидно, что 7>3,863. Это означает, что при значимом уровне α=0,05 нам необходимо отклонить нулевую гипотезу, что означает, что CV-значения четырех показателей существенно различаются.

(3) Чтобы дополнительно продемонстрировать существенную разницу между четырьмя показателями, вводится тест Бонферрони–Данна. Для значимого уровня α=0,05 мы можем рассчитать соответствующее критическое расстояние CD_α = 2,569. На рис. 3.3 показаны результаты тестирования с α = 0,05 для четырех наборов данных, где точками обозначены средние значения четырех показателей, а отрезки линии обозначают область применения CD_α. Если два отрезка линии частично перекрываются по оси y, то существенной разницы между этими двумя показателями нет.

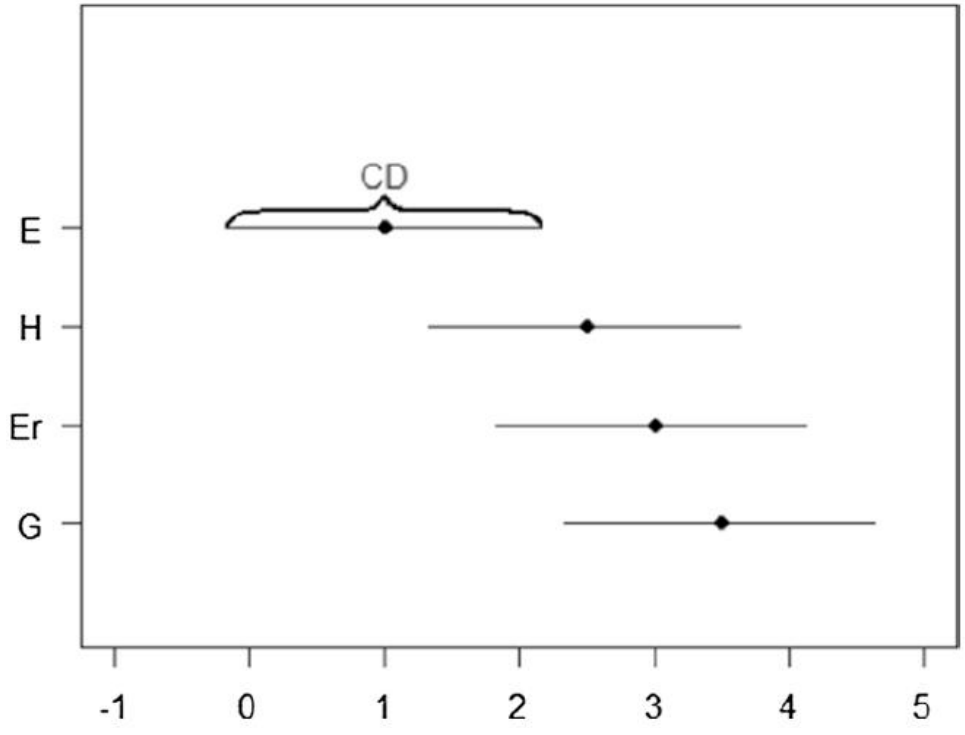


Рис. 3.3. Тест Бонферрони–Данна, основанный на CV-значениях четырех измерений

(4) На рис. 3.3 получены следующие результаты:

(а) Показатели E статистически лучше, чем показатели H, G или E_r;

((б) Статистических различий между H, G и E_r нет.

3.5. Пример уменьшения атрибутов HIS

В этом разделе мы приводим пример применения предложенных мер по уменьшению атрибутов в HIS.

3.5.1 Уменьшение атрибутов в HIS

Определение 3.5.1 Пусть (U, A) - HIS. Дано, что $B \subseteq A$. Тогда B называется согласованным подмножеством A, если $R^*_B = R^*_A$.

Определение 3.5.2 Пусть (U, A) - HIS. Дано, что $a \in B \subseteq A$. Тогда a называется независимым в B, если $R^*_B \neq R^*_{B-\{a\}}$.

Определение 3.5.3 Пусть (U, A) - HIS. Дано, что $B \subseteq A$. Тогда B называется независимым подмножеством A, если для любого $a \in B$, a независимо в B.

Определение 3.5.4 Пусть (U, A) - HIS. Дано, что $B \subseteq A$. Тогда B называется восстановлением A, если B является одновременно последовательным и независимым.

В работе множество всех координационных подмножеств (соответственно, все редукции знаний) A обозначается как co(A) (соответственно, red(A)).

Очевидно,

$B \in red(A)$ Û $B \in co(A)$ и $\forall C \subset B$, $C \notin co(A)$.

Утверждение 3.5.5 Пусть (U, A) - HIS. Дано, что $B \subseteq A$. Тогда следующие условия эквивалентны:

(1) $B \in co(A)$;

(2) G(B)=G(A);

(3) H(B)=H(A);

(4) $E_r(B)=E_r(A)$;

(5) E(B)=E(A).

Доказательство

(1) ⇒ (2). Очевидно.

(2) ⇒ (1). Предположим, что G(B)= G(A). Тогда из Утверждения 3.2, имеем $\frac{1}{n^2}\sum_{i=1}^{n}\left|[x_i]_{R_B^*}\right|=\frac{1}{n^2}\sum_{i=1}^{n}\left|[x_i]_{R_A^*}\right|$

Тогда $\sum_{i=1}^{n}\left|[x_i]_{R_B^*}\right|-\sum_{i=1}^{n}\left|[x_i]_{R_A^*}\right|=0$

Заметим, что $R_A^* \subseteq R_B^*$. Тогда $\forall i,\ [x_i]_{R_A^*} \subseteq [x_i]_{R_B^*}$. Следовательно,

$\forall i, |[x_i]_{R_B^*}-[x_i]_{R_A^*}| \geq 0$

Так что, $\forall i$, $|[x_i]_{R_B^*}-[x_i]_{R_A^*}|=0$. Следовательно, $\forall i$, $[x_i]_{R_B^*}=[x_i]_{R_A^*}$. Тогда, $R_B^*=R_A^*$. Поэтому $B \in co(A)$.

(2) ⇔ (5). Может быть получено из Утверждения 3.3.15.

(1) ⇒ (3). Очевидно.

(3) ⇒ (1). Пусть H(B)=H(A) . Тогда из Утверждения 3.3.11, имеем

$$-\sum_{i=1}^{n}\frac{1}{n}\log_2\frac{\left|[x_i]_{R_B^*}\right|}{n}=-\sum_{i=1}^{n}\frac{1}{n}\log_2\frac{\left|[x_i]_{R_A^*}\right|}{n}$$

Тогда $\sum_{i=1}^{n}\log_2\frac{\left|[x_i]_{R_B^*}\right|}{\left|[x_i]_{R_A^*}\right|}=0.$

Заметим, что $R_B^* \subseteq R_A^*$. Отсюда $\forall i$, $\left|[x_i]_{R_B^*}\right|=\left|[x_i]_{R_A^*}\right|$

Поэтому $R_B^*=R_A^*$. Следовательно $B \in co(A)$.

(3) ⇔ (4). Может быть получено из Утверждения 3.3.18.

Утверждение доказано.

Следствие 3.5.6 Пусть (U, A) - HIS. Дано, что $B \subseteq A$. Тогда следующие условия эквивалентны:

(1) $B \in red(A)$;

(2) $G(B)=G(A)$ и $\forall a \in B$, $G(B-\{a\}) \neq G(A)$;

(3) $H(B)=H(A)$ и $\forall a \in B$, $H(B-\{a\}) \neq H(A)$;

(4) $E_r(B)=E_r(A)$ и $\forall a \in B$, $E_r(B-\{a\}) \neq Er(A)$;

(5) $E(B)=E(A)$ и $\forall a \in B$, $E(B-\{a\}) \neq E(A)$.

Доказательство Может быть получено из Утверждения 3.5.5.

3.5.2. Алгоритмы редукции в HIS

Ниже мы рассмотрим алгоритмы редукции, основанные на измерении неопределенности для разнородных данных. Согласно Утверждениям 3.3.15 и 3.3.18, мы имеем

$G(B) + E(B) = 1$, $E_r(B) + H(B) = \log_2 n$

где (U, A) - HIS и $B \subseteq A$. Тогда нам остается только рассмотреть алгоритмы редукции, основанные на грануляции информации и информационной энтропии соответственно.

Алгоритмы редукции HIS, основанные на грануляции информации и информационной энтропии, приведены ниже.

Алгоритм 3.1 Алгоритм редукции в HIS, основанный на грануляции информации

input: A HIS (U,A).
output: A reduct B
begin
B=A
start=1
Compute G(A)
while start **do**
for каждого атрибута aÎ B **do**
if $G(B-\{a\})=G(B)$
then

```
                    B=B−{a}
                else
                    start = 0
                end if
            end for
        end while
        return B
end
```

Алгоритм 3.2 Алгоритм редукции в HIS, основанный на информационной энтропии

```
input: A HIS (U,A).
output: A reduct B.
begin
        B=0
        start = 1
        Compute H(A)
        while start do
            for каждого атрибута aÎ A-B do
                if H(BÈ{a})¹ H(A)
                then
                    B=BÈ{a}
                else
                    start = 0
                end if
            end for
        end while
        return B
end
```

Алгоритм 3.1 применяет детализацию информации для определения атрибута, который отбрасывается из текущего выбранного согласованного набора в каждом цикле. Этот алгоритм завершается, когда редукция любых оставшихся атрибутов не приводит к уменьшению функции оценки.

При размерности |A|, временной сложности для вычисления грануляции информации, равной |A|, наихудшее время поиска для

сокращения приведет к |A|(|A|+1)/2 оценкам функции оценки. Общая временная сложность алгоритма 3.1 равна $O(|A|^2)$.

Алгоритм 3.2 использует информационную энтропию для получения атрибутов, которые добавляются к текущему выбранному согласованному набору в каждом цикле. Этот алгоритм завершается, когда добавление любого оставшегося атрибута не приводит к уменьшению вычислительной функции. При размерности |A| временная сложность вычисления информационной энтропии равна |A|, наихудшее время поиска для сокращения приведет к |A|(|A|+1)/2 вычислениям расчетной функции. Общая временная сложность алгоритма 2 равна $O(|A|^2)$.

3.5.3 Кластер-анализ

В этом подразделе рассматривается кластеризация на основе редуцированной HIS.

Предложенные алгоритмы сокращения тестируются на пяти разнородных наборах данных из UCI, описанных в таблице 3.12. Каждый из этих пяти наборов данных может быть представлен как HIS без атрибутов принятия решения. Разработан тест, чтобы выявить важные характеристики. Благодаря кластерному анализу метод является эффективным.

На основе алгоритма 3.1 разработана стратегия эвристического поиска, позволяющая найти сокращение из заданных пяти наборов данных. Процесс начинается со всего набора атрибутов, а затем удаляет атрибут один за другим до тех пор, пока значения степени детализации информации полученного подмножества атрибутов и всего набора атрибутов не станут одинаковыми.

На основе алгоритма 3.2 также разработана стратегия эвристического поиска, позволяющая найти сокращение из заданных пяти наборов данных. Процесс начинается с пустого набора, а затем атрибут

добавляется один за другим, пока значения информационной энтропии полученного подмножества атрибутов и всего набора атрибутов не сравняются.

Ниже мы используем только алгоритм 3.1 для получения редукции HIS. Алгоритм 3.1 - это эвристический поиск, один поиск может привести не более чем к одному уменьшению. Мы выполнили 19 поисков и каждый раз получали по одному уменьшению для каждого набора данных, указанного в таблице 3.12, соответственно. Полученные результаты приведены в таблице 3.13, на рисунке 3.4 показан пример редукции для кластеризации KMeans для Набора 5 с двумя кластерами.

Эти результаты показывают, что алгоритм 1 может эффективно уменьшить размерность данных. Для оценки выбранных атрибутов мы используем кластеризацию k-средних в качестве функций проверки. Выбранные алгоритмы реализованы в пакете (машинное обучение на Python). В работе мы устанавливаем номера кластеров в соответствии с классами каждого набора данных.

Коэффициенты профиля до или после сокращения приведены в таблице 3.14. При указании одинакового количества кластеров для исходного HIS и уменьшенного HIS в таблице 3.14 показано, что коэффициент силуэта для уменьшенного HIS лучше, чем для исходного HIS.

Таблица 3.12

Пять наборов данных из UCI

Набор данных	Образцы	Условные атрибуты				Классы
		Все	Номинальные	Порядковые	Масштабируемые	
Набор 4	299	12	5	6	1	2
Набор 5	690	14	6	5	3	2
Набор 6	690	15	9	3	3	2
Набор 7	270	13	6	6	1	2
Набор 8	303	13	8	4	1	5

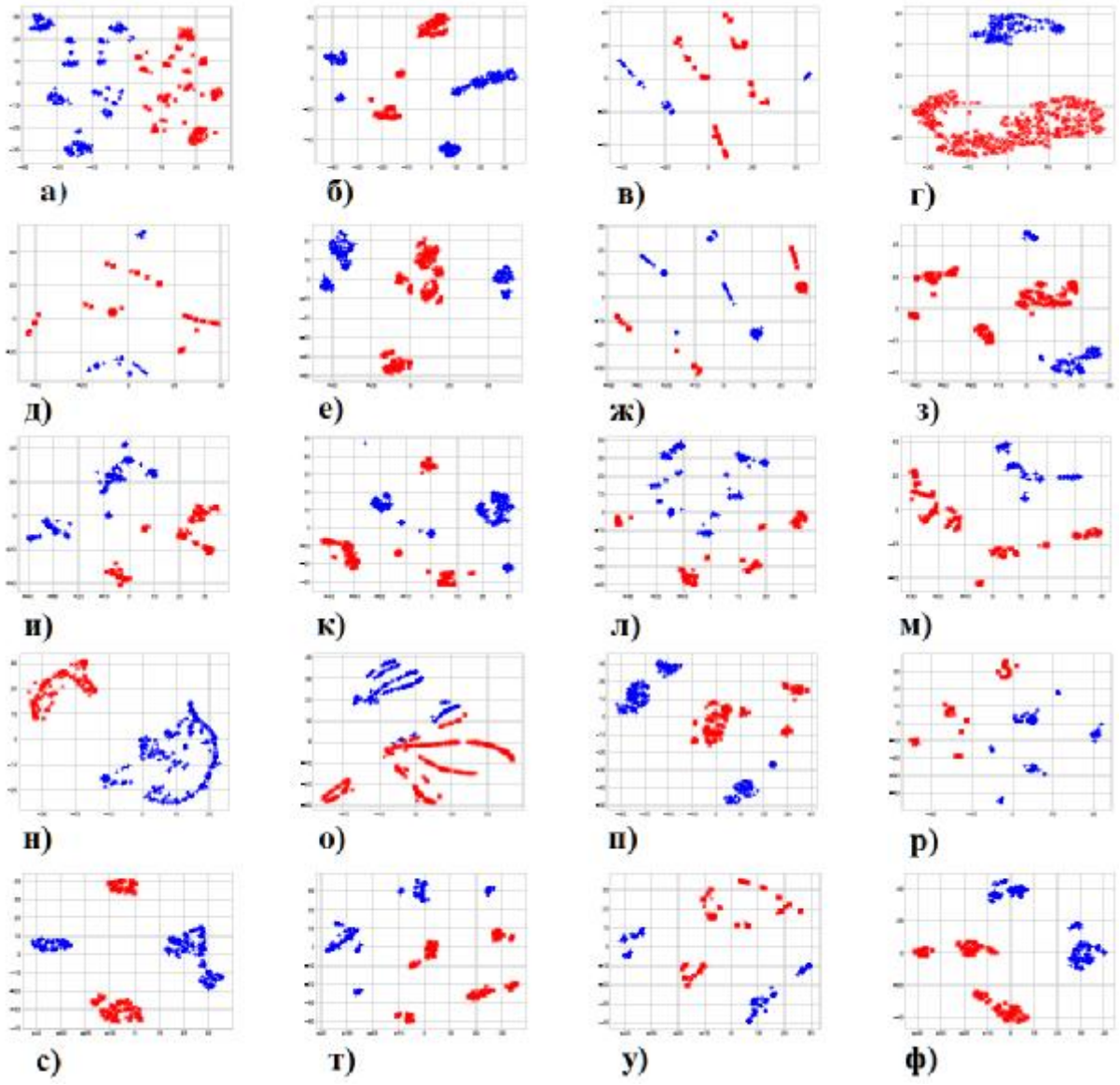


Рис. 3.4 Кластеризация с использованием KMeans на Наборе 5 с двумя кластерами: а – исходные данные; б) редукция A_1; в) редукция A_2; г) редукция A_3; д) редукция A_4; е) редукция A_5; ж) редукция A_6; з) редукция A_7; и) редукция A_8; к) редукция A_9; л) редукция A_{10}; м) редукция A_{11}; н) редукция A_{12}; о) редукция A_3; п) редукция A_{14}; р) редукция A_{15}; с) редукция A_{16}; т) редукция A_{17}; у) редукция A_{18}; ф) редукция A_{19}

Таблица 3.13

Результаты редукции пяти наборов данных из UCI

Набор данных	Редукции
Набор 4	$D_1 = \{a_1, a_4, a_5, a_7, a_{11}, a_{12}\}$ $D_2 = \{a_3, a_8\}$

Набор данных	Редукции
	$D_3 = \{a_1, a_5, a_9, a_{10}\}$ $D_4 = \{a_3, a_5, a_8, a_{10}, a_{12}\}$ $D_5 = \{a_5, a_7\}$ $D_6 = \{a_1, a_8\}$ $D_7 = \{a_1, a_4, a_7, a_8, a_9, a_{11}\}$ $D_8 = \{a_5, a_6, a_7, a_{12}\}$ $D_9 = \{a_1, a_6, a_7\}$ $D_{10} = \{a_1, a_5, a_7, a_{10}\}$ $D_{11} = \{a_1, a_8, a_{11}\}$ $D_{12} = \{a_2, a_7\}$ $D_{13} = \{a_1, a_3, a_5, a_{11}\}$ $D_{14} = \{a_2, a_4\}$ $D_{15} = \{a_1, a_2, a_3, a_5, a_7, a_8, a_{10}\}$ $D_{16} = \{a_1, a_2, a_3, a_6, a_{10}, a_{11}\}$ $D_{17} = \{a_1, a_3, a_5, a_7, a_9, a_{11}\}$ $D_{18} = \{a_3, a_7, a_{10}, a_{12}\}$ $D_{19} = \{a_1, a_4, a_6, a_{10}, a_{11}\}$
Набор 5	$A_1 = \{a_1, a_2, a_3, a_4, a_7, a_9, a_{10}, a_{13}, a_{14})\}$ $A_2 = \{a_1, a_4, a_5, a_{14}\}$ $A_3 = \{a_2, a_3, a_4, a_7, a_{13}\}$ $A_4 = \{a_1, a_4, a_5, a_{12}\}$ $A_5 = \{a_1, a_2, a_3, a_5, a_6, a_7, a_{13}, a_{14}\}$ $A_6 = \{a_1, a_2, a_6, a_9, a_{13}\}$ $A_7 = \{a_1, a_2, a_4, a_5, a_{10}, a_{12}\}$ $A_8 = \{a_1, a_5, a_6, a_7, a_{10}, a_{11}, a_{12}\}$ $A_9 = \{a_2, a_3, a_4, a_8, a_9, a_{10}, a_{12}, a_{13}\}$ $A_{10} = \{a_1, a_4, a_5, a_7, a_9, a_{11}, a_{12}, a_{13}, a_{14}\}$ $A_{11} = \{a_1, a_5, a_6, a_{13}, a_{14}\}$ $A_{12} = \{a_1, a_{10}, a_{13}, a_{14}\}$ $A_{13} = \{a_1, a_6, a_7, a_{13}\}$ $A_{14} = \{a_2, a_3, a_5, a_6, a_7, a_8, a_9, a_{10}, a_{12}, a_{13}, a_{14}\}$ $A_{15} = \{a_1, a_4, a_6, a_9, a_{14}\}$ $A_{16} = \{a_1, a_3, a_5, a_7, a_9\}$ $A_{17} = \{a_2, a_6, a_7, a_8, a_9, a_{10}, a_{11}\}$ $A_{18} = \{a_1, a_5, a_6, a_9, a_{10}\}$ $A_{19} = \{a_2, a_5, a_6, a_8, a_{11}, a_{13}\}$
Набор 6	$B_1 = \{a_1, a_7, a_{11}\}$ $B_2 = \{a_2, a_3, a_5, a_6, a_{10}, a_{12}, a_{13}, a_{14}\}$ $B_3 = \{a_2, a_5, a_{14}\}$ $B_4 = \{a_1, a_4, a_6, a_7, a_{11}, a_{14}\}$ $B_5 = \{a_1, a_{15}\}$ $B_6 = \{a_1, a_2, a_7, a_{14}\}$

Набор данных	Редукции
	$B_7 = \{a_1, a_2, a_4, a_7, a_8, a_{10}, a_{12}, a_{15}\}$ $B_8 = \{a_2, a_3, a_4, a_7, a_8, a_9, a_{10}, a_{11}, a_{12}, a_{13}, a_{15}\}$ $B_9 = \{a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}\}$ $B_{10} = \{a_1, a_4, a_5, a_6, a_8, a_9, a_{10}, a_{11}, a_{12}\}$ $B_{11} = \{a_2, a_3, a_5, a_7, a_8, a_9, a_{10}, a_{11}, a_{13}, a_{15}\}$ $B_{12} = \{a_1, a_4, a_5, a_7, a_{10}, a_{12}\}$ $B_{13} = \{a_2, a_3, a_5, a_7, a_9, a_{13}\}$ $B_{14} = \{a_1, a_4, a_6, a_9, a_{15}\}$ $B_{15} = \{a_2, a_6, a_{14}, a_{15}\}$ $B_{16} = \{a_2, a_3, a_5, a_6, a_9, a_{11}, a_{13}, a_{14}\}$ $B_{17} = \{a_1, a_3, a_9\}$ $B_{18} = \{a_1, a_2, a_7, a_{15}\}$ $B_{19} = \{a_1, a_7\}$
Набор 7	$C_1 = \{a_1, a_2, a_3\}$ $C_2 = \{a_2, a_3, a_7, a_8, a_{12}, a_{13}\}$ $C_3 = \{a_2, a_6, a_7, a_{10}, a_{11}, a_{13}\}$ $C_4 = \{a_1, a_2, a_5, a_8\}$ $C_5 = \{a_1, a_6, a_{12}\}$ $C_6 = \{a_2, a_4, a_8, a_{13}\}$ $C_7 = \{a_1, a_2, a_3, a_6, a_8, a_{10}, a_{12}, a_{13}\}$ $C_8 = \{a_4, a_7, a_{10}, a_{13}\}$ $C_9 = \{a_2, a_4, a_{11}\}$ $C_{10} = \{a_2, a_6, a_7, a_8, a_{12}, a_{13}\}$ $C_{11} = \{a_1, a_4, a_7, a_8, a_{10}\}$ $C_{12} = \{a_1, a_4, a_8, a_9\}$ $C_{13} = \{a_3, a_5, a_6\}$ $C_{14} = \{a_3, a_8\}$ $C_{15} = \{a_2, a_{10}\}$ $C_{16} = \{a_1, a_2, a_9, a_{10}, a_{11}, a_{13}\}$ $C_{17} = \{a_2, a_7, a_8, a_{13}\}$ $C_{18} = \{a_1, a_2, a_3, a_4, a_5, a_{12}, a_{13}\}$ $C_{19} = \{a_1, a_3, a_6\}$
Набор 8	$E_1 = \{a_2, a_7, a_{12}, a_{13}\}$ $E_2 = \{a_1, a_2, a_6, a_{13}\}$ $E_3 = \{a_3, a_9, a_{11}\}$ $E_4 = \{a_3, a_5, a_6, a_{11}, a_{12}\}$ $E_5 = \{a_2, a_4, a_6, a_7, a_8, a_{10}, a_{11}, a_{12}\}$ $E_6 = \{a_2, a_4, a_{10}, a_{11}\}$ $E_7 = \{a_1, a_{12}\}$ $E_8 = \{a_1, a_3, a_5, a_{10}, a_{11}\}$ $E_9 = \{a_2, a_4, a_6, a_{10}\}$ $E_{10} = \{a_1, a_2, a_4, a_6, a_7, a_9, a_{13}\}$

Набор данных	Редукции
	$E_{11} = \{a_1, a_{12}\}$ $E_{12} = \{a_4, a_7, a_{13}\}$ $E_{13} = \{a_4, a_5, a_{10}, a_{11}, a_{12}\}$ $E_{14} = \{a_1, a_3, a_6, a_7, a_8, a_9\}$ $E_{15} = \{a_2, a_4, a_5, a_6, a_7, a_9, a_{11}\}$ $E_{16} = \{a_1, a_9, a_{12}, a_{13}\}$ $E_{17} = \{a_2, a_{11}\}$ $E_{18} = \{a_1, a_3\}$ $E_{19} = \{a_1, a_2, a_5, a_7, a_9, a_{11}, a_{12}, a_{13}\}$

Таблица 3.14

Коэффициенты профиля (S.c.) до и после редукции

Набор 5		Набор 6		Набор 7	
HIS	S.c.	HIS	S.c.	HIS	S.c.
Raw	0.2150	Raw	0.4812	Raw	0.5154
A_1	0.3686	B_1	0.8687	C_1	0.5556
A_2	0.5161	B_2	0.5668	C_2	0.5658
A_3	0.6732	B_3	0.7729	C_3	0.6360
A_4	0.4581	B_4	0.5241	C_4	0.6652
A_5	0.3787	B_5	0.9823	C_5	0.7821
A_6	0.4596	B_6	0.8620	C_6	0.5973
A_7	0.4133	B_7	0.8067	C_7	0.6453
A_8	0.3673	B_8	0.7982	C_8	0.7058
A_9	0.3725	B_9	0.5071	C_9	0.6565
A_{10}	0.2509	B_{10}	0.5516	C_{10}	0.6349
A_{11}	0.5827	B_{11}	0.8141	C_{11}	0.7656
A_{12}	0.8752	B_{12}	0.8012	C_{12}	0.6468
A_{13}	0.6951	B_{13}	0.8334	C_{13}	0.6252
A_{14}	0.3730	B_{14}	0.5890	C_{14}	0.6763
A_{15}	0.3998	B_{15}	0.6354	C_{15}	0.8135
A_{16}	0.4412	B_{16}	0.5917	C_{16}	0.7531
A_{17}	0.3587	B_{17}	0.5548	C_{17}	0.6760
A_{18}	0.4330	B_{18}	0.8631	C_{18}	0.6675
A_{19}	0.3861	B_{19}	0.8706	C_{19}	0.6039
Набор 4		Набор 8			
HIS	S.c.	HIS	S.c.		
Raw	0.1969	Raw	0.2765		
D_1	0.3918	E_1	0.54970		
D_2	0.8083	E_2	0.6599		

D_3	0.4685	E_3	0.5072
D_4	0.6403	E_4	0.3649
D_5	0.4611	E_5	0.3447
D_6	0.4773	E_6	0.7240
D_7	0.4614	E_7	0.6979
D_8	0.6037	E_8	0.5044
D_9	0.7207	E_9	0.5532
D_{10}	0.6447	E_{10}	0.5171
D_{11}	0.7129	E_{11}	0.6990
D_{12}	0.8802	E_{12}	0.8519
D_{13}	0.6475	E_{13}	0.5768
D_{14}	0.5939	E_{14}	0.4491
D_{15}	0.4222	E_{15}	0.3511
D_{16}	0.2919	E_{16}	0.5250
D_{17}	0.5906	E_{17}	0.9731
D_{18}	0.6387	E_{18}	0.6989
D_{19}	0.2901	E_{19}	0.3855

Архитектура программной системы управления гетерогенной информационной системой, отличающаяся использованием отношения эквивалентности на множестве объектов для измерения неопределенности системы, и обеспечивающая редукцию многомерных анализируемых атрибутов на основе грануляции информации и информационной энтропии, представлена на рис. 3.5.

3.6. Выводы к главе 3

Чтобы разобраться с нечетким отношением, было рассмотрено приблизительное равенство между двумя нечеткими множествами. На основе этого приблизительного равенства было построено отношение эквивалентности для множества объектов HIS. И разбиение на этом множестве объектов HIS было вызвано этим отношением эквивалентности. На основе этого разбиения изучается измерение неопределенности для HIS. Численный эксперимент и анализ эффективности показывают, что предложенные меры подходят для HIS. В качестве применения

предложенных мер было приведено уменьшение атрибутов в HIS. В будущем мы применим предложенные меры к интеллектуальному анализу данных.

Рис. 3.5. Архитектура программной системы управления гетерогенной информационной системой

Таким образом, предложена архитектура программной системы управления гетерогенной информационной системой, отличающаяся использованием отношения эквивалентности на множестве объектов для измерения неопределенности системы, и обеспечивающая редукцию многомерных анализируемых атрибутов на основе грануляции информации и информационной энтропии.

Источники к главе 3

3.1. Beaubouef T, Petry FE, Arora G (1998) Information-theoretic measures of uncertainty for rough sets and rough relational databases. Inf Sci 109:185–195.

3.2. Chen YM, Wu KS, Chen XH, Tang CH, Zhu QX (2014) An entropy-based uncertainty measurement approach in neighborhood systems. Inf Sci 279:239–250.

3.3. Dunn OJ (1961) Multiple comparisons among means. J Am Stat Assoc 56(293):52–64.

3.4. Ditsch I, Gediga G (1998) Uncertainty measures of rough set prediction. Artif Intell 106(1):109–137.

3.5. Dai JH, Wang WT, Hao QX, Tian W (2012) Uncertainty measurement for interval-valued decision systems based on extended conditional entropy. Knowl-Based Syst 27:443–450.

3.6. Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat 11(1):86–92.

3.7. Hu QH, Yu DR, Liu JF, Wu CX (2008) Neighborhood rough set based heterogeneous feature subset selection. Inf Sci 178:3577–3594.

3.8. Hu QH, Yu DR, Xie ZX (2006) Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recogn Lett 27(5):414–423.

3.9. Li ZW, Liu YY, Li QG, Qin B (2016) Relationships between knowledge bases and related results. Knowl Inf Syst 49:171–195.

3.10. Li ZW, Zhang PF, Ge X, Xie NX, Zhang GQ, Wen CF (2019) Uncertainty measurement for a fuzzy relation information system. IEEE Trans Fuzzy Syst 27(12):2338–2352.

3.11. Li ZW, Huang D, Liu XF, Xie NX, Zhang GQ (2020a) Information structures in a covering information system. Inf Sci 507:449–471.

3.12. Li ZW, Liu XF, Dai JH, Chen JL, Fujita H (2020b) Measures of uncertainty based on Gaussian kernel for a fully fuzzy information system. Knowl-Based Syst 196:105791.

3.13. Li ZW, Zhang GQ, Wu WZ, Xie NX (2020c) Measures of uncertainty for knowledge bases. Knowl Inf Syst 62:611–637.

3.14. Liang JY, Qian YH (2008) Information granules and entropy theory in information systems. Sci China (Ser F) 51:1427–1444.

3.15. Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht.

3.16. Pawlak Z, Skowron A (2007) Rudiments of rough sets. Inf Sci 177:3–27.

3.17. Pawlak Z, Skowron A (2007) Rough sets: some extensions. Inf Sci 177:28–40.

3.18. Pawlak Z, Skowron A (2007) Rough sets and Boolean reasoning. Inf Sci 177:41–73.

3.19. Shannon C (1948) A mathematical theory of communication. Bell Syst Techn J 27:379–423.

3.20. Sanchez MA, Castro JR, Castillo O, Mendoza O, Rodriguez-Diaz A, Melin P (2017) Fuzzy higher type information granules from an uncertainty measurement. Granul Comput 2:95–103.

3.21. Sun BZ, Ma WM, Chen DG (2014) Rough approximation of a fuzzy concept on a hybrid attribute information system and its uncertainty measure. Inf Sci 284:60–80.

3.22. Wang CZ, Huang Y, Shao MW, Chen DG (2019) Uncertainty measures for general fuzzy relations. Fuzzy Sets Syst 360:82–96.

3.23. Wang XD, Song YF (2018) Uncertainty measure in evidence theory with its applications. Appl Intell 48:1672–1688.

3.24. Xie NX, Liu M, Li ZW, Zhang GQ (2019) New measures of uncertainty for an interval-valued information system. Inf Sci 470:156–174.

3.25. Yao YY (2003) Probabilistic approaches to rough sets. Expert Syst 20:287–297.

3.26. Yu B, Guo LK, Li QG (2019) A characterization of novel rough fuzzy sets of information systems and their application in decision making. Expert Syst Appl 122:253–261.

3.27. Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353.

3.28. Zhang XX, Chen DG, Tsang EC (2017) Generalized dominance rough set models for the dominance intuitionistic fuzzy information systems. Inf Sci 378:1–25.

3.29. Zeng AP, Li TR, Liu D, Zhang JB, Chen HM (2015) A fuzzy rough set approach for incremental feature selection on hybrid information systems. Fuzzy Sets Syst 258:39–60.

3.30. Zhang GQ, Li ZW, Wu WZ, Liu XF, Xie NX (2018) Information structures and uncertainty measures in a fully fuzzy information system. Int J Approx Reason 101:119–149.

3.31. Zhang X, Mei CL, Chen DG, Li JH (2016) Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy. Pattern Recogn 56:1–15.

4. Особенности управления гетерогенными данными

4.1. Комплексный подход к обработке гетерогенных данных с активным обучением

4.1.1. Извлечение и категоризация текстовых данных

По мере роста проникновения интернет-услуг резко возросло число пользователей Интернета, что привело к увеличению разнообразия материалов, к которым можно получить доступ через Интернет. Извлечение и категоризация текстовых данных является важной областью исследований в современном мире.

Неоднородность данных и высокая размерность - два главных препятствия на пути решения этой проблемы, однако выбор функций может помочь. Электронная почта, онлайн-страницы, электронные документы и SMS-сообщения - все это формы коммуникации, которые в значительной степени основаны на тексте. Невозможно оспорить тот факт, что контролируемое обучение является наиболее распространенным и эффективным методом преодоления трудностей машинного обучения.

Можно классифицировать текст, используя контролируемые методы машинного обучения, которые основаны на ранее помеченных или аннотированных экспертами наборах данных, чтобы вывести общее правило, применимое к большинству наборов данных. За последнее десятилетие алгоритмический прогресс был огромным. Однако производительность модели в значительной степени зависит от качества и объема данных, что требует затрат времени, денег и человеческих ресурсов, и, следовательно, могут возникать проблемы с анализом разнородных данных.

Стало возможным использовать подходы машинного обучения для оценки разнородных текстовых данных, улучшив необходимые процессы предварительной обработки, которые учитывают особые качества

текстовых данных. Для правильного анализа этих данных необходимы такие методы предварительной обработки, как языковая маркировка, нормализация и стемминг.

Поскольку контролируемое обучение требует от специалистов-людей ручного аннотирования больших объемов данных, оно является дорогостоящим и отнимает много времени. Точность и стабильность контролируемого моделирования обеспечивается при одновременном снижении затрат на операции по маркировке, выполняемые человеком, за счет активного обучения [4.1].

При работе с большими объемами разнородных данных ручная классификация нецелесообразна, поскольку требует времени, денег и ресурсов. В последние годы широкое распространение получила технология машинного обучения в исследованиях по классификации текстов, позволяющая автоматизировать категоризацию разнообразных данных. То есть, если некоторые примеры выбраны случайным образом и помечены опытным экспертом, то контролируемый алгоритм обучения может использовать этот обучающий набор для воспроизведения меток для всей коллекции примеров.

В результате этой работы мы можем использовать методы активного обучения для получения необходимой степени точности без необходимости изучать или помечать весь набор данных. Это дает возможность использовать активное обучение для улучшения анализа проектов машинного обучения на основе разнородных данных, которые были очень сложными с точки зрения финансовых ресурсов, а также технических и контекстуальных знаний, необходимых для работы людей-комментаторов.

Традиционные алгоритмы машинного обучения требуют подачи в алгоритм большого количества помеченных данных, как показано на рис. 4.1. Вместо того чтобы классифицировать весь набор данных в целом, в

случае активного обучения мы просим эксперта-человека назвать лишь несколько элементов. В соответствии с некоторыми мерами алгоритм последовательно выбирает наиболее информативные образцы и отправляет их эксперту по маркировке, который выдает реальные этикетки для тех образцов, которые были подвергнуты сомнению, как показано на рис. 4.2.

Это приводит нас к разработке предлагаемого подхода, который позволяет нам управлять процессом отбора таким образом, что нам нужно помечать только небольшую выборку текста для достижения определенной степени точности классификации, а не случайным образом выбирать экземпляры, составляющие текстовые данные. Это актуальная проблема активного обучения; табл. 4.1 иллюстрирует разницу между традиционным машинным обучением (ML) и ML с активным обучением.

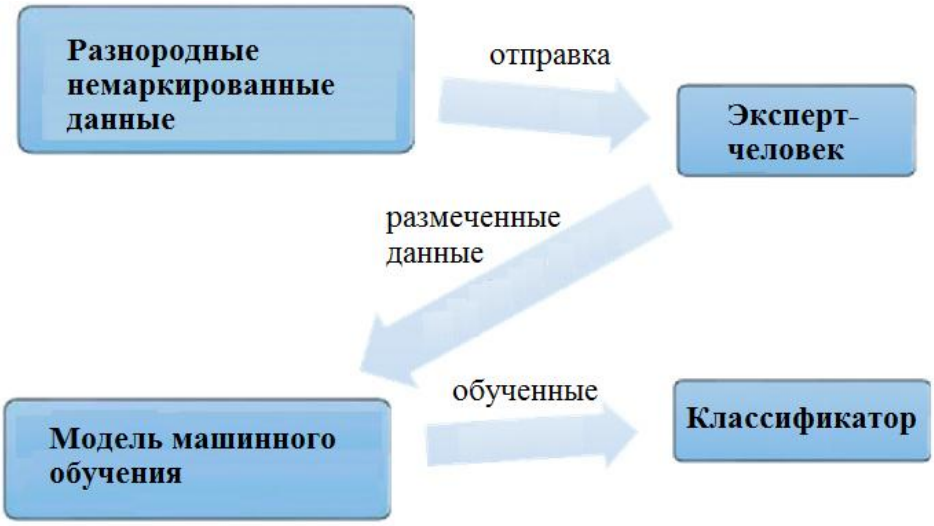


Рис. 4.1. Процесс использования традиционных методов машинного обучения для обработки немаркированных данных

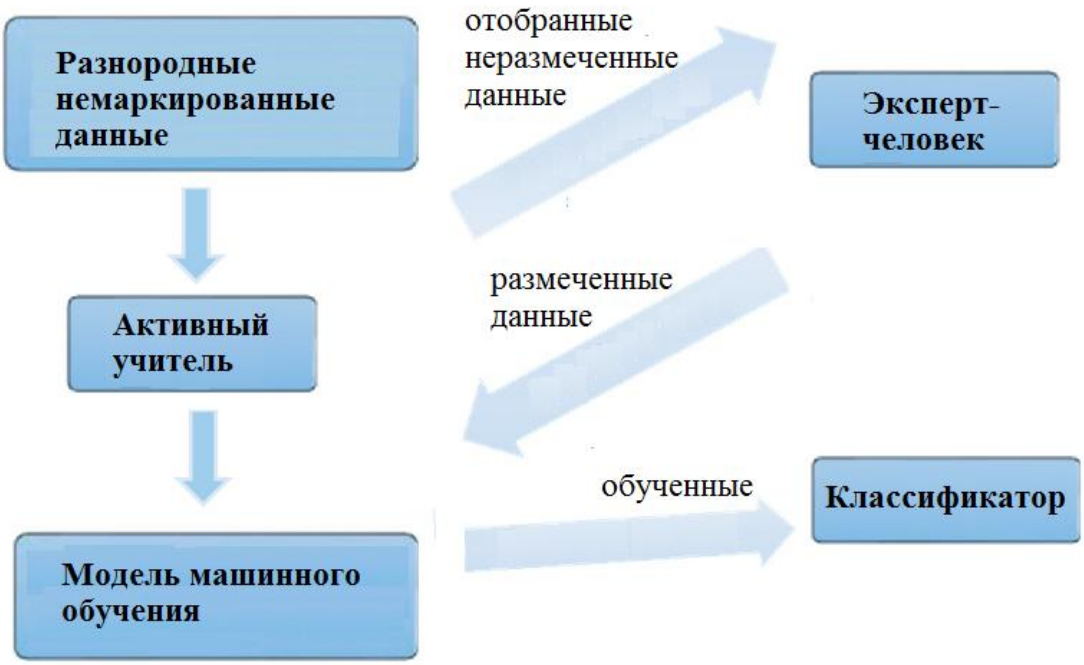


Рис. 4.2. Процесс активного машинного обучения для обработки немаркированных данных

Таблица 4.1

Сравнение между традиционным ML и ML с активным обучением

	Традиционное ML	ML с активным обучение
Выборка данных	Случайным образом выбираются данные из немаркированного пула для обучения и тестирования	Выбор данных по (стратегии запроса) для набора тестов
Маркировка	Назначение меток выбранным данным с помощью эксперта-человека	
Обучение	Обучение модели ML	
Оценка	Проверка производительности модели	

Как видно из табл. 4.1, различие между машинным обучением и машинным обучением с активным обучением заключается в методе запроса, используемом для выбора обучающего набора данных для маркировки из немаркированного пула данных.

Обучающий ансамбль - это важный компонент машинного обучения, который включает в себя ряд обучающих алгоритмов для повышения способности и точности прогнозирования.

Выделение признаков необходимо для групповой классификации, чтобы преобразовать текст в числовые представления. В результате производительность категоризации значительно снижается. Групповые классификаторы объединяют множество базовых классификаторов в единый надежный классификатор в контексте машинного обучения. Для предварительной обработки данных используются методы обработки данных на естественном языке, такие как мешок слов (BOW), периодичность использования термина - обратная частота документа (TF-IDF) и word2vec.

В качестве классификаторов, использованных в предлагаемом методе ансамбля, использовались методы опорных векторов (SVM), логистической регрессии (LR), наивного Байеса и случайного леса (RF). Ансамблевый классификатор, использующий TF-IDF, превзошел по эффективности методику извлечения признаков BOW или word2vec.

4.1.2. Состояние проблемы

Сокращение количества маркированных образцов необходимо для анализа и классификации разнородных данных из-за высокой стоимости ручной маркировки, и несколько недавних исследований были посвящены этой теме. Использование контролируемых методов классификации данных рассматривается как препятствие, и, следовательно, вместо этого используются неконтролируемые методы, такие как кластеризация. Поскольку анализ разнородных данных является актуальной темой среди ученых, каждый год проводится несколько исследований.

В результате на точность модели классификации сильно влияет ручная маркировка случайных выборок данных, известная как пассивное

обучение. Точность модели низкая из-за использования краудсорсинга для идентификации немаркированных данных. Вот несколько обзоров литературы, которые имеют отношение к нашей работе.

Согласно [4.3, 4.4], метод k-ближайших соседей (KNNS), TF-IDF, наивный байесовский алгоритм, дерево решений, гибридные подходы и методы опорных векторов были представлены в качестве широкого метода автоматической классификации текста. Также обсуждались управление большим количеством функций и работа с неструктурированным текстом, устранение недостающей информации и выбор эффективной стратегии машинного обучения для обучения текстовых классификаторов. В ходе тестирования алгоритм TFIDF дал правильные результаты. Оба варианта дали удовлетворительные результаты. Для достижения лучших результатов эта система должна быть расширена и усовершенствована.

Согласно [4.5-4.7], статистические методы, такие как наивный байесовский анализ, k-NNS и SVMS, являются наиболее подходящими подходами для получения результатов классификации текстов. В этой работе классификатор классифицировал текст с хорошими результатами и пассивным подходом к обучению посредством случайного отбора образцов перед процессом маркировки аннотатором или экспертной системой, что является методом, который имеет слабые стороны. Активные методы обучения также позволяют специалистам-специалистам начинать с небольшого набора данных, которые помогут в создании первоначальной модели. Машинное обучение выбирает наиболее полезные образцы данных на основе результатов моделирования, тем самым повышая точность. Этот метод применяется до тех пор, пока не будет достигнут требуемый уровень точности.

Несмотря на то, что во многих областях, таких как [4.8, 4.9], в большинстве классификаторов были достигнуты высочайшие уровни точности, групповые классификаторы являются успешными, особенно в

текстовой классификации. В литературе было разработано множество подходов к извлечению, включая BOW, TF-IDF и word2vec, для классификации ансамблей.

BOW - один из самых популярных и часто используемых методов выделения признаков в процессе классификации. BOW в основном предназначен для представления текста в наборе значений N, известном как векторные числа, отражая количество слов и частот в этих векторах [4.10]. В [4.11] предлагался комплексный подход к классификации текста, в котором для выделения признаков использовался BOW, и использовались наивный байесовский классификатор, линейный дискриминантный анализ и метод опорных векторов; результаты были многообещающими.

TF-IDF является одним из важных методов расширенной парадигмы BOW-подхода. Несколько более ранних исследований подтвердили предпочтение TF-IDF для определенных классификаторов перед другими технологиями извлечения признаков [4.12]. Влияние четырех методов определения характеристик, таких как word2vec, TF-IDF, doc2vec и счетчиков SVM, LR, NB и RF, было изучено в [4.13]. Например, функции TF-IDF и counter-vector обеспечивают высочайшую точность, метод опорных векторов и логистическая регрессия с помощью TF-IDF или counter-векторизатора достигли максимальной точности, и во многих более ранних публикациях TF-IDF рассматривается для функционального извлечения для групповых классификаций [4.14, 4.15].

В случае несбалансированных наборов данных классов в [4.16] рекомендуется использовать word2vec с совокупными классификаторами. Кроме того, в [4.17] используется word2vec для категоризации текста с помощью совокупного классификатора, который интегрирует модели нейронных сетей.

Согласно [4.18, 4.19], классификатор достиг удовлетворительных

результатов в классификации текста с помощью наивного Байеса, k-ближайших соседей и метода опорных векторов, а также с помощью предварительной обработки, выбора признаков, семантических методов и методов машинного обучения, таких как NB, дерево решений и SVM-классификация.

4.1.3. Активный подход к обучению и классификация разнородных текстов

Активное обучение вызывает большой интерес у исследователей, которые сокращают количество времени, затрат и усилий на маркировку данных во многих приложениях за счет повышения эффективности контролируемых алгоритмов машинного обучения на этапе обучения с минимальным участием человека; выбирая наименьший возможный объем обучающих данных, это обеспечивает высокую эффективность классификации в процессе обучения и на этапе тестирования [4.20]. Активное обучение обычно используется с контролируемыми методами машинного обучения, чтобы еще больше свести к минимуму усилия по маркировке данных и, как следствие, сократить объем человеческой работы, необходимой для поиска соответствующей информации и принятия решений.

Активное обучение - это реактивный, повторяющийся метод, который позволяет проводить высокопроизводительную классификацию с использованием данных с небольшой маркировкой. В то время как пассивное обучение работает с фиксированным набором помеченных образцов, предоставляемых алгоритму обучения, который используется для построения модели, парадигмы активного обучения требуют, чтобы алгоритм обучения отбирал результаты обучения путем выбора наиболее информативных примеров [4.21]. Активное обучение обычно используется, когда доступны большие объемы немаркированных данных.

Подход к активному обучению основан на концепции, согласно которой алгоритм обучения выбирает выборки данных, которые будут помечены вручную, вместо того, чтобы полагаться на случайную выборку или предопределенные критерии отбора экспертом-человеком. Существуют различные типы активного обучения, такие как активное обучение на основе пула и поточное активное обучение.

Активное обучение на основе пула данных является наиболее распространенной формой активного обучения, процесс обучения и маркировки выполняется пакетами, извлекаемыми из пула 20 немаркированных данных. Как показано на рис. 4.3, алгоритм обучения обучается с использованием помеченных образцов после маркировки каждой партии и выбора новых образцов для обучения, что может повысить эффективность обучения; алгоритм может быть записан так, как показано на рис. 4.4. Но данные передаются алгоритму в виде потока со всеми обучающими выборками при активном обучении на основе потока. Каждый случай независимо передается алгоритму для рассмотрения. Этот пример должен быть помечен или не помечен непосредственно алгоритмом. Некоторые примеры обучения в этом пуле помечаются экспертом-человеком, и непосредственно перед отображением следующего примера алгоритм получает метку.

Качество используемых данных оказывает значительное влияние на результаты задачи классификации. Процесс маркировки данных является ключевым препятствием для классификации, особенно при нынешней доступности ресурсов данных. Следовательно, неоднородность данных стала новой проблемой для обработки данных, выбор наиболее полезных данных зависит от используемой меры неопределенности [4.23]. Алгоритм активного обучения в выборке на основе пула выбирает экземпляры, которые вносят свой вклад в увеличивающуюся обучающую выборку, наиболее полезными выборками из которой являются те, которые

наименее достоверны.

Рис. 4.3. Процесс активного обучения классификации текстов

```
Begin
    ε – предельная ошибка обучения
    Разделить данные на немаркированный пул P
        и тестовый набор S;
    Разделить P на пакеты;
    Случайным образом из P выбрать k образцов;
    T= {выбранные образцы };
    While Ошибка обучения > ε do
        Обучить модель с использованием T;
        Протестировать обученную модель на S;
        Измерить качество обучения;
        Для e∈P вычислить неопределенность;
        Выбрать k наиболее информативных
            образцов на основе показателя
            неопределенности;
        Переместить k образцов в T;
        Удалить эти k образцов из P
    End While
End
```

Рис. 4.4. Алгоритм 1 активного обучения на основе пула

Стратегии запроса, которые включают выборку с неопределенностью, являются одними из наиболее распространенных. При

выборке с неопределенностью [4.24] наиболее неопределенной выборкой будут те, которые имеют наименьшую разницу с наибольшей вероятностью между двумя прогнозами, которым больше всего доверяют.

Используются выборки с неопределенностью, которые выбирают или запрашивают некоторые из наших немаркированных пулов данных для наименее определенных выборок, затем помеченные выборки данных классифицируются с помощью классификатора ансамблей, и все больше немаркированных выборок постоянно помечаются, пока точность модели не будет увеличена до приемлемого уровня; после обработки каждой партии выборок с использованием метода обучения методы на основе полного набора обучающих данных постоянно проводится сравнение.

Подход к активному обучению проиллюстрирован на рис. 4.5 следующим образом: Используя метод запроса, созданный моделью активного обучения, эксперт-человек запрашивает немаркированные экземпляры, которые модель активного обучения выбирает и представляет эксперту-человеку. Эксперт помечает эти экземпляры и возвращает их в модель активного обучения. После каждого обновления активная обучающая модель переобучается. Эта процедура повторяется до тех пор, пока не будет выполнено условие остановки, например, максимальное количество итераций или минимальный порог изменения точности классификации.

Цель активных методов обучения состоит в создании модели с использованием как можно меньшего количества маркированных примеров, чтобы свести к минимуму привлечение экспертов без ущерба для эффективности классификации; это приведет к сокращению затрат и времени на маркировку образцов. Предлагаемая ансамблевая модель, основанная на активном обучении, настолько интеллектуальна, что может выбирать любые данные, которые модель хочет обработать, но все же рекомендуется обучать модель на выборках, которые значительно влияют

на ее производительность.

Рис. 4.5. Обзор подхода к активному обучению

4.1.3.1. Методология

В этом разделе рассматривается общий подход, использованный в данном исследовании, который включает в себя предлагаемый подход, сбор и предварительную обработку наборов данных, извлечение признаков, групповые методы классификации и оценки.

В нашем новом подходе, основанном на пуле, процесс маркировки выполняется пакетами, выбранными из пула немаркированных данных; после маркировки каждого пакета алгоритм обучается с использованием этих пакетов; и этот процесс повторяется с набором новых образцов до тех пор, пока эффективность обучения не улучшится.

Мы используем выборку с неопределенностью, при которой из массива немаркированных данных отбираются наименее достоверные образцы. Затем эти образцы обрабатываются специалистом, и этот процесс повторяется до тех пор, пока не будут помечены все партии. После каждой партии точность проверяется до тех пор, пока не будет достигнута требуемая степень точности. Затем помеченные образцы данных классифицируются. Этот процесс повторяется для маркировки большего

количества немаркированных образцов до тех пор, пока точность модели не повысится до приемлемого уровня.

4.1.3.2. Сбор данных

Собранные наборы данных имеют различные размеры и источники; были использованы пять разнородных наборов данных из различных областей, таких как набор данных о реформе здравоохранения, набор данных Сандера Франдсена, набор финансовых фраз о банке, набор данных о сборе SMS-спама и набор данных о продажах учебников. В этом разделе мы приводим краткое описание каждого набора данных.

Мы использовали множество наборов данных. Первым из них был набор данных о реформе здравоохранения (HCR), в котором были просмотрены твиты, содержащие хэштег. Этот набор данных содержит более 1050 твитов. Второй набор данных - это набор данных из библиотеки Crowdflower "Данные для всех". Этот набор данных содержит более 850 твитов. Третий набор данных, банк финансовых фраз, содержит комментарии к заголовкам финансовых новостей. Этот набор данных содержит коллекцию из более чем 870 заголовков новостей. Четвертый набор данных, SMS-текстовые сообщения из базы данных UCI, содержит набор из более чем пяти тысяч SMS-текстовых сообщений для обнаружения спама, а последний набор данных представляет собой набор данных о продажах в качестве учебного пособия для проекта, управляемого Dataquest. Этот набор данных содержит более 2150 твитов.

4.1.3.3. Предварительная обработка

Этап предварительной обработки включает в себя методы удаления избыточных и незначительных данных для минимизации объема объекта. Этот шаг повышает точность алгоритма машинного обучения [4.25].

Предварительная обработка является основным этапом анализа больших данных и крайне важной процедурой подготовки данных. Очистка данных является важным этапом, и на нем необходимо обучить все модели ML. Для повышения качества анализа данных проводится предварительная обработка и интеграция данных. Эти методы снижают неопределенность информации и предлагают превосходные аналитические критерии, улучшают понимание сложности информации и делают процесс анализа данных надежным и эффективным; результаты существенно отличаются друг от друга без очистки и предварительной обработки. В данной работе рассматриваются пять популярных этапов подготовки текстовых данных [4.26], включающие очистку данных, удаление стоп-слов, разбиение на строчные буквы, токенизацию и создание основы. Ниже приводится краткое описание этих этапов:

- Очистка данных: Для улучшения качества данных были удалены нежелательные замечания, такие как точки, цифры, короткие фразы и некоторые символы.

- Удаление стоп-слов: Стоп-слова - это высокочастотные слова, которые не зависят от определенной темы, например предлоги. В исследованиях по анализу текста стоп-слова часто считаются избыточными и бесполезными, поскольку они часто повторяются, но не дают ценного смысла.

- Преобразование в нижний регистр: поскольку как строчные, так и прописные буквы не имеют четкого значения, все буквы со знаком преобразуются в буквы меньшего размера.

- Токенизация: Это процесс разделения текста на значимые сегменты, такие как слова или фразы.

- Стемминг: это операция по получению корневых форм всех слов. При обработке естественных языков стемминг необходим для

рассмотрения многих форм производных слов как единой основы.

4.1.3.4. Особенности извлечения

Этот этап включает анализ текста для выявления определенных признаков, которые описываются в виде вектора, понятного для классификации. Представление текста является важным и мощным этапом обработки естественного языка.

BOW - это наиболее распространенный прием, который использовался при групповом обучении в предыдущих исследованиях, включая различные стратегии извлечения. Частота употребления BOW и TF-IDF будет рассматриваться как дискретные символы. Порядок слов и их значение не принимаются во внимание. Word2vec решает эту проблему, собирая дополнительные данные о сходстве слов. Таким образом, экспериментальными методами являются BOW, TF-IDF и word2vec. Методы BOW и TF-IDF использовались с использованием библиотеки scikit-learn, а библиотека Genism использовалась для импорта моделей word2vec с использованием языка программирования Python. Было проведено обучение по групповой классификации, и набор тестов был использован для целей оценки.

4.1.4.5. Групповая классификация

Был создан групповой классификатор и оценен с использованием всех созданных функций. Для достижения основной цели этого исследования нам необходимо определить основные классификаторы и методику объединения таких классификаторов для создания модели ансамбля. Среди различных комбинированных методов голосование является эффективным, простым и часто используемым способом коллективного обучения. Для каждого из базовых наборов

классификаторов необходимо провести простейшее голосование, называемое большинством голосов, равный вклад и один голос.

Наиболее частые голосования являются конечным результатом классификации по ансамблевой модели. Таким образом, процедура голосования большинством голосов была применена в качестве комбинации. SVM, логистическая регрессия, наивный байесовский анализ и случайный лес были положены в основу модели ensemble, поскольку она широко известна и популярна при классификации разнородных текстовых данных. Классификатор ensemble был обучен с использованием обучающего подмножества каждого признака. Показатель точности рассматривался в качестве критерия оценки в тестовых подмножествах каждой функции для целей оценки.

4.1.4.6. Показатели оценки производительности

Смешанная матрица - это наиболее распространенный метод, используемый для обобщения эффективности метода классификации [4.25]. Соотношение правильно классифицированных случаев (TP + TN)/(TP + TN + FP + FN) может быть описано как несколько действий, основанных на результатах смешанной матрицы, таких как точность правильно классифицированных случаев. TP, FN, FP и TN обозначают количество истинно положительных результатов, ложноотрицательных результатов, ложноположительных результатов и истинно отрицательных результатов, а конкретные измерения используются для проверки каждой модели и оценки ее точности, отзывчивости и оценки F1 для каждого классификатора.

Точность (4.1), погрешность (4.2), реактивность (4.3) и оценка F1 (4.4) - вот наши оценочные показатели:

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \quad (4.1)$$

$$Precision = TP / (TP + FP) \quad (4.2)$$

$$Recall = TP / (TP + FN) \quad (4.3)$$

$$F1 = 2 \times Precision \times Recall/(Precision + Recall) \quad (4.4)$$

TP, FN, FP и TN определяются количеством реальных положительных случаев, количеством ложноотрицательных случаев и количеством истинно отрицательных случаев.

1. Точность является наиболее понятным показателем эффективности, и определяется процент правильной классификации элементов для всех оцениваемых объектов.

2. Определены правильные положительные результаты, отличающиеся от истинно положительных и ложноотрицательных.

3. Положительная точность делится на истинные и ложные срабатывания.

4. F1-оценка включает в себя расчет точности модели, использование точности и памяти; поэтому это идеальный шаг для анализа текста.

4.1.5. Результаты экспериментов

4.1.5.1. Цель и задачи эксперимента

В этом исследовании предлагается использовать активное обучение для разработки новой комплексной методики, улучшающей анализ разнообразных данных. Этот метод обеспечивает быстрый и эффективный способ обработки немаркированных данных, которые считаются разнородными, сокращая затраты на обработку и маркировку за счет грамотного отбора образцов без ущерба для точности классификации. Эту проблему можно устранить, используя большое количество немаркированных образцов, а затем выбирая партию, подходящую для классификации ML, на основе вероятностей и методов запроса, что является трудоемкой и чреватой ошибками задачей на этапе обучения. В этом исследовании использовались предоставленные данные и был

опробован рекомендованный метод анализа и понимания данных. Были использованы все ранее указанные наборы данных. Программирование было реализовано в среде Anaconda с использованием Jupyter notebook, которая представляет собой универсальную среду разработки на языке Python. Эксперименты проводились на устройстве со следующими характеристиками: платформа Windows 10, 8 ГБ оперативной памяти, процессор Core i5. Мы оценили модель ensemble и другие традиционные методы классификации, используя функции BOW, TF-IDF и word2vec.

Результаты показывают, насколько эффективен этот параметр для ансамблевой модели во всех наборах данных. Кроме того, функция BOW является наименее эффективным выбором во всех наборах данных для ансамблевой модели.

4.1.5.2. Обсуждение и оценка эксперимента

В этом разделе мы представляем полученные экспериментальные результаты; оцениваем производительность на основе результатов различных традиционных классификаторов ML и предложенной модели ensemble, которая с использованием TF-IDF достигла наивысшей точности; а также запускаем классификаторы для всех наборов данных и сравниваем производительность на основе отзыва, оценки F1 и точности.

Экспериментальные результаты выбранного набора данных были подтверждены с помощью этого измерения точности, запоминания и оценки F1. Результаты показали, что предложенный подход способен значительно повысить производительность.

В табл. 4.2 приведены результаты по точности традиционного классификатора ML и проведено сравнение выходных данных каждого классификатора с результатами ансамблевой модели, использованной в наборе данных HCR с помощью TF-IDF при разработке модели.

В результате этой оценки было определено, что классификатор моделей ensemble является наиболее точным из всех классификаторов ML. Окончательные результаты эксперимента показывают, что ансамблевый классификатор достиг наилучшей производительности среди всех других классификаторов с точностью 85%, погрешностью 0,84%, повторением 85% и F1-мерой 85%, используя методы извлечения признаков (TF-IDF). Напротив, классификатор логистической регрессии достигает самой низкой общей точности - 78%, погрешности - 79%, отзыва - 79% и F1-меры - 79% среди всех классификаторов. Таким образом, этот результат считается хорошим, в соответствии с объемом нашей выборки данных.

В табл. 4.3 приведены результаты по точности традиционного ML-классификатора и ансамблевой модели, а также сравниваются выходные данные каждого классификатора, использующего набор данных SF, с данными TFIDF при построении модели.

В результате этой оценки было определено, что классификатор моделей ensemble является наиболее точным из всех классификаторов ML. Окончательные результаты эксперимента показывают, что ансамблевый классификатор достиг наилучшей производительности среди всех других классификаторов с точностью 84%, погрешностью 0,85%, повторением 84% и F1-мерой 86, используя методы извлечения признаков (TF-IDF). Напротив, наивный байесовский классификатор достигает самой низкой общей точности - 78%, погрешности - 78%, отзывчивости - 80% и F1-меры - 78% среди всех классификаторов. Таким образом, этот результат считается хорошим, в соответствии с объемом нашей выборки данных.

В табл. 4.4 приведены результаты по точности традиционного классификатора ML и точности модели ensemble, а также результаты сравнения каждого классификатора с использованием набора данных FPB и TFIDF при построении модели.

В результате этой оценки было определено, что классификатор

моделей ensemble является наиболее точным из всех классификаторов ML. Окончательные результаты эксперимента показывают, что ансамблевый классификатор достиг наилучшей производительности среди всех других классификаторов с точностью 86%, погрешностью 0,85%, повторением 85% и F-мерой 86%, используя методы извлечения признаков (TF-IDF). В отличие от этого, классификатор случайных лесов достигает самой низкой общей точности (79%), погрешности (79%), отзыва (79%) и F1-показателя (79%) среди всех классификаторов. Таким образом, этот результат считается хорошим, в соответствии с объемом нашей выборки данных.

Таблица 4.2

Сравнительный анализ результатов традиционных классификаторов и ансамблевой модели: набор данных HCR с TF-IDF

Метод	T	Precision	Recall	F1-score
Логистическая регрессия	0.78	0.79	0.79	0.79
Случайный лес	0.80	0.79	0.79	0.78
SVM	0.79	0.79	0.80	0.79
Наивный Байес	0.80	0.78	0.79	0.78
Ансамблевая модель	0.85	0.84	0.85	0.85

Таблица 4.3

Сравнительный анализ результатов работы классификаторов: набор данных SF и TF-IDF

Метод	Accuracy	Precision	Recall	F1-score
Логистическая регрессия	0.79	0.79	0.80	0.79
Случайный лес	0.79	0.78	0.79	0.79
SVM	0.80	0.79	0.80	0.79
Наивный Байес	0.78	0.78	0.80	0.78
Ансамблевая модель	0.84	0.85	0.84	0.86

Таблица 4.4

Сравнительный анализ результатов традиционных классификаторов и ансамблевой модели: набор данных FPB с использованием TF-IDF

Метод	Accuracy	Precision	Recall	F1-score
Логистическая регрессия	0.80	0.79	0.80	0.79
Случайный лес	0.79	0.79	0.79	0.79
SVM	0.81	0.80	0.80	0.80
Наивный Байес	0.81	0.78	0.80	0.78
Ансамблевая модель	0.86	0.85	0.85	0.86

Таблица 4.5

Сравнительный анализ результатов традиционных классификаторов и ансамблевой модели: набор данных SMS с использованием TF-IDF

Метод	Accuracy	Precision	Recall	F1-score
Логистическая регрессия	0.78	0.79	0.80	0.79
Случайный лес	0.79	0.79	0.79	0.79
SVM	0.80	0.80	0.80	0.80
Наивный Байес	0.81	0.78	0.79	0.78
Ансамблевая модель	0.85	0.85	0.84	0.85

Таблица 4.6

Сравнительный анализ результатов работы традиционных классификаторов и ансамблевой модели: набор данных T BOOK с TF-IDF

Метод	Accuracy	Precision	Recall	F1-score
Логистическая регрессия	0.80	0.79	0.80	0.79
Случайный лес	0.81	0.79	0.79	0.79
SVM	0.79	0.80	0.80	0.80
Наивный Байес	0.80	0.78	0.79	0.78
Ансамблевая модель	0.86	0.86	0.85	0.86

В табл. 4.5 показаны результаты по точности традиционного ML-классификатора и по точности ансамблевой модели, а также сравниваются выходные данные каждого классификатора с использованием набора

данных SMS и TFIDF во время построения модели.

В результате этой оценки было определено, что классификатор моделей ensemble является наиболее точным из всех классификаторов ML. Окончательные результаты эксперимента показывают, что ансамблевый классификатор достиг наилучшей производительности среди всех других классификаторов с точностью 85%, погрешностью 0,85%, повторением 84% и F-мерой 85%, используя методы извлечения признаков (TF-IDF). Напротив, классификатор логистической регрессии достигает самой низкой общей точности - 78%, погрешности - 79%, отзыва - 80% и измерения Fmeasure - 79% среди всех классификаторов. Таким образом, этот результат считается хорошим, в соответствии с объемом нашей выборки данных.

В табл. 4.6 приведены результаты по точности традиционного классификатора ML и точности ансамблевой модели, а также результаты сравнения каждого классификатора с использованием набора данных из учебника и TF-IDF при построении модели.

В результате этой оценки было определено, что классификатор моделей ensemble является наиболее точным из всех классификаторов ML. Окончательные результаты эксперимента показывают, что ансамблевый классификатор достиг наилучшей производительности среди всех других классификаторов с точностью 86%, погрешностью 0,86%, повторением 85% и F-мерой 86%, используя методы извлечения признаков (TF-IDF). В отличие от этого, классификатор SVM достигает самой низкой общей точности - 79%, погрешности - 80%, отзыва - 80% и F-измерения - 80% среди всех классификаторов. Таким образом, этот результат считается хорошим в соответствии с объемом нашей выборки данных.

4.1.6. Итоги рассмотрения подхода к разработке и обнаружению наилучшей модели ML для анализа разнородных текстовых данных,

основанного на активном обучении и ансамблевых методах для классификаторов машинного обучения

В работе представлен новый подход, который фокусируется на разработке и обнаружении наилучшей модели ML для анализа разнородных текстовых данных, основанный на активном обучении и ансамблевых методах для классификаторов машинного обучения для решения проблем анализа разнородных данных и сокращения объемов данных. В этой работе для группового обучения были выбраны три метода извлечения данных: BOW, TF-IDF и word2vec. На пяти экспериментальных наборах данных было исследовано несколько ситуаций для применения трех стратегий извлечения с использованием модели ансамбля. Результаты эксперимента показывают, что метод TF-IDF повышает точность анализа совокупности в большей степени, чем другие; результаты подтвердили, что предложенный подход достигает цели исследования, обеспечивая высокопроизводительный анализ гетерогенных текстовых данных. Результаты показывают, что предложенный метод является превосходным и в большинстве случаев дает лучшие результаты, чем обычные алгоритмы ML. Исследования показывают, что для повышения общей точности модели важно использовать комплексные методы с активным обучением. Огромные преимущества активного обучения помогают свести к минимуму затраты и время, необходимые для обучения модели классификации разнородных наборов данных. Ожидается, что в будущем этот концептуальный тест будет распространен на более широкий набор данных, чтобы сосредоточиться на разработке тестов и выявлении фраз с помощью Bert, заменив word2vec на Quick Text. Более того, изменив метод комбинирования, мы сможем изучить более одной формы групповой классификации.

4.2. Оценка неопределенностей нулевого значения базы данных на основе искусственного интеллекта

4.2.1. Потеря информации и неопределенность как проблема современных БД

Из-за сложности объективного мира потеря информации и неопределенность являются обычным явлением. В качестве инструмента для отображения реального мира база данных использует нулевые значения, чтобы выразить проблему отсутствия информации. Для решения проблемы нулевого значения в базе данных с неопределенностью предлагается алгоритм оценки нулевого значения на основе искусственного интеллекта. Сначала анализируются характеристики неопределенной базы данных, затем строится модель поиска потерянной информации, а оценка пустого значения базы данных завершается выбором признаков и преобразованием данных, кластеризацией искусственного интеллекта, вычислением степени влияния, оценкой шага пустого значения и другими методами. Наконец, он анализирует временную сложность алгоритма и устраняет проблему низкого эффекта оценки традиционных алгоритмов. Результаты, подтвержденные экспериментальными данными, показывают, что предложенный алгоритм обладает более высокой точностью, чем традиционный алгоритм. Это показывает, что этот алгоритм может эффективно оценивать нулевое значение в базе данных с неопределенностью и имеет высокую практическую ценность для применения, а также может обеспечить теоретическую ценность для соответствующих исследований

В последние годы, в связи с активным развитием компьютерных технологий и сетевых информационных технологий, растет число различных информационных систем, которые выполняют все больше задач по хранению и сбору данных. Особенно в среде больших данных, при

быстром росте объема бизнес-данных организаций, все виды данных генерируются и обрабатываются с беспрецедентной скоростью [4.19]. Поэтому вопрос о том, как извлекать эффективную информацию из накопленных массивных данных, стал актуальным в академических кругах.

С развитием теории реляционных баз данных различные системы реляционных баз данных широко используются в различных областях социальной жизни, особенно в области интеллектуального анализа данных. Однако данные в реальных базах данных часто содержат шум, ошибки по умолчанию и неоднозначность, что влияет на достоверность интеллектуального анализа данных. Поэтому вопрос о том, как точно оценить значение null в процессе предварительной обработки данных, является важной темой исследования [4.20]. Перед лицом этой проблемы обычно существует несколько способов ее решения:

1. удаление записей с нулевыми значениями;

2. замена нулевого значения постоянным значением;

3. выбор среднего значения вместо нулевого значения в диапазоне значений null;

4. в диапазоне значений null вместо нулевого значения используется случайное значение;

5. статистическая функция распределения исходных данных, а затем в соответствии с функцией распределения сгенерировать значение замены нулевого значения.

Однако вышеприведенные методы не могут идеально решить все проблемы с нулевыми значениями, процесс вычисления сложен, а тенденция кластеризации исходных данных игнорируется, и эффект оценки нулевых значений может быть получен не очень хорошо [4.21, 4.22].

С этой целью предложено несколько методов оценки нулевых

значений базы данных, в [4.23] предлагается общий метод оценки граничных значений для неопределенных показателей модели данных. В соответствии с характеристиками баз данных о транзакциях с неопределенными весами сначала разрабатывается общая система оценки граничных значений для часто используемых модельных показателей, а затем предлагается метод быстрой оценки верхних границ модельных показателей в соответствии с этой системой. Наконец, оцениваются верхние границы двух типичных модельных показателей, чтобы проиллюстрировать их осуществимость. Экспериментальные результаты показывают, что, хотя этот метод может реализовывать оценку нулевых значений данных, эффект оценки не очень хорош, когда требуется одинаковое время локального поиска и одинаковое время измерения риска. В [4.24] предлагается эффективный метод оценки нулевого значения в реляционных базах данных. Сначала метод анализирует данные в таблице данных, чтобы найти набор атрибутов, связанных с оценочным атрибутом. В этом процессе используются только данные, предоставляемые самими данными. Информация, позволяющая избежать ошибки, вызванной субъективностью, когда эксперт определяет условные атрибуты. Во-вторых, выполняется нечеткая кластеризация в соответствии с полученным набором атрибутов для получения разделения исходных данных, а затем выдается оценочное пустое значение в таблице взаимосвязей на основе оцененного кластера и метода линейной регрессии. Наконец, средняя абсолютная погрешность используется для измерения точности оценки алгоритма. Экспериментальные результаты показывают, что результат применения этого метода имеет высокую точность, но существует проблема низкого эффекта.

Для решения вышеуказанных проблем предлагается основанный на искусственном интеллекте алгоритм оценки нулевых значений для неопределенных баз данных. Этот метод использует принцип ошибки для

определения порядка оценки нулевых значений для каждого столбца. С помощью интеллектуального анализа данных определяется набор атрибутов, связанный с оцененными атрибутами. Исходные данные разбиваются на нечеткие кластеры с помощью метода искусственного интеллекта, а нулевое значение оценивается методом линейной регрессии внутри каждого кластера.

4.2.2. Анализ характеристик неопределенной базы данных

Неопределенные данные - это общий термин, обозначающий данные, которые не имеют полной уверенности в модели данных. В действительности данные являются детерминированными. Причина получения неопределенных данных заключается в ограниченности их собственных знаний, что приводит к существованию неопределенных данных в модели данных. В результате следующие факторы приведут к возникновению неопределенности в данных:

1. при описании реального мира в модели данных;

2. при изменении или преобразовании данных в модели данных;

3. при манипулировании данными в модели данных.

Термин “неопределенные данные” используется для обозначения данных, которые не полностью соответствуют всем моделям обработки данных. В целом, неопределенные данные в основном включают следующие категории:

1. Вероятностные данные: те данные, которые оцениваются как истинные или ложные с определенным значением вероятности, называются вероятностными данными.

2. Неточные данные: такого рода данные доступны в моделях данных, но они не очень понятны. Например, данные могут быть диапазонами или не содержать значений.

3. Нечеткие данные: в модели данных такие данные выражаются

неопределенно в количестве или единицах измерения.

4. Несогласованные данные: данные с разными атрибутами true и false в разное время могут изменяться с течением времени.

5. Неоднозначные данные: некоторые данные в модели данных могут приводить к неоднозначности.

Целью системы баз данных является предоставление пользователям необходимой им информации, а информация, с которой они взаимодействуют, является результатом преобразования описания информации реального мира [4.25]. Иными словами, в центре внимания системы баз данных находится интерактивная информация с пользователями.

С момента своего появления реляционная база данных широко используется в различных областях благодаря своей простой и понятной структуре данных, гибкости, независимости, целостности, меньшей избыточности и удобству применения. Но в практическом применении это также отражает некоторые недостатки реляционной базы данных, такие как введение нулевого значения для решения проблемы неопределенной обработки данных в реальном мире [4.26].

Нулевое значение в реляционной базе данных представляет собой неизвестное значение, которое не является ни числом 0, ни пустой строкой, ни каким-либо другим значимым значением. В ранней базе данных не существовало понятия нулевого значения. Все значения были определены и доступны для понимания. С помощью значения null мы можем выразить данные, которые нам неизвестны, неопределенные данные и, конечно же, данные о человеческих ошибках. Введение значения null делает представление данных в реляционной базе данных более полным и в некоторой степени решает проблему неопределенности [4.27].

4.2.3. Построение модели поиска потерянных данных

Построение модели поиска потерянных данных TIR в основном предназначено для описания имитационного и абстрактного процесса поиска потерянных данных. Прежде всего, технология TIR используется для получения информации из базы данных, а цель поиска заключается в получении информации, тесно связанной с ключевыми словами, за определенный период времени.

Чтобы лучше удовлетворить потребности в поиске потерянной информации, основными целями модели поиска потерянных данных являются определение поиска потерянных данных, определение результатов поиска, расчет релевантности результатов поиска и т.д. В соответствии с характеристиками базы данных, модель поиска потерянных данных определяется как форма из четырех кортежей, которая представлена [Q, D, R, S], где Q представляет информацию запроса; D представляет модель данных; R представляет поиск потерянных данных; S представляет механизм оценки информации запроса и поиска результаты.

С развитием сетевых технологий данные в основном хранятся в базе данных, но для них используется атрибут времени. Чтобы лучше выразить атрибут времени данных, данные в базе данных представлены в виде графика временных данных.

Граф временных данных представлен формулой $G=(V_t, E_t)$, где, V_t представляет набор временных узлов, а E_t представляет набор временных ребер.

Временным узлом является v_t, выраженный как $v_t=[v, (ts_{vt}, te_{vt})]$, где v представляет идентификацию временного узла, $[ts_{vt}, te_{vt}]$ представляет полуоткрытый временной интервал, а E - время действия данных.

Временной интервал e_t выражается как $e_t=[u_t, v_t, (ts', te')]$, [ts', te'] - эффективное время поиска.

Временные данные показаны на рис. 4.6. Как показано на рис. 4.6,

информация имеет время действия и время транзакции. В процессе поиска потерянных данных в основном учитывается время действия информации.

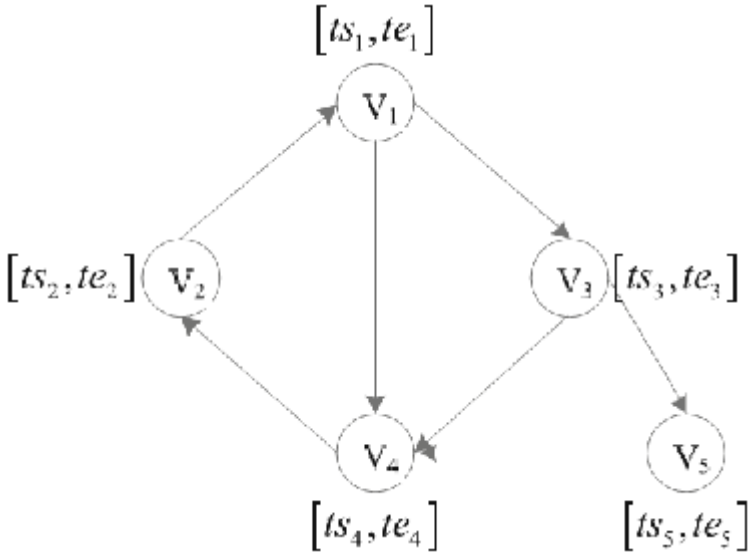


Рис. 4.6. Граф временных данных

На основе модели поиска потерянных данных, в соответствии с временными ограничениями и ключевыми словами в информации запроса, поддерево результатов извлекается из графика временных данных. В общем, в случае запроса с отсутствующими данными мы получим множество поддеревьев результатов запроса. Чтобы получить наиболее похожую информацию о пропущенных данных, мы сортируем поддеревья результатов запроса по принципу вычисления сходства, и первое из них - это поддеревья поиска похожих пропущенных данных. Поэтому очень важно рассчитать сходство информации о потерянных данных.

В общем, чем меньше вес временной границы, тем больше сходство. Конкретный метод расчета веса временной границы заключается в следующем.

Чтобы обеспечить извлечение как можно большего числа узлов с ключевыми словами и обеспечить тесную связь результатов поиска с информацией о потерянных данных в запросе, рассчитаем вес структуры узлов. Вес структуры узла представляет важность узла в графе временных

данных, который вводится в формулу расчета веса ребра, и формула расчета веса ребра получается следующим образом:

$$W(Q, e_t) = \frac{1}{IR_{(k,u)} + IR_{(k,v)}} \times W_e(u, v) \quad (4.5)$$

где $W_e(u, v)$ представляет вес структуры узла.

Временная граница имеет значение своевременности, и веса разных временных границ также различны. Следовательно, в процессе поиска потерянных данных конечное время временной границы должно соответствовать времени запроса потерянных данных. Вес временного ребра устанавливается в соответствии со временем запроса потерянных данных, и формула его расчета такова:

$$W(Q, e_t) = 1 - \frac{|I_c \cap I_t|}{I_c} \quad (4.6)$$

где I_c представляет время запроса потерянных данных, I_t - эффективное время временного интервала.

С помощью приведенной выше формулы получается полное значение веса временного интервала, позволяющее оценить сходство потерянных данных и обеспечить поддержку данных для последующей оценки нулевого значения в базе данных неопределенности.

4.2.4. Алгоритмы оценки пространственных значений, основанные на искусственном интеллекте

При реальном применении базы данных проблема отсутствия данных практически неизбежна, что приводит к проблеме нулевого значения. Оценка нулевого значения стала основным направлением исследований в области обработки нулевых значений, и появилось большое количество методов оценки нулевого значения базы данных [4.28]. Большинство из этих методов используют часть полных данных в таблице базы данных в качестве обучающего набора, извлекают знания из

обучающего набора с помощью машинного обучения или некоторой теории мягких вычислений, выводят правила принятия решений или модели и, наконец, оценивают нулевое значение в соответствии с правилами или моделями.

Существует множество широко используемых алгоритмов оценки нулевого значения, таких как метод грубого набора, метод облачной модели и метод, основанный на генетическом алгоритме. У этих алгоритмов есть свои преимущества, но есть и некоторые очевидные недостатки. Метод приблизительного набора в основном основан на соотношении совместимости между данными, которое заполняется совместимыми значениями кортежей. Однако, если кортеж несовместим с другими кортежами или значения атрибутов, соответствующие совместимым кортежам, отсутствуют, то оценка не может быть дана. Метод облачной модели в основном основан на генерации случайных точек вблизи положения равновесия подчиненным генератором облаков для соответствия исходному распределению данных, что вызовет некоторую “случайность” оценочного значения и повлияет на результаты алгоритма [4.29]. Основным недостатком оценки нулевого значения на основе генетического алгоритма является то, что он требует анализа семантики естественного языка для эффективного кодирования, а алгоритм требует длительного времени итерации и обладает плохой масштабируемостью при большом объеме данных. Для более точной оценки пустого значения в реляционной базе данных на основе модели поиска потерянных данных и сходства потерянных данных предлагается метод оценки пустого значения в базе данных с неопределенностью, основанный на искусственном интеллекте. Конкретный процесс реализации заключается в следующем:

Шаг 1: Выбор объекта и преобразование данных.

1.1. Выбор объекта, алгоритм сокращения атрибутов, основанный на

грубом наборе, используется для уменьшения атрибутов исходной таблицы данных и получения набора ключевых атрибутов после сокращения.

1.2. Преобразование данных, в основном, относится к предварительной обработке данных, что упрощает использование формы данных. Сначала семантические атрибуты естественного языка нумеруются, чтобы их было удобно использовать для интеллектуального анализа данных. Затем используется формула нечеткого числа для нормализации числовой информации и упрощения вычислений [12].

Шаг 2: Кластеризация с искусственным интеллектом.

Наборы ненулевых атрибутов, связанные с атрибутами с нулевыми значениями, полученными на шаге 1, используются для кластеризации. При объединении похожих данных разные данные разбиваются на разные кластеры. На рис. 4.7 показана совместимость объектов из разных наборов атрибутов.

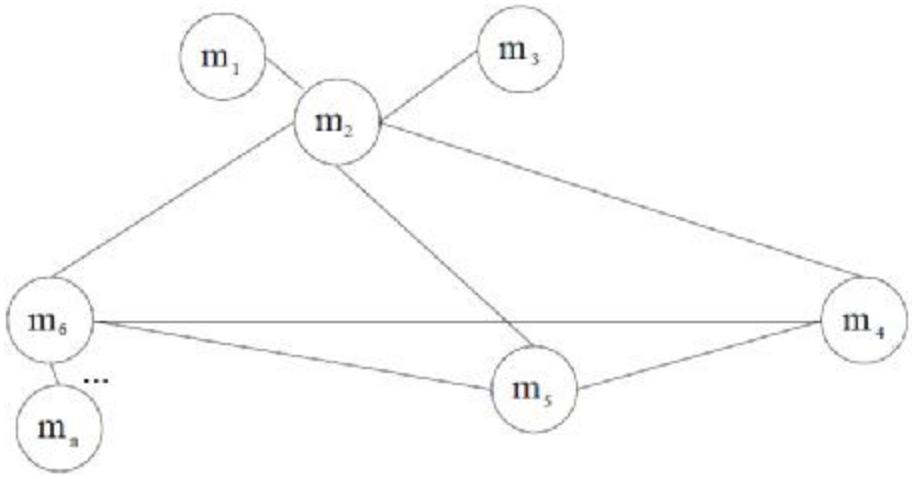


Рис. 4.7. Соотношение совместимости объектов в разных наборах атрибутов

Как показано на рис. 4.7, учитывая, что разные атрибуты имеют разный вес влияния на столбцы с нулевыми значениями, вводятся соответствующие веса:

$$w = \frac{r^2 - W(Q, e_t)¢}{\sum_{k=1}^{m} r_1^2} \qquad (4.7)$$

В формуле m - количество атрибутов в наборе непустых атрибутов, относящихся к атрибутам с нулевыми значениями; r - коэффициенты корреляции атрибутов с нулевыми значениями; w - отношение коэффициентов корреляции атрибутов с нулевыми значениями и сумма коэффициентов корреляции всех связанных атрибутов. и атрибуты с нулевыми значениями, что отражает весомость влияния атрибутов с нулевыми значениями. После кластеризации с использованием искусственного интеллекта будет получен центр кластеризации.

Шаг 3: Вычисление влияния [4.31, 4.32].

После кластеризации данных в несколько кластеров для каждого кластера влияние различных независимых переменных на зависимые переменные различно. Коэффициент регрессии искусственного интеллекта используется для расчета влияния различных независимых переменных на зависимые переменные [4.33]. Сначала коэффициент нечеткой корреляции используется для представления степени корреляции между атрибутами, затем определяется коэффициент независимой переменной и, наконец, получается степень влияния атрибутов.

Формула для расчета степени корреляции выглядит следующим образом:

$$z_{a,b} = \frac{\sum_{i=1}^{n} \left[(a_i - \bar{a}) \left(b_i - \bar{b} \right) \right]}{\sqrt{\sum_{i=1}^{n} (a_i - \bar{a})^2 \sum_{i=1}^{n} \left(b_i - \bar{b} \right)^2}} \qquad (4.8)$$

В формуле $\bar{a}$ и $\bar{b}$ представляют собой выборочное среднее значение a, b нечеткого множества. Формула для определения коэффициента независимой переменной выглядит следующим образом:

$$COD = \pm \frac{r^2}{\sum_{k=1}^{m} r_1^2} \quad (4.9)$$

Шаг 4: Оценка нулевых значений.

Сначала вычисляется евклидово расстояние между кортежем и центром каждого кластера, и для получения оценочного значения используется алгоритм оценки нулевого значения [4.34].

Ниже описан алгоритм оценки нулевого значения для базы данных с неопределенностью, основанный на искусственном интеллекте. Если число записей в таблице данных равно N, то число записей, содержащих нулевые значения, равно N_{null}. Номер ключевого атрибута после сокращения атрибута равен m, а номер кластеризации после разделения равен C. Затем временная сложность алгоритма анализируется следующим образом:

1. Выбор признака и преобразование данных. На этом этапе используется алгоритм сокращения признака, основанный на различимой матрице в грубом наборе, а временная сложность алгоритма составляет $O(N)^2$;

2. Кластеризация с использованием алгоритма искусственного интеллекта для оценки нулевого значения в базе данных с неопределенностью, временная сложность алгоритма равна O(N).

3. Вычисление степени корреляции, чтобы получить общее время, затрачиваемое на вычисление, оценивается линейнюй сложностью [4.35].

4. Вычисление нулевых значений и последующая обработка. На этом этапе оценивается небольшое количество нулевых значений, содержащихся в таблицах базы данных, с временной сложностью O(Nnull), это бесконечно малая величина высокого порядка O(N), которой можно пренебречь [4.36].

Таким образом, алгоритм искусственного интеллекта оценки

нулевого значения в неопределенной базе данных обладает высокой точностью оценки.

4.2.5. Экспериментальный анализ алгоритма оценки нулевого значения в неопределенной базе данных, основанной на искусственном интеллекте

Для проверки рациональности алгоритма оценки нулевого значения в неопределенной базе данных, основанной на искусственном интеллекте, были проведены экспериментальная проверка и анализ.

4.2.5.1. Данные и окружение эксперимента

Окружение эксперимента: Операционная система - Windows 7, процессор Intel Core i7-3770 с тактовой частотой 3,40 ГГц, 8 ГБ оперативной памяти, язык C+ на платформе разработки Microsoft Visual Studio 2012 для реализации алгоритма оценки нулевого значения базы данных неопределенности на основе искусственного интеллекта.

В эксперименте использовался набор данных Algae о водорослях, классический набор данных для интеллектуального анализа данных. Атрибуты приведены в табл. 4.7.

Таблица 4.7

Набор данных Algae

Переменная	Порядок
Сезон	{Весна, Лето, Осень, Зима}
Размер	{Маленький, Средний, Большой}
Скорость	{Низкая, Средняя, Высокая}
PH	Положительное действительное число
CL	Положительное действительное число
NO_3	Положительное действительное число
NH_4	Положительное действительное число
PO_4	Положительное действительное число

В таблице показано, что под влиянием независимых переменных набор данных содержит 184 фрагмента данных, 164 из которых используются в качестве обучающего набора данных, а остальные 20 - в качестве тестового набора.

4.2.5.2. Шаги эксперимента

Эксперименты проводятся на наборах данных о водорослях. Конкретный процесс заключается в следующем:

Шаг 1: Преобразование данных и выбор функции: поскольку сезон, размер и скорость являются текстовыми переменными, их нельзя обрабатывать напрямую, изменив сезонные значения “весна, лето, осень, зима” на “1, 2, 3, 4”; размер “малый, средний, большой” на “1, 2, 3”; скорость от “низкой”, “средней", "высокой" до "1, 2, 3”. Затем мы используем алгоритм выбора объектов, чтобы исключить четыре атрибута: сезон, CL, NH_4 и PO_4. Остальные атрибуты представляют более 95% объектов данных.

Шаг 2: Кластеризация оставшихся атрибутов, использование индикаторов кластеризации для получения лучшего эффекта кластеризации, разделение исходного набора данных и определение центра кластеризации.

Шаг 3: Для каждого типа данных расчет влияния независимых переменных на зависимые переменные.

Шаг 4: Оценка нулевого значения.

На рис. 4.8 представлена структурная схема этапов эксперимента.

4.2.5.3. Результаты эксперимента и их анализ

Чтобы изучить обоснованность алгоритма оценки нулевого значения на основе искусственного интеллекта для неопределенных баз данных, время локального поиска и время измерения риска используются в

качестве критериев для сравнения результатов оценки традиционных справочных алгоритмов [4.23] и алгоритмов на основе искусственного интеллекта.

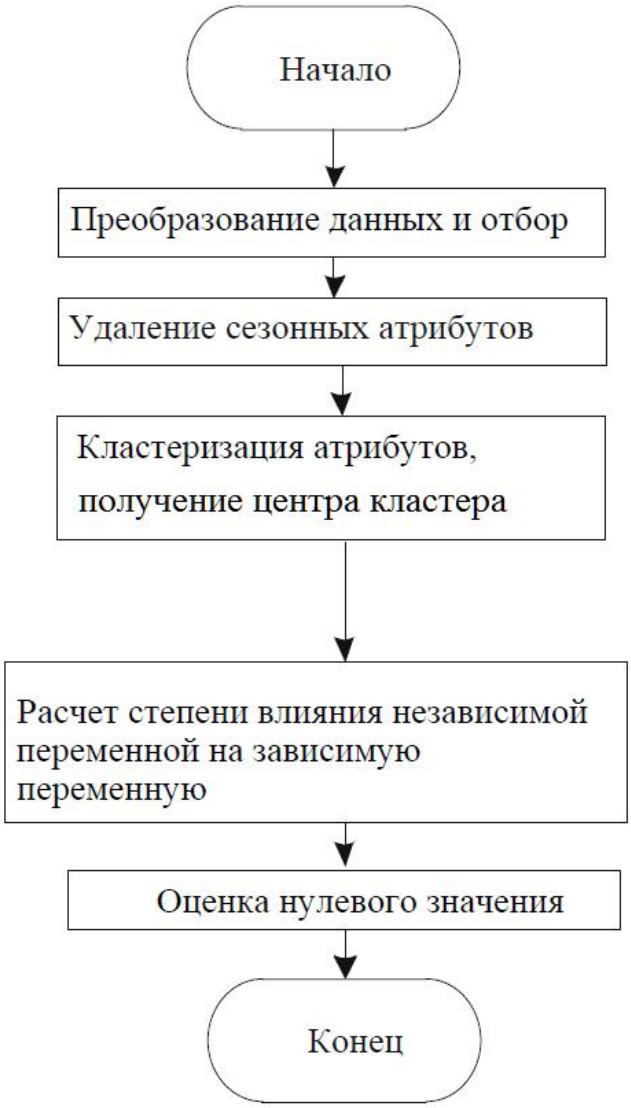


Рис. 4.8. Структурная схема этапов эксперимента

1. Локальное время поиска

Когда время локального поиска остается неизменным, традиционный алгоритм [5] сравнивается с оценочным эффектом, основанным на алгоритме искусственного интеллекта, и результаты показаны на рис. 4.9.

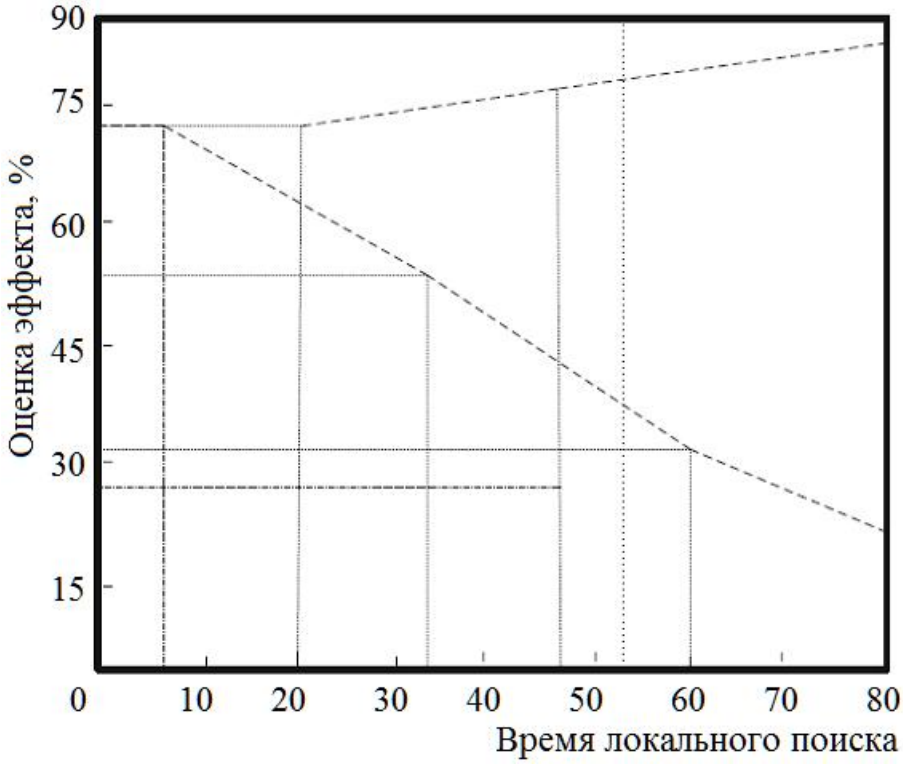


Рис. 4.9. Оценка эффекта двух методов, предоставляющих одинаковое время локального поиска: ---- традиционный алгоритм [5]; ---- алгоритм, основанный на искусственном интеллекте

На рис. 4.9. показано, что алгоритм на основе искусственного интеллекта на 2% лучше, чем традиционный алгоритм [4.23], когда время локального поиска составляет 10 с; алгоритм на основе искусственного интеллекта на 8% лучше, чем традиционный алгоритм [4.23], когда время локального поиска составляет 20 с; алгоритм на основе искусственного интеллекта - эффективность алгоритма, основанного на ссылках, на 20% выше, чем у традиционного алгоритма [4.23], когда время локального поиска составляет 30 с; а эффективность алгоритма, основанного на искусственном интеллекте, на 40 с выше, чем у традиционного алгоритма [4.23], когда время локального поиска составляет 40 с. По сравнению с традиционным эталонным алгоритмом [4.23], алгоритм на основе искусственного интеллекта имеет более высокую оценку эффективности на 32%; алгоритм на основе искусственного интеллекта имеет более

высокую оценку эффективности на 39% при времени локального поиска 50 с; алгоритм на основе искусственного интеллекта имеет более высокую оценку эффективности на 53% при времени локального поиска 60 с. с; алгоритм, основанный на искусственном интеллекте, дает более высокую оценку эффективности на 58% при времени локального поиска 70 с; а алгоритм, основанный на искусственном интеллекте, дает более высокую оценку эффективности на 80 с при времени локального поиска 80 с. Алгоритм, основанный на искусственном интеллекте, на 60% эффективнее традиционного эталонного алгоритма [4.23]. Следовательно, при том же времени локального поиска алгоритм, основанный на искусственном интеллекте, лучше, чем традиционный алгоритм ссылок [4.23], оценивает эффект.

2. Время измерения риска

Когда время измерения риска совпадает, традиционный алгоритм [4.23] сравнивается с эффектом оценки, основанным на алгоритме искусственного интеллекта, и результат показан на рис. 4.10.

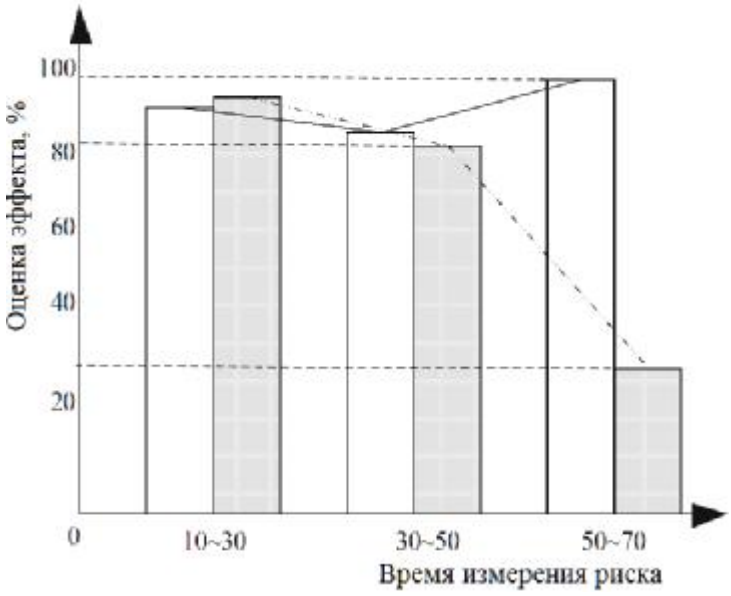


Рис. 4.10. Влияние двух методов оценки на одно и то же время измерения риска: □ - традиционный алгоритм [5]; ■ - алгоритм, основанный на искусственном интеллекте

На рис. 4.10 показано, что при времени измерения риска 10-30 с эффективность оценки традиционного алгоритма [4.23] составляет 90%, а алгоритма на основе искусственного интеллекта - 86%. Когда время измерения риска составляет 30-50 с, эффективность оценки традиционного алгоритма [5] составляет 80%, а алгоритма, основанного на искусственном интеллекте, - 83%. Когда время измерения риска составляет 50-70 с, эффективность оценки традиционного алгоритма [4.23] составляет 28%, а алгоритма, основанного на искусственном интеллекте - 97%. Можно сделать вывод, что при тех же временных условиях измерения риска алгоритм, основанный на искусственном интеллекте, лучше, чем традиционный эталонный алгоритм [4.23], оценивает эффект. Это связано с тем, что метод основан на анализе характеристик неопределенной базы данных, преобразовании данных и использовании технологии искусственного интеллекта для кластеризации данных, а также на расчете степени воздействия и, в конечном счете, на оценке нулевого значения базы данных с помощью вышеуказанных шагов для усиления эффекта оценки нулевого значения.

Подводя итог, можно сказать, что эффект нулевой оценки предложенного алгоритма лучше, чем у традиционного алгоритма из литературы [4.23], при том же времени локального поиска или при тех же условиях измерения риска, что указывает на высокую прикладную ценность предложенного алгоритма.

4.2.5.4. Заключение по эксперименту

Таким образом, алгоритм оценки нулевого значения в базе данных с неопределенностью, основанный на искусственном интеллекте, эффективен. При том же времени локального поиска максимальный эффект оценки алгоритма искусственного интеллекта составляет 86%, а

при том же времени измерения риска максимальный эффект оценки алгоритма искусственного интеллекта составляет 97%.

4.2.6. Итоги рассмотрения алгоритма оценки нулевого значения в базе данных с неопределенностью, основанного на искусственном интеллекте

Недетерминированная база данных основана на строгих математических понятиях. Она имеет единую концепцию и простую и понятную структуру данных. Ее самое большое преимущество заключается в том, что связь между объектами может быть выражена с помощью отношения, то есть неопределенная база данных может описывать саму себя. Множество преимуществ позволяют базе данных indefinite занимать доминирующее положение на рынке и широко использоваться.

С постепенным расширением области применения базы данных uncertain, объем обработки данных и возможности базы данных uncertain востребованы в различных областях. Например, в области научных вычислений, применения датчиков и систем обучения знаниям база данных необходима для обработки неопределенных данных. Однако большинство неопределенных баз данных могут обрабатывать только точные данные, не имея комплексного метода обработки неопределенных данных. В настоящее время единственным способом решения этой проблемы является использование алгоритма искусственного интеллекта для оценки нулевого значения неопределенных баз данных.

Этот метод использует алгоритм искусственного интеллекта для классификации исходных данных, принимая во внимание нечеткий характер классификации данных, и вводит взвешенные значения в соответствии с различными измерениями выборочных данных для учета вклада кластеризации, что делает результаты кластеризации более

точными. Затем модель оценки нулевого значения строится с помощью множественной линейной регрессии, что делает метод оценки нулевого значения более эффективным и точным в среднем на 7.6%.

4.3. Выводы по главе 4

Создан алгоритм обнаружении наилучшей модели машинного обучения для анализа гетерогенных данных, отличающийся применением активного обучения и ансамблевых методах для классификаторов при решении проблем анализа разнородных данных, обеспечивающий сокращение объемов данных и повышение точность анализа совокупности данных.

Разработан алгоритм идентификации нулевых значений в гетерогенных базах данных, отличающийся предварительной классификацией исходных данных на основе взвешенных значений и обеспечивающий более эффективную и точную оценку нулевых значений в среднем на 7.6%..

Источники к главе 4

4.1. Goudjil M, Koudil M, Bedda M, et al. A novel active learning method using SVM for text classification. Int J Automation Comput 2018; 15(3): 290–298.

4.2. Wang Y, Zhou Z, Jin S, et al. Comparisons and selections of features and classifiers for short text classification. IOP Conf Ser Mater Sci Eng 2017; 261: 012018.

4.3. Mohammad AH, Alwada'n T and Al-Momani O. Arabic text categorization using support vector machine, Naive Bayes and neural network. GSTF J Comput (Joc) 2016; 5(1): 108–115.

4.4. Trstenjak B, Mikac S and Donko D. KNN with TF-IDF based framework for text categorization. Proced Eng 2014; 69: 1356–1364.

4.5. Hartmann J, Huppertz J, Schamp C, et al. Comparing automated text classification methods. Int J Res Marketing 2019; 36(1): 20–38.

4.6. Perikos I and Hatzilygeroudis I. Aspect based sentiment analysis in social media with classifier ensembles. In: IEEE/ ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, China, 24–26 May 2017, pp. 273–278.

4.7. Miller B, Linder F and Mebane WR. Active learning approaches for labeling text: review and assessment of the performance of active learning approaches. Polit Anal 2020; 28(4): 532–551.

4.8. Budhi GS, Chiong R, Pranata I, et al. Using machine learning to predict the sentiment of online reviews: a new framework for comparative analysis. Arch Comput Methods Eng 2021; 28(4): 2543–2566.

4.9. Al-Fairouz EI and Al-Hagery MA. The most efficient classifiers for the students' academic dataset. Int J Adv Comput Sci Appl 2020; 11(9): 501–506.

4.10. Onan A, Korukoglu S and Bulut H. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Syst Appl 2016; 62: 1–16.

4.11. Alhaj YA, Wickramaarachchi WU, Hussain A, et al. Efficient feature representation based on the effect of words frequency for Arabic documents classification. In: Proceedings of the 2nd International Conference on Telecommunications and Communication Engineering, 2018, Beijing China, 28-30 November 2018, pp. 397–401.

4.12. Lilleberg J, Zhu Yand Zhang Y. Support vector machines andword2vec for text classification with semantic features. In: IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing, Beijing, China, 6–8 July 2015, pp. 136–140.

4.13. Abdelaal HM, Elmahdy AN, Halawa AA, et al. Improve the automatic classification accuracy for Arabic tweets using ensemble methods. J

Electr Syst Inf Techn 2018; 5(3): 363–370.

4.14. Wang G, Sun J, Ma J, et al. Sentiment classification: the contribution of ensemble learning. Decis Support Systems 2014; 57: 77–93.

4.15. Al-Hagery MA, Al-assaf MA and Al-kharboush FM. Exploration of the best performance method of emotions classification for arabic tweets. Indonesian J Electr Eng Comput Sci 2020; 19(2): 1010–1020.

4.16. Al-Azani S and El-Alfy E-SM. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. Proced Comput Sci 2017; 109: 359–366.

4.17. Xu S, Liang H and Baldwin T. Unimelb at semeval-2016 tasks4a and 4b: An ensemble of neural networks and a word2vec based model for sentiment classification. In: Proceedings of the 10th international workshop on semantic evaluation. SemEval 2016, San Diego, California, 16-17 June 2016, pp. 183–189.

4.18. Luo X. Efficient english text classification using selected machine learning techniques. Alexandria Eng J 2021; 60(3): 3401–3409.

4.19. Kowsari K, Jafari Meimandi K, Heidarysafa M, et al. Text classification algorithms: a survey. Information2019; 10(4):150.

4.20. Tharwat A and Schenck W. Balancing exploration and exploitation: a novel active learner for imbalanced data. Knowledge-Based Syst, 2020; 210: 106500.

4.21. Birman Y, Hindi S, Katz G, et al. Cost-effective ensemble models selection using deep reinforcement learning. Infor Fus 2021; 77: 133–148.

4.22. Wang M, Fu K, Min F, et al. Active learning through label error statistical methods. Knowledge-Based Syst, 2020; 189: 105140.

4.23. Kumar P and Gupta A. Active learning query strategies for classification, regression, and clustering: a survey. J Comput Sci Techn 2020; 35(4): 913–945.

4.24. Brunner U and Stockinger K. Entity matching on unstructureddata: an active learning approach. In: 2019 6th Swiss Conference on Data Science (SDS), Bern, Switzerland, 14–14 June 2019, pp. 97–102.

4.25. Stieglitz S, Mirbabaie M, Ross B, et al. Social media analytics - Challenges in topic discovery, data collection, and data preparation. Int J Info Manage 2018; 39: 156–168.

4.26. Wu J-Y, Hsiao Y-C and Nian M-W. Using supervised machine learning on large-scale online forums to classify course related Facebook messages in predicting learning achievement within the personal learning environment. Interact Learn Environ 2020; 28(1): 65–80.

4.27. Alam M.K., Aziz A.A., Latif S.A. et al. Error-aware data clustering for in-network data reduction in wireless sensor networks// Sensors 2020, 20(4), 1011.

4.28. Zhang T.A. Dynamic threats assessment based on intuitionistic

fuzzy set under missing data condition// Fire Control Command Control 2018, 43(8), 93-97.

4.29. Kim K. Identifying the structure of cities by clustering using a new similarity measure based on smart card data// IEEE Trans. Intell. Transp. Syst. 2020, 21(5), 2002-2011.

4.30. Wang J., Liu F.X., Jin C.J. General bound estimation method for pattern measures over uncertain datasets// J. Comput. Appl. 2018, 38(01), 165-170.

4.31. Liu L., Wang L.S., Wu F. An efficient method for estimating null values in relational database// Comput. Technol. Autom. 2016, 35(03), 110-114.

4.32. Li H. RFID tag number estimation algorithm based on sequential linear Bayes method// J. Comput. Appl. 2018, 38(11), 3287-3292.

4.33. Gao J.M. Adaptive deduplication simulation in privacy protection database// Comput. Simul. 2019, 36(01), 239-242.

4.34. Song X.P. Global estimates of ecosystem service value and change: taking into account uncertainties in satellite-based land cover data// Ecol. Econ. 2018, 143(1), 227-235.

4.35. Faghih M., Mirzaei M., Adamowski J. et al. Uncertainty estimation in flood inundation mapping: an application of non-parametric bootstrapping// River Res. Appl. 2017, 33(4), 611-619.

4.36. Wang Y., Tao W., Yan Z., Wei R. Uncertainty analysis of dynamic thermal rating basedon environmental parameter estimation// EURASIP J. Wirel. Commun. Netw. 2018, 2018(1), 1-10. https://doi.org/10.1186/s13638-018-1181-7.

4.37. Ju H., Zhang G., Cui J. et al. A novel algorithm for pose estimation based on generalizedorthogonal iteration with uncertainty-weighted measuring error of feature points// J. Mod. Opt. 2018, 65(3), 331-341.

4.38. Frédérique S., Bernard C., Paolo D.G. High-resolution humidity profiles retrieved from wind profiler radar measurements// Atmos. Meas. Tech. 2018, 11(3), 1669-1688.

4.39. Fu W., Liu S., Srivastava G. Optimization of big data scheduling in social networks// Entropy 2019, 21(9), 902.

4.40. Forsberg E.M., Huan T., Rinehart D. et al. Data processing, multi-omic pathwaymapping, and metabolite activity analysis using XCMS Online// Nat. Protoc. 2018, 13(4), 633-651.

4.41. Liu S., Liu D., Srivastava G., et al. Overview and methods of correlation filter algorithms in object tracking// Complex Intell. Syst. 2020. https://doi.org/10.1007/s40747-02000161-4

4.42. Koch D.C.L., Jean-Paul G., Xue M. et al. Terahertz frequency modulated continuouswave imaging advanced data processing for art painting analysis// Opt. Express 2018, 26(5), 5358.

4.43. Yang Y.G., Guo X.P., Xu G. et al. Reducing the communication complexity of quantum private database queries by subtle classical post-

processing with relaxed quantum ability// Comput. Secur. 2019, 81(3), 15-24.

4.44. Liu S., Li Z., Zhang Y. et al. Introduction of key problems in long-distance learning and training// Mob. Netw. Appl. 2019, 24(1), 1-4.

Заключение

Целью работы являлась разработка разработка моделей и методов управления гетерогенными данными информационных систем с многомерными атрибутами на основе мягкой максиминной оценки и активного обучения.

В процессе выполнения диссертационного исследования получены следующие основные результаты:

1. Проведен анализ проблем управления гетерогенными данными информационных систем с многомерными атрибутами на основе мягкой максиминной оценки и активного обучения.

2. Разработана мягкая максиминная оценка для гетерогенных данных, содержащих уникальные вариационные компоненты, обеспечивающая сохранение статистических свойств и лучшую вычислительную эффективность

3. Предложена архитектура гетерогенной программной системы, обеспечивающая эффективную редукцию многомерных анализируемых атрибутов.

4. Создан алгоритм выбора наилучшей модели машинного обучения для анализа гетерогенных данных, обеспечивающий сокращение объемов данных и повышение точности анализа совокупности данных.

5. Разработан алгоритм идентификации нулевых значений в гетерогенных базах данных, обеспечивающий более эффективную и точную оценку нулевых значений в среднем на 7.6%.

6. Элементы программного обеспечения зарегистрированы в ФИПС.

Рекомендации и перспективы дальнейшей разработки темы

1. Результаты исследования рекомендуются к применению в задачах управления гетерогенными данными информационных систем с многомерными атрибутами.

2. Дальнейшая разработка темы будет направлена на практическую реализацию теоретических и алгоритмических результатов, интеграцию в наиболее распространенные распределенные системы. Развитие результатов будет направлено на улучшение модифицируемости и реконфигуруемости программных систем.

Список использованных источников

1. Атласов Д.И. Нечеткая энтропия для разнородных данных в гетерогенных информационных системах// Сб. тр. VI Всеросс. НПК «Информационные технологии в экономике и управлении». – Махачкала, 2024. С. 65-70.

2. Атласов Д.И., Васми И., Коптелова А.С., Кочегаров А.В. Оценка и оптимизация систем с гетерогенными данными с учетом показателей эффективности на основе интегрированного алгоритма// Моделирование, оптимизация и информационные технологии. 2025; 13(3). URL: https://moitvivt.ru/ru/journal/pdf?id=2014. DOI: 10.26102/2310-6018/2025.50.3.025.

3. Атласов Д.И., Кравец О. Я. Извлечение надежного сигнала из гетерогенных данных// Современные инновации, системы и технологии - Modern Innovations, Systems and Technologies, 2024, 4(1), 0122–0132. https://doi.org/10.47813/2782-2818-2024-4-1-0122-0132.

4. Атласов Д.И., Кравец О.Я. Идентификация параметров функциональной зависимости прироста ресурсоотдачи процесса решения задач управления информационно-коммуникационными системами// Информационные технологии моделирования и управления, №2(132), 2023. – С. 110-116.

5. Атласов Д.И., Кравец О.Я. Комплексный подход к обработке гетерогенных данных с активным обучением// Информационные технологии моделирования и управления, №4(142), 2025. – С. 304-312.

6. Атласов Д.И., Кравец О.Я. Разработка системы обнаружения аномалий на основе изолированного леса// Интеллектуальные информационные системы: тр. Междунар. НПК, посв. 40-летию кафедры САПРИС. - Воронеж, 2024. – с. 180-184.

7. Атласов Д.И., Кравец О.Я. Экспериментальное исследование по устранению неопределенности данных в гетерогенных информационных системах// Экономика и менеджмент систем управления, №1(55), 2025. – С. 22-31.

8. Атласов Д.И., Кравец О.Я. Экспериментальное исследование по устранению неопределенности данных в гетерогенных информационных системах// Экономика и менеджмент систем управления, №1(55), 2025. – С. 22-31.

9. Атласов Д.И., Кравец О.Я. Эффективный метод получения необходимой информации из набора данных, представляющих собой совокупность неоднородных групп// Оптимизация и моделирование в автоматизированных системах: тр. Междунар. молодежной научной школы. – Воронеж: ВГТУ, 2023. С. 66-70.

10. Атласов Д.И., Кравец О.Я., Пашкевич А.С. Управление

специальной информацией в беспроводной сети на основе гетерогенной последовательности данных// Системы управления и информационные технологии, №1(91), 2023. – С. 49-55.

11. Атласов Д.И., Красновский Е.Е., Сараев П.В. Пути интерполяции мягкой максиминной оценки для гетерогенных данных// Системы управления и информационные технологии, №4(94), 2023. С. 27-30.

12. Атласов Д.И., Сотников Д.В., Васми Ихаб А Васми, Хуссейн Али Иед, Линкина А.В. Типовой интерфейс облачных вычислений. Свидетельство о регистрации программы для ЭВМ № 2025681822 от 18.08.2025. - М.: Роспатент, 2025.

13. Атласов Д.И., Сотников Д.В., Кравец О.Я., Красновский Е.Е. Оценка неопределенности нулевых значений базы данных на основе искусственного интеллекта// Системы управления и информационные технологии, №2.1(100), 2025. С. 4-11.

14. Бостром Н. Искусственный интеллект. Возможные пути, опасности и стратегии. – М.: Манн, Иванов и Фербер, 2015.

15. Бринк Х., Ричардс Дж., Феверолф М. Машинное обучение. – СПб.: Питер, 2017.

16. Бурнаев Е. и др. Многодисциплинарная оптимизация, анализ данных и автоматизация инженерных расчетов с помощью программного комплекса pSeven// CAD/CAM/CAE Observer #4 (88), 2014.

17. Вьюгин В. Математические основы машинного обучения и прогнозирования. - М.: МЦНМО, 2013

18. Маркус Г. Искусственный интеллект: перезагрузка: Как создать машинный разум, которому действительно можно доверять – М.: Альпина Диджитал, 2021.

19. Сильвер Н. Сигнал и Шум. Почему одни прогнозы сбываются, а другие — нет. – М.: Азбука-Аттикус, КоЛибри, 2015.

20. Сотников Д.В., Атласов Д.И., Кравец О.Я., Красновский Е.Е. Исследование метода оптимизации данных для эксплуатации и сопровождения базы знаний программного обеспечения на основе облачных вычислений// Системы управления и информационные технологии, №2(100), 2025. С. 37-42.

21. Хей Дж. Введение в методы байесовского статистического вывода. — М.: Финансы и статистика, 1987.

22. Abdelaal HM, Elmahdy AN, Halawa AA, et al. Improve the automatic classification accuracy for Arabic tweets using ensemble methods. J Electr Syst Inf Techn 2018; 5(3): 363–370.

23. Ai Q., Azizi V., Chen X., Zhang Y. 2018. Learning heterogeneous knowledge base embeddings for explainable recommendation// Algorithms 11 (9), 137.

24. Alam M.K., Aziz A.A., Latif S.A. et al. Error-aware data clustering

for in-network data reduction in wireless sensor networks// Sensors 2020, 20(4), 1011.

25. Al-Azani S and El-Alfy E-SM. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. Proced Comput Sci 2017; 109: 359–366.

26. Al-Fairouz EI and Al-Hagery MA. The most efficient classifiers for the students' academic dataset. Int J Adv Comput Sci Appl 2020; 11(9): 501–506.

27. Al-Hagery MA, Al-assaf MA and Al-kharboush FM. Exploration of the best performance method of emotions classification for arabic tweets. Indonesian J Electr Eng Comput Sci 2020; 19(2): 1010–1020.

28. Alhaj YA, Wickramaarachchi WU, Hussain A, et al. Efficient feature representation based on the effect of words frequency for Arabic documents classification. In: Proceedings of the 2nd International Conference on Telecommunications and Communication Engineering, 2018, Beijing China, 28-30 November 2018, pp. 397–401.

29. Antoine B. et al. 2013. Translating embeddings for modeling multi-relational data// Adv. Neural Inf. Process. Syst. 26.

30. Atlasov D.I. An experimental study of an integrated approach to heterogeneous data processing with active learning// Modern informatization problems in simulation and social technologies (MIP-2026'SCT): Proc. of the XXXI-th Int. Open Science Conf. - Yelm, WA, USA: Science Book Publishing House, 2026. – pp. 20-28.

31. Atlasov D.I., Kravets O.Ja. Computational features of soft maximin estimation interpolation for heterogeneous data// Modern informatization problems in the technological and telecommunication systems analysis and synthesis (MIP-2024'AS): Proceedings of the XXIX-th International Open Science Conference. - Yelm, WA, USA: Science Book Publishing House, 2024. Pp. 120-124.

32. Atlasov D.I., Kravets O.Ja. Theoretical foundations for measuring the uncertainty of heterogeneous data// Modern informatization problems in simulation and social technologies (MIP-2025'SCT): Proc. of the XXX-th Int. Open Science Conf. - Yelm, WA, USA: Science Book Publishing House, 2025. – pp. 11-23.

33. Atlasov D.I., Kravets O.Ja. To the formulation of the problem of extracting a common signal from heterogeneous data of heterogeneous information systems// Modern informatization problems in simulation and social technologies (MIP-2023'SCT): Proceedings of the XXVIII-th International Open Science Conference (Yelm, WA, USA, January 2023). - Yelm, WA, USA: Science Book Publishing House, 2023. – Pp. 8-13.

34. Bao J., Zheng Yu, Wilkie D., Mokbel M. 2015. Recommendations in location-based social networks: a survey// GeoInformatica 19 (3), 525–565.

35. Beaubouef T, Petry FE, Arora G (1998) Information-theoretic measures of uncertainty for rough sets and rough relational databases. Inf Sci 109:185–195.

36. Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1), 183–202.

37. Birman Y, Hindi S, Katz G, et al. Cost-effective ensemble models selection using deep reinforcement learning. Infor Fus 2021; 77: 133–148.

38. Bishop C. Pattern Recognition and Machine Learning. - Springer, 2006.

39. Brunner U and Stockinger K. Entity matching on unstructureddata: an active learning approach. In: 2019 6th Swiss Conference on Data Science (SDS), Bern, Switzerland, 14–14 June 2019, pp. 97–102.

40. Bryan P., Al-Rfou R., Skiena S. 2014. Deepwalk: online learning of social representations// Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 701–710.

41. Budhi GS, Chiong R, Pranata I, et al. Using machine learning to predict the sentiment of online reviews: a new framework for comparative analysis. Arch Comput Methods Eng 2021; 28(4): 2543–2566.

42. Bühlmann, P., & Meinshausen, N. (2016). Magging: Maximin aggregation for inhomogeneous large-scale data. Proceedings of the IEEE, 104(1), 126–135.

43. Buis, P. E., & Dyksen, W. R. (1996). Efficient vector and parallel manipulation of tensor products. ACM Transactions on Mathematical Software (TOMS), 22(1), 18–23.

44. Burke R., Vahedian F., Mobasher B. 2014. Hybrid recommendation in heterogeneous networks// Int. Conf. on User Modeling, Adaptation, and Personalization. Springer, pp. 49–60.

45. Cai D., Qian S., Quan F., Xu C. 2021. Heterogeneous hierarchical feature aggregation network for personalized micro-video recommendation// IEEE Trans. Multimed. 24, 805–818.

46. Cai X., Han J., Yang L. 2018. Generative adversarial network based heterogeneous bibliographic network representation for personalized citation recommendation// Thirty-second AAAI Conf. on Artificial Intelligence.

47. Chang J. et al. 2020. Bundle recommendation with graph convolutional networks// Proc. of the 43rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 1673–1676.

48. Chen C. et al. 2021. Graph heterogeneous multi-relational recommendation// Proc. of the AAAI Conf. on Artificial Intelligence, 35, pp. 3958–3966.

49. Chen J. et al. 2019. Citation recommendation based on weighted heterogeneous information network containing semantic linking// 2019 IEEE Int. Conf. on Multimedia and Expo (ICME). IEEE, pp. 31–36.

50. Chen L. et al. 2018. Heterogeneous neural attentive factorization machine for rating prediction// Proc. of the 27th ACM Int. Conf. on Information and Knowledge Management, pp. 833–842.

51. Chen Li et al. 2021. Sequence-aware heterogeneous graph neural collaborative filtering// Proc. of the 2021 SIAM Int. Conf. on Data Mining (SDM). SIAM, pp. 64–72.

52. Chen Li, et al. 2021. Package recommendation with intra- and inter-package attention networks// The 44th Int. ACM SIGIR Conf. on Research & Development in Information Retrieval.

53. Chen YM, Wu KS, Chen XH, Tang CH, Zhu QX (2014) An entropy-based uncertainty measurement approach in neighborhood systems. Inf Sci 279:239–250.

54. Chen, X., Lu, Z., & Pong, T. K. (2016). Penalty methods for a class of non-lipschitz optimization problems. SIAM Journal on Optimization, 26(3), 1465–1492.

55. Currie, I. D., Durban, M., & Eilers, P. H. (2006). Generalized linear array models with applications to mul-tidimensional smoothing. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(2), 259–280.

56. Dai JH, Wang WT, Hao QX, Tian W (2012) Uncertainty measurement for interval-valued decision systems based on extended conditional entropy. Knowl-Based Syst 27:443–450.

57. De Boor, C. (1979). Efficient computer manipulation of tensor products. ACM Transactions on Mathematical Software (TOMS), 5(2), 173–182.

58. Ditsch I, Gediga G (1998) Uncertainty measures of rough set prediction. Artif Intell 106(1):109–137.

59. Dong Y., Chawla N.V., Swami A. 2017. metapath2vec: scalable representation learning for heterogeneous networks// Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 135–144.

60. Dunn OJ (1961) Multiple comparisons among means. J Am Stat Assoc 56(293):52–64.

61. Faghih M., Mirzaei M., Adamowski J. et al. Uncertainty estimation in flood inundation mapping: an application of non-parametric bootstrapping// River Res. Appl. 2017, 33(4), 611-619.

62. Fan S. et al. 2019. Metapath-guided heterogeneous graph neural network for intent recommendation// Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 2478–2486.

63. Fanaee T. H., & Gama, J. (2023). Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence, 2(2), 113–127.

64. Feng W., Wang J. 2012. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems// Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp.

1276–1284.

65. Forsberg E.M., Huan T., Rinehart D. et al. Data processing, multi-omic pathwaymapping, and metabolite activity analysis using XCMS Online// Nat. Protoc. 2018, 13(4), 633-651.

66. Fouss F. et al. 2007. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation// IEEE Trans. Knowl. Data Eng. 19 (3), 355–369.

67. Frédérique S., Bernard C., Paolo D.G. High-resolution humidity profiles retrieved from wind profiler radar measurements// Atmos. Meas. Tech. 2018, 11(3), 1669-1688.

68. Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat 11(1):86–92.

69. Fu T.-y., Lee W.-C., Lei Z. 2017. Hin2vec: explore meta-paths in heterogeneous information networks for representation learning// Proc. of the 2017 ACM on Conf. on Information and Knowledge Management, pp. 1797–1806.

70. Fu W., Liu S., Srivastava G. Optimization of big data scheduling in social networks// Entropy 2019, 21(9), 902.

71. Gao J.M. Adaptive deduplication simulation in privacy protection database// Comput. Simul. 2019, 36(01), 239-242.

72. Goudjil M, Koudil M, Bedda M, et al. A novel active learning method using SVM for text classification. Int J Automation Comput 2018; 15(3): 290–298.

73. Grinvald, A., & Bonhoeffer, T. (2002). Optical imaging of electrical activity based on intrinsic signals and on voltage sensitive dyes: The methodology.

74. Grover A., Leskovec J. 2016. node2vec: scalable feature learning for networks// Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 855–864.

75. Guerraoui R. et al. 2017. Heterogeneous recommendations: what you might like to read after watching interstellar// Proc. of the VLDB Endowment, 10, pp. 1070–1081, 10.

76. Guo C., Liu X. 2015. Automatic feature generation on heterogeneous graph for music recommendation// Proc. of the 38th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 807–810.

77. Guy I. et al. 2010. Social media recommendation based on people and tags// Proc. of the 33rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 194–201.

78. Han X. et al. 2018. Aspect-level deep collaborative filtering via heterogeneous information networks// IJCAI, pp. 3393–3399.

79. Han X. et al. 2018. Representation learning with depth and breadth for recommendation using multi-view data// Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Int. Conf. on Web and Big Data.

Springer, pp. 181–188.

80. Hartmann J, Huppertz J, Schamp C, et al. Comparing automated text classification methods. Int J Res Marketing 2019; 36(1): 20–38.

81. He X. et al. 2017. Neural collaborative filtering// Proc. of the 26th Int. Conf. on World Wide Web, pp. 173–182.

82. Hu B. et al. 2018. Leveraging meta-path based context for top-n recommendation with a neural co-attention model// Proc. of the 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 1531–1540.

83. Hu J. et al. 2016. Recexp: a semantic recommender system with explanation based on heterogeneous information network// Proc. of the 10th ACM Conf. on Recommender Systems, pp. 401–402.

84. Hu QH, Yu DR, Liu JF, Wu CX (2008) Neighborhood rough set based heterogeneous feature subset selection. Inf Sci 178:3577–3594.

85. Hu QH, Yu DR, Xie ZX (2006) Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recogn Lett 27(5):414–423.

86. Jamali M., Lakshmanan L. 2013. Heteromf: recommendation in heterogeneous information networks using context dependent factor models// Proc. of the 22nd Int. Conf. on World Wide Web, pp. 643–654.

87. Jannach D., Zanker M., Felfernig A., Gerhard F. 2010. Recommender Systems - an Introduction. - Cambridge University Press.

88. Jeh G., Widom J. 2002. Simrank: a measure of structural-context similarity// Proc. of the Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 538–543.

89. Jeh G., Widom J. 2003. Scaling personalized web search// Proc. of the 12th Int. Conf. on World Wide Web, pp. 271–279.

90. Ji H. et al. 2021. Large-scale comb-k recommendation// Proc. of Web Conf. 2021, pp. 2512–2523.

91. Ji H. et al. 2021. Who you would like to share with? a study of share recommendation in social e-commerce// Proc. of the AAAI Conf. on Artificial Intelligence, vol. 35, pp. 232–239.

92. Ji Y. et al. 2020. Temporal Heterogeneous Interaction Graph Embedding for Next-Item Recommendation// Lecture Notes in Computer Science, vol 12459. Springer, Cham. https://doi.org/10.1007/978-3-030-67664-3_19.

93. Jiang Z. et al. 2018b. Cross-language citation recommendation via hierarchical representation learning on heterogeneous graph// The 41st Int. ACM SIGIR Conf. on Research & Development in Information Retrieval, pp. 635–644.

94. Jin J. et al. 2020. An efficient neighborhood-based interaction model for recommendation on heterogeneous graph// Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 75–84.

95. Jin L. et al. 2020. Heterogeneous graph embedding for cross-domain recommendation through adversarial learning// Int. Conf. on Database Systems

for Advanced Applications. Springer, pp. 507–522.

96. Ju H., Zhang G., Cui J. et al. A novel algorithm for pose estimation based on generalizedorthogonal iteration with uncertainty-weighted measuring error of feature points// J. Mod. Opt. 2018, 65(3), 331-341.

97. Kim K. Identifying the structure of cities by clustering using a new similarity measure based on smart card data// IEEE Trans. Intell. Transp. Syst. 2020, 21(5), 2002-2011.

98. Kipf T.N., Welling M. 2016. Semi-Supervised Classification with Graph Convolutional Networks. - arXiv preprint arXiv:1609.02907.

99. Koch D.C.L., Jean-Paul G., Xue M. et al. Terahertz frequency modulated continuouswave imaging advanced data processing for art painting analysis// Opt. Express 2018, 26(5), 5358.

100. Kowsari K, Jafari Meimandi K, Heidarysafa M, et al. Text classification algorithms: a survey. Information2019; 10(4):150.

101. Kravets O.Ja., Atlasov D.I. et al. Designing the architecture of a distributed system for information monitoring of IoT and IIoT infrastructures traffic// International Journal on Information Technologies and Security, vol. 16, no. 1, 2024, pp. 49-56. https://doi.org/10.59035/BTBI7690.

102. Kumar P and Gupta A. Active learning query strategies for classification, regression, and clustering: a survey. J Comput Sci Techn 2020; 35(4): 913–945.

103. Lee S. et al. 2013. Pathrank: ranking nodes on a heterogeneous graph for flexible hybrid recommender systems// Expert Syst. Appl. 40 (2), 684–697.

104. Li H. RFID tag number estimation algorithm based on sequential linear Bayes method// J. Comput. Appl. 2018, 38(11), 3287-3292.

105. Li ZW, Huang D, Liu XF, Xie NX, Zhang GQ (2020a) Information structures in a covering information system. Inf Sci 507:449–471.

106. Li ZW, Liu XF, Dai JH, Chen JL, Fujita H (2020b) Measures of uncertainty based on Gaussian kernel for a fully fuzzy information system. Knowl-Based Syst 196:105791.

107. Li ZW, Liu YY, Li QG, Qin B (2016) Relationships between knowledge bases and related results. Knowl Inf Syst 49:171–195.

108. Li ZW, Zhang GQ, Wu WZ, Xie NX (2020c) Measures of uncertainty for knowledge bases. Knowl Inf Syst 62:611–637.

109. Li ZW, Zhang PF, Ge X, Xie NX, Zhang GQ, Wen CF (2019) Uncertainty measurement for a fuzzy relation information system. IEEE Trans Fuzzy Syst 27(12):2338–2352.

110. Liang JY, Qian YH (2008) Information granules and entropy theory in information systems. Sci China (Ser F) 51:1427–1444.

111. Lilleberg J, Zhu Yand Zhang Y. Support vector machines andword2vec for text classification with semantic features. In: IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing,

Beijing, China, 6–8 July 2015, pp. 136–140.

112. Liu L., Wang L.S., Wu F. An efficient method for estimating null values in relational database// Comput. Technol. Autom. 2016, 35(03), 110-114.

113. Liu S. et al. 2020. A heterogeneous graph neural model for cold-start recommendation// Proc. of the 43rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 2029–2032.

114. Liu S., Li Z., Zhang Y. et al. Introduction of key problems in long-distance learning and training// Mob. Netw. Appl. 2019, 24(1), 1-4.

115. Liu S., Liu D., Srivastava G., et al. Overview and methods of correlation filter algorithms in object tracking// Complex Intell. Syst. 2020. https://doi.org/10.1007/s40747-02000161-4

116. Liu X. et al. 2014. Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation// Proc. of the 23rd Acm Int. Conf. on Conf. on Information and Knowledge Management, pp. 121–130.

117. Liu Z. et al. 2020. Basconv: aggregating heterogeneous interactions for basket recommendation with graph convolutional neural network// Proc. of the 2020 SIAM Int. Conf. on Data Mining. SIAM, pp. 64–72.

118. Lu C.-T. et al. 2016. Item Recommendation for Emerging Online Businesses// IJCAI, pp. 3797–3803.

119. Lu Y. et al. 2020. Social Influence Attentive Neural Network for Friend-Enhanced Recommendation// Lecture Notes in Computer Science, vol 12460. Springer, Cham. https://doi.org/10.1007/978-3-030-67667-4_1.

120. Lund, A. (2018). glamlasso: Penalization in large scale generalized linear array models. R package version 3.0.

121. Lund, A. (2021). SMME: Soft maximin estimation for large scale heterogeneous data. R package version 1.0.1.

122. Lund A., Mogensen S.W., Hansen N.R. Soft Maximin Estimation for Heterogeneous Data, 2022. - https://arxiv.org/pdf/1805.02407.

123. Luo C. et al. 2014. Hete-cf: social-based collaborative filtering recommendation using heterogeneous relations// 2014 IEEE Int. Conf. on Data Mining. IEEE, pp. 917–922.

124. Luo X. Efficient english text classification using selected machine learning techniques. Alexandria Eng J 2021; 60(3): 3401–3409.

125. MacKay D.J.C. Information Theory, Inference, and Learning Algorithms. - Cambridge University Press, 2003.

126. Meinshausen, N., & Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. The Annals of Statistics, 43(4), 1801–1830.

127. Miller B, Linder F and Mebane WR. Active learning approaches for labeling text: review and assessment of the performance of active learning approaches. Polit Anal 2020; 28(4): 532–551.

128. Mohammad AH, Alwada'n T and Al-Momani O. Arabic text categorization using support vector machine, Naive Bayes and neural network.

GSTF J Comput (Joc) 2016; 5(1): 108–115.

129. Nandanwar S. et al. 2018. Fusing diversity in recommendations in heterogeneous information networks// Proc. of the Eleventh ACM Int. Conf. on Web Search and Data Mining, pp. 414–422.

130. Nguyen P.T.-A. et al. 2016. A general recommendation model for heterogeneous networks. IEEE Trans. Knowl. Data Eng. 28 (12), 3140–3153.

131. Ni L., Cohen W.W. 2010. Relational retrieval using a combination of path-constrained random walks// Mach. Learn. 81 (1), 53–67.

132. Niu X. et al. 2020. A dual heterogeneous graph attention network to improve long-tail performance for shop search in e-commerce// Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 3405–3415.

133. Niu X. et al. 2021. A Dual Heterogeneous Graph Attention Network to Improve Long-Tail Performance for Shop Search in E-Commerce// KDD '20: Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining. – P. 3405 – 3415.

134. Onan A, Korukoglu S and Bulut H. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Syst Appl 2016; 62: 1–16.

135. Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht.

136. Pawlak Z, Skowron A (2007) Rough sets and Boolean reasoning. Inf Sci 177:41–73.

137. Pawlak Z, Skowron A (2007) Rough sets: some extensions. Inf Sci 177:28–40.

138. Pawlak Z, Skowron A (2007) Rudiments of rough sets. Inf Sci 177:3–27.

139. Perikos I and Hatzilygeroudis I. Aspect based sentiment analysis in social media with classifier ensembles. In: IEEE/ ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, China, 24–26 May 2017, pp. 273–278.

140. Ren X. et al. 2014. Cluscite: effective citation recommendation by information network-based clustering// Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 821–830.

141. Resnick P., Varian H.R., 1997. Recommender systems// Commun. ACM 40 (3), 56–58.

142. Roland, P. E., Hanazawa, A., Undeman, C., Eriksson, D., Tompa, T., Nakamura, H., Valentiniene, S., & Ahmed, B. (2006). Cortical feedback depolarization waves: A mechanism of top-down influence on early visual areas. Proceedings of the National Academy of Sciences, 103(33), 12586–12591.

143. Roll, J. (2008). Piecewise linear solution paths with application to direct weight optimization. Automatica, 44(11), 2732–2737.

144. Rothenhäusler, D., Meinshausen, N., Bühlmann, P., & Peters, J.

(2021). Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2), 215–246.

145. Sanchez MA, Castro JR, Castillo O, Mendoza O, Rodriguez-Diaz A, Melin P (2017) Fuzzy higher type information granules from an uncertainty measurement. Granul Comput 2:95–103.

146. Schafer J.B., Konstan J.A., Riedl J. 2001. E-commerce recommendation applications// Data Min. Knowl. Discov. 5 (1), 115–153.

147. Schlichtkrull M. et al. 2018. Modeling relational data with graph convolutional networks// European Semantic Web Conf.. Springer, pp. 593–607.

148. Shannon C (1948) A mathematical theory of communication. Bell Syst Techn J 27:379–423.

149. Shi C. et al. 2012. Heterecom: a semantic-based recommendation system in heterogeneous networks// Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1552–1555.

150. Shi C. et al. 2014. Hetesim: a general framework for relevance measure in heterogeneous networks// IEEE Trans. Knowl. Data Eng. 26 (10), 2479–2492.

151. Shi C. et al. 2015. Semantic path based personalized recommendation on weighted heterogeneous information networks// Proc. of the 24th ACM Int. on Conf. on Information and Knowledge Management, pp. 453–462.

152. Shi C. et al. 2016. A survey of heterogeneous information network analysis// IEEE Trans. Knowl. Data Eng. 29 (1), 17–37.

153. Shi C. et al. 2016. Integrating heterogeneous information via flexible regularization framework for recommendation// Knowl. Inf. Syst. 49 (3), 835–859.

154. Shi C. et al. 2018. Heterogeneous information network embedding for recommendation// IEEE Trans. Knowl. Data Eng. 31 (2), 357–370.

155. Singh A.P., Gordon J.G. 2008. Relational learning via collective matrix factorization// Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 650–658.

156. Song X.P. Global estimates of ecosystem service value and change: taking into account uncertainties in satellite-based land cover data// Ecol. Econ. 2018, 143(1), 227-235.

157. Stieglitz S, Mirbabaie M, Ross B, et al. Social media analytics - Challenges in topic discovery, data collection, and data preparation. Int J Info Manage 2018; 39: 156–168.

158. Su Y. et al. 2019. Hrec: heterogeneous graph embedding-based personalized point-of-interest recommendation// Int. Conf. on Neural Information Processing. Springer, pp. 37–49.

159. Sun BZ, Ma WM, Chen DG (2014) Rough approximation of a fuzzy concept on a hybrid attribute information system and its uncertainty measure. Inf Sci 284:60–80.

160. Sun Y. et al. 2011. Pathsim: meta path-based top-k similarity search in heterogeneous information networks// Proc. of the VLDB Endowment, 4, pp. 992–1003, 11.

161. Sun Y., Yu Y., Han J. 2009. Ranking-based clustering of heterogeneous information networks with star network schema// Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 797–806.

162. Sun Z. et al. 2018. Recurrent knowledge graph embedding for effective recommendation// Proc. of the 12th ACM Conf. on Recommender Systems, pp. 297–305.

163. Sun Z. et al. 2019. Research commentary on recommendations with side information: a survey and research directions// Electron. Commer. Res. Appl. 37, 100879.

164. Tang J. et al. 2015. Pte: predictive text embedding through large-scale heterogeneous text networks// Proc. of the 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1165–1174.

165. Tharwat A and Schenck W. Balancing exploration and exploitation: a novel active learner for imbalanced data. Knowledge-Based Syst, 2020; 210: 106500.

166. Trstenjak B, Mikac S and Donko D. KNN with TF-IDF based framework for text categorization. Proced Eng 2014; 69: 1356–1364.

167. Tseng, P., & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming, 117(1-2), 387–423.

168. Tseng, P., & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization.

169. Velickovic P. et al. 2017. Graph Attention Networks// arXiv preprint arXiv:1710.10903.

170. Vijaikumar M., Shevade S,, Narasimha M.M. 2020. Gamma: a graph and multi-view memory attention mechanism for top-n heterogeneous recommendation. Adv. Knowl. Discov// Data Mining 28, 12084.

171. Wang CZ, Huang Y, Shao MW, Chen DG (2019) Uncertainty measures for general fuzzy relations. Fuzzy Sets Syst 360:82–96.

172. Wang G, Sun J, Ma J, et al. Sentiment classification: the contribution of ensemble learning. Decis Support Systems 2014; 57: 77–93.

173. Wang J., Liu F.X., Jin C.J. General bound estimation method for pattern measures over uncertain datasets// J. Comput. Appl. 2018, 38(01), 165-170.

174. Wang M, Fu K, Min F, et al. Active learning through label error statistical methods. Knowledge-Based Syst, 2020; 189: 105140.

175. Wang X. et al. 2019. Heterogeneous graph attention network// The World Wide Web Conf., pp. 2022–2032.

176. Wang XD, Song YF (2018) Uncertainty measure in evidence theory with its applications. Appl Intell 48:1672–1688.

177. Wang Y, Zhou Z, Jin S, et al. Comparisons and selections of features and classifiers for short text classification. IOP Conf Ser Mater Sci Eng 2017; 261: 012018.

178. Wang Y. et al. 2020. Disenhan: disentangled heterogeneous graph attention network for recommendation// Proc. of the 29th ACM Int. Conf. on Information & Knowledge Management, pp. 1605–1614.

179. Wang Y., Tao W., Yan Z., Wei R. Uncertainty analysis of dynamic thermal rating basedon environmental parameter estimation// EURASIP J. Wirel. Commun. Netw. 2018, 2018(1), 1-10. https://doi.org/10.1186/s13638-018-1181-7.

180. Wang Z. et al. 2019. Unified embedding model over heterogeneous information network for personalized recommendation// IJCAI, pp. 3813–3819.

181. Wang, Do., Xu G., Deng S. 2017. Music recommendation via heterogeneous information graph embedding// 2017 Int. Joint Conf. on Neural Networks (IJCNN). IEEE, pp. 596–603.

182. Wright, S.J., Nowak, R.D., & Figueiredo, M.A. (2009). Sparse reconstruction by separable approximation. IEEE Transactionson Signal Processing, 57(7), 2479–2493.

183. Wu C. et al. 2021. User-as-graph: User Modeling with Heterogeneous Graph Pooling for News Recommendation// Proc. of the Thirtieth Int. Joint Conf. on Artificial Intelligence. 10.24963/ijcai.2021/224

184. Wu J-Y, Hsiao Y-C and Nian M-W. Using supervised machine learning on large-scale online forums to classify course related Facebook messages in predicting learning achievement within the personal learning environment. Interact Learn Environ 2020; 28(1): 65–80.

185. Xiao Y. et al. 2013. Collaborative filtering with entity similarity regularization in heterogeneous information networks// IJCAI HINA 27.

186. Xiao Y. et al. 2013. Recommendation in heterogeneous information networks with implicit user feedback// Proc. of the 7th ACM Conf. on Recommender Systems, pp. 347–350.

187. Xiao Y. et al. 2014. Personalized entity recommendation: a heterogeneous information network approach// Proc. of the 7th ACM Int. Conf. on Web Search and Data Mining, pp. 283–292.

188. Xie NX, Liu M, Li ZW, Zhang GQ (2019) New measures of uncertainty for an interval-valued information system. Inf Sci 470:156–174.

189. Xu F. et al. 2019. Relation-aware graph convolutional networks for agent-initiated social e-commerce recommendation// Proc. of the 28th ACM Int. Conf. on Information and Knowledge Management, pp. 529–538.

190. Xu S, Liang H and Baldwin T. Unimelb at semeval-2016 tasks4a and 4b: An ensemble of neural networks and a word2vec based model for sentiment classification. In: Proceedings of the 10th international workshop on semantic

evaluation. SemEval 2016, San Diego, California, 16-17 June 2016, pp. 183–189.

191. Yang L., Zhang Z., Cai X., Guo L. 2019. Citation recommendation as edge prediction in heterogeneous bibliographic network: a network representation approach// IEEE Access 7, 23232–23239.

192. Yang X. et al. 2014. A survey of collaborative filtering based social recommender systems// Comput. Commun. 41 (1–10).

193. Yang Y.G., Guo X.P., Xu G. et al. Reducing the communication complexity of quantum private database queries by subtle classical post-processing with relaxed quantum ability// Comput. Secur. 2019, 81(3), 15-24.

194. Yao YY (2003) Probabilistic approaches to rough sets. Expert Syst 20:287–297.

195. Yu B, Guo LK, Li QG (2019) A characterization of novel rough fuzzy sets of information systems and their application in decision making. Expert Syst Appl 122:253–261.

196. Yu J. et al. 2018. Adaptive implicit friends identification over heterogeneous network for social recommendation// Proc. of the 27th ACM Int. Conf. on Information and Knowledge Management, pp. 357–366.

197. Yuan F. et al. 2016. Semantic proximity search on graphs with metagraph-based learning// 2016 IEEE 32nd Int. Conf. on Data Engineering (ICDE). IEEE, pp. 277–288.

198. Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353.

199. Zeng AP, Li TR, Liu D, Zhang JB, Chen HM (2015) A fuzzy rough set approach for incremental feature selection on hybrid information systems. Fuzzy Sets Syst 258:39–60.

200. Zhang GQ, Li ZW, Wu WZ, Liu XF, Xie NX (2018) Information structures and uncertainty measures in a fully fuzzy information system. Int J Approx Reason 101:119–149.

201. Zhang J/ et al. 2008. Recommendation over a heterogeneous social network// 2008 the Ninth Int. Conf. on Web-Age Information Management. IEEE, pp. 309–316.

202. Zhang T.A. Dynamic threats assessment based on intuitionistic fuzzy set under missing data condition// Fire Control Command Control 2018, 43(8), 93-97.

203. Zhang X, Mei CL, Chen DG, Li JH (2016) Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy. Pattern Recogn 56:1–15.

204. Zhang XX, Chen DG, Tsang EC (2017) Generalized dominance rough set models for the dominance intuitionistic fuzzy information systems. Inf Sci 378:1–25.

205. Zhang Y. et al. 2018. Learning over Knowledge-Base Embeddings for Recommendation// arXiv preprint arXiv:1803.06540.

206. Zhao H. et al. 2017. Meta-graph based recommendation fusion over

heterogeneous information networks// Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 635–644.

207. Zhao H. et al. 2019. Motif enhanced recommendation over heterogeneous information network// Proc. of the 28th ACM Int. Conf. on Information and Knowledge Management, pp. 2189–2192.

208. Zhao Z. et al. For recommendation// Proc. of the 25th ACM SIGKDD Int. 2020. Hetnerec: heterogeneous network embedding based recommendation. Knowl. Conf. on Knowledge Discovery & Data Mining, pp. 2347–2357.