

Корчмит Тамара Дмитриевна
*студент, Сибирский государственный университет науки и технологий имени
академика М. Ф. Решетнёва,
Россия, г. Красноярск*

ОГРАНИЧЕНИЯ И ПРЕИМУЩЕСТВА ЛОГИЧЕСКОГО АНАЛИЗА ДАННЫХ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ

Цель. Анализ ограничений и преимуществ логического анализа данных (LAD) при решении задачи кластеризации по сравнению с классическими алгоритмами (k-means, DBSCAN). Метод. Проведён обзор и систематизация литературных источников, посвящённых принципам формирования логических правил, критериям валидации, влиянию бинаризации и уровня шума на стабильность правил, а также вычислительной сложности LAD-методов. Результат. Систематизированы ключевые преимущества LAD (интерпретируемость, прозрачность правил, возможность верификации экспертом) и основные ограничения (чувствительность к бинаризации, высокая вычислительная сложность, деградация при зашумлённых данных). Выводы. LAD рекомендуется применять на хорошо структурированных данных умеренного размера с низким уровнем шума; для крупномасштабных задач целесообразно использование гибридных схем.

ABSTRACT

Background. Analysis of limitations and advantages of Logical Analysis of Data (LAD) in the clustering problem compared to classical algorithms (k-means, DBSCAN). Methods. A review and systematization of literature sources on the principles of logical rule formation, validation criteria, the influence of binarization and noise level on rule stability, as well as the computational complexity of LAD methods, was conducted. Result. The key advantages of LAD (interpretability, rule transparency, verifiability by an expert) and main limitations (sensitivity to binarization, high computational complexity, degradation under noisy data) are systematized. Conclusion. LAD is recommended for well-structured data of moderate size with low noise levels; for large-scale problems, hybrid schemes are advisable.

Ключевые слова: логический анализ данных, кластеризация, ограничения, преимущества, интерпретируемость, бинаризация, вычислительная сложность.

Keywords: logical data analysis, clustering, limitations, advantages, interpretability, binarization, computational complexity.

Введение

Современные методы кластеризации, такие как k-means и DBSCAN, стали стандартными инструментами анализа данных благодаря своей вычислительной эффективности и способности обрабатывать большие объемы информации. Однако в прикладных областях, например в медицине или социологии, результаты этих алгоритмов часто оказываются сложными для интерпретации: центроиды или границы плотности не всегда транслируются в понятные специалисту правила, что затрудняет принятие решений. Логический анализ данных (LAD) предлагает альтернативный подход, представляя кластеры в виде прозрачных логических правил, которые могут быть легко верифицированы и осмыслены экспертом. Несмотря на это, применение LAD для задачи кластеризации изучено недостаточно систематически, а отсутствие четких критериев, когда такой подход дает преимущество, а когда сталкивается с непреодолимыми вычислительными ограничениями, сдерживает его внедрение в практику.

Целью данной работы является анализ ограничений и преимуществ метода LAD для кластеризации, а также установление условий, при которых этот метод может обеспечить более интерпретируемые решения по сравнению с классическими алгоритмами. Основная идея исследования заключается в сравнительном анализе LAD и традиционных методов, таких как k-means и DBSCAN, с акцентом на интерпретируемость получаемых решений и выявление границ применимости каждого из подходов.

Для достижения поставленной цели в работе решаются следующие задачи:

1) провести обзор и систематизацию существующих подходов к кластеризации на основе LAD; 2) на основе анализа литературных источников выполнить

теоретическое сравнение LAD-кластеризации с классическими методами; 3) систематизировать известные ограничения LAD, связанные с вычислительной сложностью, способами бинаризации и влиянием шума; 4) выделить ключевые преимущества LAD в контексте интерпретируемости.

Основная часть

1. Принципы формирования логических правил в задачах обучения без учителя

Логические правила в логическом анализе данных (LAD) представляют собой конъюнкции бинарных признаков, которые индуцируются из предварительно бинаризованных данных и описывают условия принадлежности объектов к тому или иному кластеру. В задачах обучения без учителя построение таких правил происходит без использования меток, что принципиально отличает их от классических задач классификации. Теоретической основой данного подхода во многом служит формальный анализ понятий (FCA), который, как отмечается в литературе, «был введён в начале 1980-х годов Рудольфом Вилле как математическая теория для формализации понятий и концептуального мышления» [13, с.2]. FCA позволяет идентифицировать интересные кластеры (формальные понятия) в множестве объектов и их атрибутов и организовывать их в структуру, называемую решёткой понятий, что созвучно целям LAD-кластеризации.

Ключевым этапом процедуры LAD является генерация и отбор правил, обеспечивающих покрытие объектов кластера при одновременной минимизации пересечений между правилами. Для контроля размера и компактности кластеров применяются пороги покрытия (доля объектов, описываемых правилом) и ограничения на длину правила (число конъюнктов). Такая настройка позволяет балансировать между полнотой описания кластера и интерпретируемостью полученных правил.

2. Преимущества LAD: интерпретируемость и прозрачность

Метрические алгоритмы кластеризации, такие как k-средних, и плотностные методы, например DBSCAN, предоставляют результаты в виде

центроидов или границ плотности, которые не всегда отражают семантическую структуру данных и сложны для содержательной верификации экспертом. Логический анализ данных (LAD) принципиально отличается от этих подходов, предлагая явные логические условия в форме булевых правил, выделяющих каждый кластер. Как отмечается в литературе, «в настоящее время индексы качества кластеризации не учитывают представления пользователя о желаемом результате кластеризации и семантику кластерной структуры в процессе оценки качества» [7, с.10]; LAD позволяет преодолеть этот недостаток, поскольку сгенерированные правила могут быть непосредственно проверены на соответствие предметным знаниям.

К основным преимуществам LAD в задаче кластеризации можно отнести:

Прозрачность решений — каждое правило может быть представлено в виде простой конъюнкции условий, описывающей общие свойства элементов кластера;

Локальность описаний — в отличие от центроидов k -средних, которые дают усреднённый образ кластера, LAD предоставляет точную причину принадлежности объекта;

Возможность верификации экспертом — специалист способен оценить каждое условие и при необходимости скорректировать модель.

3. Ограничения LAD: бинаризация, шум и вычислительная сложность

Несмотря на преимущества в интерпретируемости, LAD имеет существенные ограничения по сравнению с классическими методами.

3.1. Бинаризация признаков. Необходимость дискретизации непрерывных признаков может приводить к потере информации о точных градациях значений. Как указывается в литературе, «в использовании деревьев решений существуют следующие ограничения: подобные модели неприемлемы при обработке большого количества непрерывных величин; кроме того, если список условий длинный и сложный, дерево решений получается необозримым и непонятным» [10, с.5]; эта проблема прямо переносится на LAD в условиях неоптимальной бинаризации.

3.2. Чувствительность к шуму. Уровень случайного шума в признаках проявляется как закономерное снижение покрытия кластеров и одновременный рост фрагментации индуцированных правил. Согласно результатам, представленным в литературе [2,4], при превышении определённых пороговых значений шума наблюдается резкая деградация стабильности правил.

3.3. Вычислительная сложность. Время генерации логических правил растёт экспоненциально с увеличением размерности пространства признаков и объёма выборки. «Для обработки больших массивов данных возникает потребность разработки алгоритмов для реализации на высокопроизводительных вычислительных системах» [7] [11, с.2063].

4. Критерии валидации и оценка качества

Качество кластеризации оценивается двумя группами метрик: внешними для сравнения с эталонной разметкой (например, ARI и NMI) и внутренними для анализа компактности и разделимости кластеров (например, силуэтный коэффициент). Как отмечается в исследовании [13], высокое значение индекса силуэта соответствует наиболее информативным подгруппам. Интерпретируемость результатов LAD формализуется через пороги, включающие максимальную длину правила, показатель простоты и семантическую читаемость, оцениваемую экспертами.

Заключение

Выполненный в работе теоретический анализ и обзор литературы позволяют сделать следующие выводы. Логический анализ данных (LAD) обеспечивает существенное преимущество в интерпретируемости кластерных решений по сравнению с традиционными методами, формируя кластеры в виде прозрачных логических правил. Это особенно ценно в прикладных областях, требующих объяснимости, например в медицине и социологии.

Вместе с тем, как систематизировано в работах [2,4,7], эффективность LAD ограничена следующими факторами:

- высокой вычислительной сложностью (экспоненциальный рост времени поиска правил);

- чувствительностью к способу бинаризации признаков;
- деградацией стабильности правил при увеличении уровня шума.

Таким образом, LAD демонстрирует наилучшие результаты на хорошо структурированных данных умеренного размера с низким уровнем шума. Для практического применения рекомендуется использовать гибридные схемы, сочетающие предварительную фильтрацию шума и метрическую инициализацию с последующим логическим уточнением границ кластеров. Дальнейшие исследования должны быть направлены на создание масштабируемых версий алгоритмов LAD и разработку адаптивных методов бинаризации.

Список литературы:

1. Бурнашев Р.А., Сергеев Я.В., Назипова А.Ф. Методы гранулирования нечётких временных рядов для анализа данных // Онтология проектирования. 2025. №3. С. 404-417.
2. Журавлева В.В., Маничева А.С. Упрощенный показатель силуэта для определения качества кластерных структур // Известия АлтГУ. Математика и механика. 2022. №4. С. 110-114.
3. Коннов И.В., Кашина О.А., Гильманова Э.И. Решение задачи кластеризации методами оптимизации на графах // Ученые записки Казанского университета. 2019. №3. С. 423-437.
4. Кочергин М.И., Ганджа Т.В. Информационные технологии в технико-экономических системах. Томск: ТУСУР, 2018. 23 с.
5. Лютикова Л.А. Применение логического анализа данных в программном комплексе для диагностики гастритов // Информационное общество. 2024. №4. С. 78-82.
6. Мезенцев Ю.А., Разумникова О.М., Эстрайх И.В. и др. Задачи и алгоритмы оптимальной кластеризации многомерных объектов // Компьютерные исследования и моделирование. 2024. №3. С. 673-693.

7. Сивоголовко Е.В. Оценка качества кластеризации в задачах интеллектуального анализа данных: автореферат. СПбГУ, 2014.
8. Федоров А.С., Шиков А.Н. Метод преобразования семантической сети // Вестник АГТУ. 2020. №4. С. 7-17.
9. Цициашвили Г.Ш., Осипова М.А., Лосев А.С. Алгоритмы кластеризации графов // Вестник ВГУ. 2016. №1. С. 145-149.
10. Черняховская Л.Р., Рыжов Г.И., Павлова А.Н. и др. Методические рекомендации по дисциплине «Методы искусственного интеллекта в управлении качеством». Уфа: УГАТУ, 2021.
11. Четверушкин Б.Н., Яшина М.В., Таташев А.Г. Алгоритм кластеризации графа // XIV всероссийское совещание по проблемам управления. Москва, 2024. С. 2062-2066.
12. Шкитронов М.Е. Классификация алгоритмов машинного обучения // Современная наука. 2025. №10. С. 204-208.
13. Масич И.С., Резова Н.Л., Шкаберина Г.Ш. и др. Subgroup discovery in machine learning problems with formal concepts analysis and test theory algorithms // Algorithms. 2023. №5. С. 2-25.