МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра Дифференциальных уравнений и математической экономики

Алгоритмы разреженной кластеризации для данных большой размерности

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы

направление 38.03.05 – Бизнес-информатика

механико-математического факультета

Хомякова Артёма Константиновича

Научный руководитель
профессор, д.э.н., профессор _____ В. А. Балаш
должность, уч. степень, звание подпись инициалы, фамилия

Заведующий кафедрой
зав. кафедрой, д.ф.-м.н., доцент _____ В. С. Рыхлов
уч. степень, звание подпись инициалы, фамилия

Саратов 2026

Введение. В современных задачах анализа данных всё чаще возникает необходимость работать с высокоразмерными наборами данных, в которых число признаков значительно превышает число объектов или содержит большое количество шумовых и неинформативных переменных. Такие данные встречаются в информационных технологиях, физике, химии, медицине и других прикладных областях. В этих условиях классические методы кластеризации, использующие все признаки одновременно, могут демонстрировать снижение качества, поскольку шумовые координаты и избыточные переменные искажают расстояния между объектами и затрудняют восстановление естественной структуры данных.

Одним из способов решения данной проблемы является разреженная кластеризация. Её идея заключается в том, что при построении кластерного разбиения используются не все признаки, а только те, которые действительно участвуют в формировании кластерной структуры. За счёт этого модель становится более устойчивой к шуму и одновременно более интерпретируемой, поскольку появляется возможность определить, какие признаки вносят основной вклад в разделение объектов.

Актуальность выпускной квалификационной работы обусловлена необходимостью исследования и разработки методов кластеризации, способных эффективно работать с данными большой размерности, содержащими значительную долю неинформативных признаков. Особый интерес представляет сравнение различных подходов к разреженной кластеризации, а также изучение возможностей модификации вероятностных моделей кластеризации с использованием регуляризации.

Целью выпускной квалификационной работы является исследование методов разреженной кластеризации для данных большой размерности, разработка алгоритма кластеризации на основе гауссовой смеси с L1-регуляризацией и оценка его эффективности на синтетических и реальных данных.

Для достижения поставленной цели в работе решаются следующие **задачи**:

1. рассмотреть основные понятия кластеризации, особенности высокоразмерных данных и постановку задачи разреженной кластеризации;

2. изучить существующие алгоритмы разреженной кластеризации и описать их математические основы;
3. выбрать метрики оценки качества кластеризации для использования в экспериментах;
4. разработать алгоритм разреженной кластеризации на основе гауссовой смеси с L1-регуляризацией параметров;
5. реализовать предложенный алгоритм и исследовать его свойства;
6. провести вычислительные эксперименты и сравнить предложенный алгоритм с существующими методами кластеризации;
7. проанализировать полученные результаты и определить области эффективного применения алгоритмов;
8. сформулировать выводы по результатам проведённого исследования.

Материалы исследования. В работе использовались как синтетические, так и реальные данные. Синтетические данные применялись для контролируемого исследования поведения алгоритмов при изменении отдельных характеристик выборки: числа шумовых признаков, степени перекрытия кластеров, наличия выбросов, объёма выборки, числа кластеров и формы кластерной структуры. Такой подход позволил отдельно оценить влияние каждого фактора на качество кластеризации. В качестве реального набора данных использовался набор спектрометрических данных RRUFF, позволяющий проверить работу алгоритмов в прикладных условиях, где структура данных заранее неизвестна и классы могут частично перекрываться.

Для **программной реализации** алгоритмов, генерации синтетических данных, проведения вычислительных экспериментов, визуализации и расчёта метрик качества использовались языки программирования R и Python. В работе применялись пакеты `mclust`, `fossil`, `aricode`, `clue`, `jsonlite`, `progress`, `pandas`, `numpy`, `matplotlib` и `mvtnorm`.

Структура работы. Выпускная квалификационная работа состоит из введения, трёх разделов, заключения, списка использованных источников и приложений. Первый раздел посвящён теоретическим основам разреженной кластеризации. Во втором разделе проводится анализ алгоритмов SPKM и L1-GMM, а также даётся характеристика используемых данных. Третий раз-

дел содержит описание разработанной модификации алгоритма, проведение вычислительных экспериментов и анализ полученных результатов.

Основное содержание работы. **Теоретические основы разреженной кластеризации.** В первом разделе работы рассмотрены основные понятия кластеризации и проблема высокой размерности. Кластеризация понимается как задача разбиения объектов на группы без заранее заданных меток классов. Цель такого разбиения состоит в том, чтобы объекты внутри одного кластера были похожи друг на друга, а объекты из разных кластеров различались достаточно существенно. При этом качество результата во многом зависит от выбранного представления данных и от того, насколько признаки действительно отражают скрытую структуру.

В работе отмечается, что при переходе к данным большой размерности возникают дополнительные трудности. Во-первых, с ростом числа признаков расстояния между объектами становятся менее различимыми, из-за чего методы, основанные на расстояниях, теряют способность надёжно выявлять кластерную структуру. Во-вторых, в реальных данных далеко не все признаки являются информативными. Часть координат может быть случайной, избыточной или не связанной с группировкой объектов. Использование полного пространства признаков в такой ситуации приводит к тому, что шумовые переменные начинают влиять на расстояния и смещать центры кластеров.

Разреженная кластеризация рассматривается как естественное развитие классических методов кластеризации для таких условий. Её ключевая идея состоит в одновременном построении кластеров и отборе признаков. В результате алгоритм не только формирует разбиение объектов, но и определяет, какие признаки важны для этого разбиения. Это повышает устойчивость модели и делает результаты более понятными с точки зрения интерпретации.

В теоретической части также рассмотрены подходы к построению алгоритмов разреженной кластеризации. Особое внимание уделено регуляризации, поскольку именно она является одним из основных механизмов получения разреженности. В работе описаны L1- и L2-регуляризация, а также их модификации, позволяющие учитывать структуру данных. L1-регуляризация имеет особое значение, так как она способна занулять часть параметров модели и тем самым выполнять отбор признаков.

Кроме того, в первом разделе рассмотрены алгоритмы разреженной кластеризации, основанные на вероятностных моделях и целевых функциях. К вероятностным подходам относятся методы, использующие смеси распределений. Они позволяют описывать данные через набор вероятностных компонент и оценивать принадлежность объектов к кластерам в мягкой форме. К методам, основанным на целевых функциях, относятся модификации k-средних, в которых каждому признаку может назначаться свой вес. Также в работе описаны метрики оценки качества кластеризации, применяемые в дальнейших экспериментах: ARI, NMI, JI, FMI и RI.

Анализ алгоритмов и характеристик данных. Во втором разделе работы проведён сравнительный анализ алгоритмов SPKM и L1-GMM. Эти алгоритмы представляют два разных подхода к разреженной кластеризации. SPKM основан на идеях k-средних и использует веса признаков в целевой функции. Если вес признака становится равным нулю, данный признак фактически исключается из построения кластерного разбиения. L1-GMM, напротив, относится к вероятностным методам и строится на основе гауссовой смеси. В нём разреженность достигается за счёт применения L1-регуляризации к параметрам средних компонент смеси.

С точки зрения интерпретируемости оба подхода позволяют выделять признаки, влияющие на кластерную структуру, однако делают это поразному. В SPKM важность признака выражается через его вес. В L1-GMM интерпретация строится через ненулевые значения параметров средних: если параметр по некоторому признаку зануляется, то данный признак перестаёт участвовать в различении соответствующих кластеров.

В работе отмечается, что алгоритм L1-GMM опирается на вероятностное описание данных. Это делает его удобным для анализа ситуаций, где кластеры можно описывать как компоненты смеси распределений. Однако такая модель требует оценки большего числа параметров и обычно оказывается более затратной по времени. Алгоритм SPKM проще и быстрее, но в большей степени ориентирован на разделение кластеров по различию центров и может хуже работать при сложной форме кластеров или сильном перекрытии групп.

Во втором разделе также был обоснован выбор данных для экспериментального исследования. Синтетические данные необходимы для того, чтобы исследовать поведение алгоритмов в контролируемых условиях. Они позволяют заранее задавать число кластеров, степень их разделённости, количество шумовых признаков, наличие выбросов и другие характеристики. Реальные данные, в свою очередь, позволяют проверить, сохраняют ли алгоритмы работоспособность в прикладной задаче, где структура признакового пространства заранее неизвестна. Поэтому в работе используются оба типа данных.

Модификация алгоритма разреженной кластеризации. В третьем разделе работы описана разработка и реализация модифицированного алгоритма L1-GMM. В качестве исходной идеи используется модель гауссовой смеси, в которой каждый кластер соответствует отдельной компоненте распределения. Классический подход к оцениванию параметров такой модели основан на EM-алгоритме, включающем чередование двух шагов: оценивание вероятностей принадлежности объектов к компонентам и обновление параметров модели.

Модификация алгоритма заключается во введении L1-регуляризации при обновлении параметров средних компонент смеси. После вычисления оценок средних применяется мягкое пороговое преобразование, за счёт которого малые значения параметров зануляются. Это позволяет получить разреженную структуру модели: признаки, не вносящие существенного вклада в различие между кластерами, перестают учитываться при построении кластерного решения.

В работе была выполнена самостоятельная реализация алгоритма L1-GMM. Реализация включает подготовку данных, стандартизацию признаков, инициализацию параметров, E-шаг, M-шаг, проверку сходимости и выбор лучшего решения среди нескольких запусков. Также были реализованы вспомогательные процедуры для устойчивых вычислений, предотвращения появления пустых кластеров, оценки качества кластеризации и автоматизации вычислительных экспериментов.

Перед запуском алгоритма данные стандартизируются, поскольку признаки, измеряемые в разных шкалах, могут по-разному влиять на значение штрафа и на итоговое качество кластеризации. Инициализация параметров

выполняется либо на основе алгоритма k-средних, либо случайным образом. Использование нескольких запусков позволяет уменьшить влияние случайной инициализации и выбрать решение с лучшим значением целевой функции.

Отдельное внимание в работе уделено параметру регуляризации. Этот параметр управляет силой L1-штрафа. При малых значениях регуляризация слабо влияет на модель, и большинство признаков сохраняется. При увеличении параметра возрастает число занулённых коэффициентов, то есть модель становится более разреженной. Однако слишком сильная регуляризация может приводить к исключению уже информативных признаков, вследствие чего качество кластеризации ухудшается. Поэтому выбор параметра регуляризации связан с компромиссом между качеством кластеризации и степенью упрощения модели.

Эксперименты на синтетических данных. Для исследования свойств алгоритмов была сформирована серия вычислительных экспериментов на симулированных данных. Сначала был построен базовый набор данных, используемый как отправная точка для дальнейших сравнений. В нём кластеры формировались таким образом, чтобы структура была достаточно выраженной, но не полностью тривиальной. После этого в отдельных сериях экспериментов изменялись конкретные параметры данных, что позволило оценить влияние каждого фактора на качество работы алгоритмов.

На базовом наборе данных оба алгоритма показали высокое качество кластеризации. Значения основных метрик для L1-GMM и SPKM находились на близком уровне, что свидетельствует о хорошем восстановлении исходной структуры данных. При этом L1-GMM по большинству показателей немного превосходил SPKM, однако SPKM работал заметно быстрее.

В серии экспериментов, посвящённой влиянию числа шумовых признаков, исследовалось, насколько алгоритмы устойчивы к добавлению неинформативных координат. Результаты показали, что при умеренном числе шумовых признаков оба алгоритма сохраняют высокое качество. Это подтверждает, что разреженные методы действительно способны снижать влияние признаков, не связанных с кластерной структурой. Вместе с тем при увеличении

числа шумовых координат задача становится сложнее, и качество кластеризации постепенно снижается.

В серии, связанной с изменением степени перекрытия кластеров, было показано, что рост перекрытия существенно ухудшает качество работы обоих алгоритмов. Когда кластеры расположены достаточно далеко друг от друга, методы успешно восстанавливают исходную структуру. Однако при среднем и высоком перекрытии границы между группами становятся менее выраженными, и задача кластеризации усложняется. В таких условиях SPKM в среднем может показывать немного лучшие значения отдельных метрик, но его результаты оказываются менее стабильными. L1-GMM ведёт себя более ровно, хотя при очень сильном перекрытии оба алгоритма работают слабо.

В экспериментах с выбросами анализировалась устойчивость алгоритмов к нетипичным наблюдениям. Было установлено, что небольшое количество выбросов не оказывает критического влияния на качество кластеризации. Однако при увеличении числа выбросов качество закономерно снижается. Это объясняется тем, что ни L1-GMM, ни SPKM в рассматриваемой постановке не содержат специального механизма робастности к выбросам. Следовательно, выбросы могут смещать оценки параметров и ухудшать разбиение.

В серии экспериментов, посвящённой изменению объёма выборки, исследовались качество и время работы алгоритмов при увеличении числа объектов. Было показано, что оба метода сохраняют работоспособность при росте объёма данных. При этом время работы обоих алгоритмов увеличивается, но SPKM в большинстве случаев остаётся быстрее. Для L1-GMM рост времени связан с необходимостью вычисления апостериорных вероятностей и обновления параметров вероятностной модели на каждой итерации.

В серии экспериментов с увеличением числа кластеров оценивалось, насколько алгоритмы сохраняют качество при усложнении кластерной структуры. Результаты показали, что L1-GMM способен сохранять высокое качество при увеличении числа кластеров, если структура данных остаётся достаточно хорошо выраженной. Однако усложнение структуры закономерно повышает требования к инициализации и может увеличивать вычислительные затраты.

Отдельная серия была посвящена данным с неэллиптической структурой. Такие данные сложнее для алгоритмов, основанных на предположении

о компактных или эллиптических кластерах. Результаты показали, что при переходе к неэллиптическим формам качество L1-GMM и SPKM снижается. Это связано с тем, что L1-GMM использует вероятностную модель с гауссовыми компонентами, а SPKM опирается на k-средних-подобную структуру разбиения. Следовательно, оба подхода имеют ограничения при работе со сложными геометрическими формами кластеров.

Исследование на реальном наборе данных. Для проверки алгоритмов в прикладных условиях использовался набор спектрометрических данных RRUFF. Данный набор содержит спектры минералов и представляет интерес как пример высокоразмерных данных, в которых признаки имеют сложную структуру. В работе рассматривался как поднабор из десяти наиболее представленных классов, так и полный набор данных.

На поднаборе RRUFF_top10 алгоритм L1-GMM показал конкурентоспособное качество кластеризации. Хорошо отделённые классы восстанавливались достаточно точно, тогда как близко расположенные и частично перекрывающиеся классы разделялись хуже. Это согласуется с результатами синтетических экспериментов, где было показано, что перекрытие кластеров является одним из факторов, наиболее сильно ухудшающих качество.

Сравнение L1-GMM, SPKM, kmeans и mclust показало, что на реальных данных различия между алгоритмами становятся более заметными. На поднаборе из десяти классов наиболее быстрыми оказывались kmeans и mclust, SPKM также демонстрировал высокую скорость, а L1-GMM был более затратным по времени. При переходе к полному набору RRUFF задача становилась сложнее за счёт увеличения числа классов и числа спектров. В этих условиях L1-GMM сохранял сильные результаты и оставался одним из лидеров по качеству, хотя по скорости уступал более простым методам.

Отдельно исследовалось влияние параметра регуляризации на качество кластеризации и степень разрежения модели. Было показано, что при малых значениях параметра регуляризации качество практически не изменяется, а регуляризация в основном приводит к занулению части параметров. При средних значениях начинает проявляться компромисс между качеством и разреженностью: модель становится компактнее, но метрики качества постепенно снижаются. При больших значениях параметра качество ухудшает-

ся существенно, поскольку вместе с неинформативными признаками начинают исключаться признаки, важные для восстановления структуры данных.

Для более наглядного анализа результатов была использована PCA-визуализация набора RRUFF_top10. Она показала, что алгоритм L1-GMM хорошо выделяет удалённые классы, расположенные отдельно от остальных. В то же время для групп, находящихся близко друг к другу, качество разделения ниже. Это подтверждает, что успешность кластеризации зависит не только от выбранного алгоритма, но и от геометрической структуры данных.

Таким образом, эксперименты на реальном наборе данных подтвердили применимость разработанного алгоритма к высокоразмерным спектральным данным. Вместе с тем они показали и ограничения метода: при большом числе классов, частичном перекрытии групп и сложной структуре признакового пространства задача кластеризации становится значительно труднее.

Заключение. В ходе выполнения выпускной квалификационной работы были рассмотрены методы разреженной кластеризации для высокоразмерных данных и проведено их теоретическое и экспериментальное исследование. В теоретической части были изучены основные понятия кластеризации, особенности данных большой размерности, подходы к построению алгоритмов разреженной кластеризации и метрики оценки качества кластерных решений.

В практической части работы был разработан и реализован алгоритм L1-GMM, основанный на модели гауссовой смеси с L1-регуляризацией параметров средних. Разреженность в данном алгоритме достигается за счёт зануления малых значений параметров, что позволяет уменьшать влияние неинформативных признаков и получать более компактное описание модели. Также были реализованы вспомогательные процедуры для подготовки данных, устойчивых вычислений, предотвращения пустых кластеров, оценки качества и автоматизации экспериментальных запусков.

Проведённые эксперименты на синтетических данных показали, что алгоритм L1-GMM хорошо работает на базовом наборе данных, устойчив к умеренному числу шумовых признаков и сохраняет высокое качество при увеличении числа кластеров. При этом качество кластеризации снижается при сильном перекрытии кластеров, большом числе выбросов и переходе к

данным со сложной неэллиптической структурой. Сравнение с алгоритмом SPKM показало, что L1-GMM в ряде постановок демонстрирует сопоставимое или более высокое качество, однако уступает SPKM по времени работы.

Исследование влияния параметра регуляризации показало, что L1-регуляризация позволяет уменьшать число ненулевых параметров модели без заметной потери качества при умеренных значениях параметра. Это подтверждает возможность использования L1-GMM не только для кластеризации, но и для отбора признаков. Однако при слишком сильной регуляризации качество резко ухудшается, поскольку модель начинает терять информативные признаки.

Эксперименты на реальном наборе данных RRUFF подтвердили работоспособность предложенного алгоритма на высокоразмерных спектральных данных. Было показано, что L1-GMM способен выделять хорошо отделённые классы и демонстрировать конкурентоспособное качество по сравнению с альтернативными методами. Вместе с тем при наличии близко расположенных и частично перекрывающихся групп качество разделения снижается, что отражает объективную сложность рассматриваемой задачи.

Таким образом, поставленные в работе задачи были выполнены. Были рассмотрены существующие методы разреженной кластеризации, разработан и реализован алгоритм L1-GMM, проведено его сравнение с другими подходами на синтетических и реальных данных, а также исследовано влияние параметра регуляризации на качество и степень разрежения модели. Полученные результаты показывают, что предложенный подход может использоваться для кластеризации высокоразмерных данных в задачах, где требуется не только построение кластерного разбиения, но и сокращение числа используемых признаков.

Практическая значимость полученных результатов состоит в том, что они могут служить ориентиром при выборе метода кластеризации для высокоразмерных данных с большим числом неинформативных признаков. Перспективы дальнейшего развития работы связаны с исследованием более гибких ковариационных структур, повышением вычислительной эффективности алгоритма L1-GMM и проверкой его поведения на других типах реальных высокоразмерных данных.