

Редукционная кластеризация высокоразмерных данных

А. Ю. Дорогов

Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
vaksa2006@yandex.ru

Аннотация. Предложен метод не параметрической кластеризации Больших данных, основанный на гистограммном анализе образов в признаковом пространстве. Метод позволяет локализовать кластерные зоны и моды кластеров в подпространствах признакового пространства, не используя метрики расстояния. Предложенный метод обходит «проклятие размерности» и пригоден для анализа как числовых, так и категориальных высокоразмерных данных.

Ключевые слова: кластерная структура; редукция данных; координатный профиль; гистограмма видимости; кластерный анализ

I. ВВЕДЕНИЕ

Следуя гипотезе компактности, под кластером будем понимать область признакового пространства, в которой плотность точек образов выше, чем в её окрестности. В зависимости от варианта формализации модели кластера приходят к различным методам кластерного анализа. В работе 1996 г. [1] представлен известный метод DBSCAN. Для этого метода модель кластера определяется через связанные цепочки локальных областей с малым радиусом ε , в которых количество точек образов превышает заданный порог. Базовая идея была развита в методе OPTICS [2] в этом методе пороговое значение и параметр ε адаптируются к данным. Для оценки локальной плотности точек могут быть использованы методы Mean-shift [3] на основе ядерных функций, по типу функций Гаусса. Методы Mean-shift осуществляют локальное сглаживание значений плотности и позволяют выделять моды кластеров. Вышеперечисленные методы и им подобные используют некоторую метрику расстояния. Наиболее распространённой метрикой является евклидово расстояние. Следует отметить, что в многомерном пространстве наблюдается эффект концентрации меры [4, 5], который проявляется в том, что подавляющая часть объёма многомерного шара сосредоточена в малой окрестности его границы. В результате, евклидова метрика может быть практически бесполезна для поиска подходящего значения ε . Этот эффект присутствует и в любом другом алгоритме, основанном на евклидовом расстоянии. «Проклятие размерности» проявляется также в экспоненциальном росте вычислительной сложности подобных алгоритмов.

Кластерный анализ предполагает существование подсистема измерений, которая отображает объекты в

образы, суть – точки признакового пространства. Будем полагать, что такое отображение является однозначным [6]. Однозначность по существу означает отсутствие противоречий в пространстве образов, т.е. при однозначной схеме измерения подмножества объектов отвечающих разным образам не пересекаются.

II. АЛГОРИТМ РЕДУКЦИОННОЙ КЛАСТЕРИЗАЦИИ

А. Координатные профили

Под координатным профилем будем понимать проекцию многомерной плотности точек на координатную ось признакового пространства. При выполнении гипотезы компактности, можно ожидать, что вид координатного профиля будет в той или иной мере отражать кластерную структуру признакового пространства. Иллюстрация данного положения для двумерного пространства показана на рис. 1. На практике, координатные профили могут быть получены построением гистограмм значений каждого информативного признака.

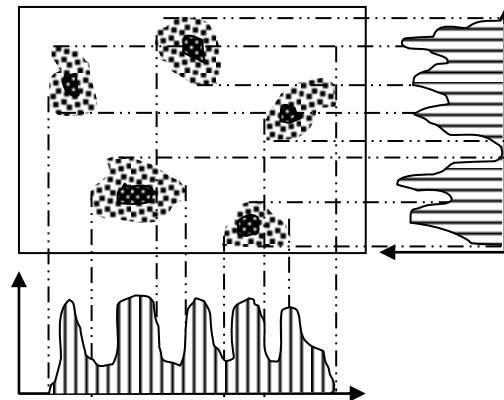


Рис. 1. Координатные профили

В. Кластерные зоны

На координатном профиле нанесём линию порогового уровня и выделим области, в которых плотность точек превышает пороговый уровень. На рис. 2 эти области показаны штриховкой на горизонтальной оси. Число построенных зон является оценкой числа кластеров. Эта величина, очевидно, зависит от позиции линии порогового уровня. Если пороговый уровень близок к нулю, то число зон уменьшается до единицы, и мы теряем информацию о кластерной структуре, при увеличении порогового уровня

число кластерных зон возрастает до максимального, но при этом снижается покрытие кластерными зонами точек признакового пространства, и мы рискуем получить кластерную структуру с низким уровнем представительности. Число мод на координатной гистограмме, очевидно, не может превысить числа кластеров.

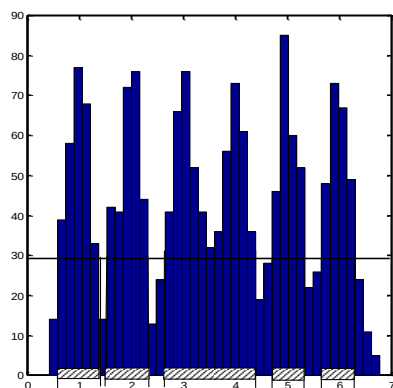


Рис. 2. Кластерные зоны координатного профиля

Точки с координатами, принадлежащими малым окрестностям какой-либо моды, потенциально являются кандидатами точек, принадлежащих кластерам. При построении профильных гистограмм определим также принадлежность точек бинам гистограммы каждого координатного профиля и назовём эти данные профильными векторами принадлежности точек.

С. Локализация кластерных зон

Для решения задачи выбора кластерных зон профиля, будем перемещать линию уровня с шагом 1 от нулевого до максимального значения гистограммы, и при каждом положении порогового уровня вычислять число кластерных зон и суммарный объем их покрытий по точкам признакового пространства. В результате получим зависимости числа кластерных зон и их покрытий от значения порогового уровня. На рис. 3 показаны примеры построения подобных зависимостей для двумерной кластерной структуры, показанной на рис. 4.

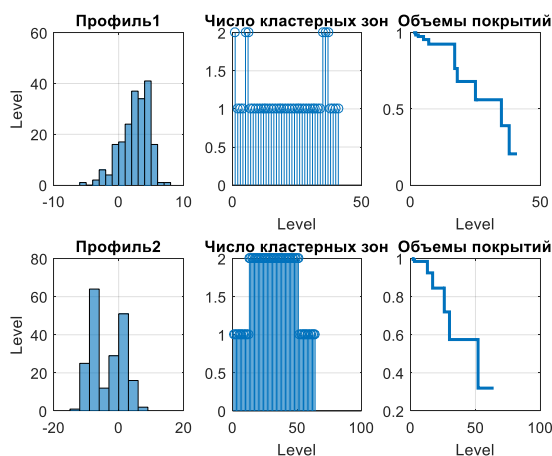


Рис. 3. Алгоритм покоординатной локализации кластерных зон

На данном рисунке координатная ось *Level* означает пороговый уровень координатного профиля. Объёмы покрытий нормированы по отношению к полному числу точек класса. Функция объёма покрытия является монотонно убывающей по значению *Level*. Функция числа кластерных зон состоит из участков с постоянным значением. Очевидно, что это справедливо для любых профилей, поэтому значения *Level* в начале каждого участка постоянства функции числа кластерных зон, являются наилучшим выбором порогового уровня, поскольку кластерные зоны в этом случае имеют наибольшее суммарное покрытие. Таким образом, для каждого профиля из построенных зависимостей получим несколько подходящих значений порогового уровня *Level*.

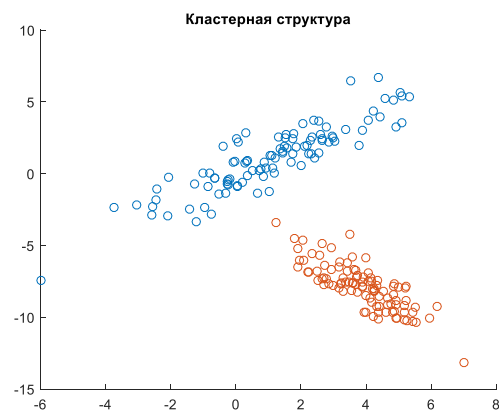


Рис. 4. Пример кластерной структуры

Протяжённость участков постоянства функции числа кластерных зон называется персистентностью [7] или устойчивостью топологической характеристики координатного профиля. При выборе порогового уровня *Level* из числа подходящих можно руководствоваться тремя критериям: персистентностью участков, числом кластерных зон и объёмом покрытия в начале участка постоянства.

Д. Редукция данных

Цель редукции заключается в уменьшении объёма данных при сохранении кластерной структуры признакового пространства. Предположим, что из числа подходящих значений пороговых уровней, выбрано наилучшим образом единственное значение *Level* для каждого профиля. По выбранным значениям *Level* построим кластерные зоны профилей и присвоим зонам индивидуальные локальные метки. Используя построенные ранее профильные векторы принадлежности, для каждого координатного профиля пометим все точки признакового пространства класса, метками кластерных зон. В отдельных профилях некоторые точки могут остаться не помеченными, если они не принадлежат ни одной кластерной зоне. Результаты соберём в таблицу (табл. 1).

В этой таблице первый столбец содержит порядковые номера образов класса, а остальные столбцы заполнены метками кластерных зон по координатным профилям.

Некоторые клетки таблицы могут оказаться пустыми. Если строка образа полностью заполнена метками зон, то данный образ проявляется в кластерных зонах всех координатных профилей и это означает, что он принадлежит некоторому кластеру в полном признаковом пространстве. Группируя все строки таблицы с выбранным набором меток, получим состав данного кластера. Строки, заполненные другим набором меток, формируют наполнение следующего кластера и т. д. Таким образом, в результате селекции уникальных строк таблицы будет полностью построена кластерная структура полного признакового пространства. Аналогично может быть выделена кластерная структура для подпространств.

ТАБЛИЦА 1 РАСПРЕДЕЛЕНИЕ ОБРАЗОВ ПО КЛАСТЕРНЫМ ЗОНАМ КООРДИНАТНЫХ ПРОФИЛЕЙ

Номер образа	Метки кластерных зон				
	Профиль 1	Профиль 2	...	Профиль m-1	Профиль m
1	*	*		*	*
2	*	*		*	*
...	...	...		...	...
N	*	*		*	*

Наглядной характеристикой наполненности таблицы является гистограмма видимости или координатной встречаемости точек [8]. Горизонтальная ось гистограммы определяет количество заполненных клеток в строке таблицы, а вертикальная ось – количество строк таблицы с одинаковой степенью наполненности. Гистограмма показывает количество точек класса, которые видны через координатные зоны в заданном числе профилей. Пример подобной гистограммы для 15-мерного признакового пространства показан на рис. 5.

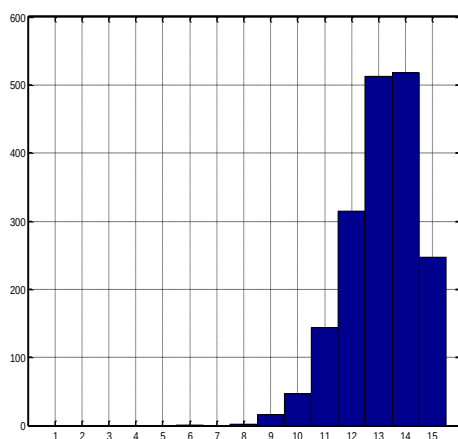


Рис. 5. Гистограмма видимости точек по кластерным зонам профилей

Последний столбец гистограммы соответствует количеству строк таблицы с полной наполненностью по строке, и определяет объем покрытия для кластерной структуры полного признакового пространства.

III. КЛАСТЕРНЫЙ АНАЛИЗ

Цель кластерного анализа заключается в том, чтобы определить все возможные кластерные структуры признакового пространства и установить соотношения между ними. Для решения данной задачи просматриваются все комбинации подходящих пороговых значений профильных зон, для каждой комбинации выполняется редукционная кластеризация и оценка покрытия найденной кластерной структуры. При больших размерностях признакового пространства, прямой перебор может потребовать значительных вычислительных затрат. Для сокращения объема вычислений, можно ввести дополнительные ограничения, например, установить минимально допустимый размер покрытия для кластерной структуры. В этом случае в процедуре перебора участвуют только те уровневые значения профильных зон, покрытия которых не меньше чем заданное покрытие кластерной структуры.

Для решения задачи классификации в большинстве случаев достаточно построить в каждом классе только одну кластерную структуру. Эта структура должна быть представительна по покрытию и иметь по возможности максимальное число кластеров. Идеальным решением является первоначальное построение всех допустимых кластерных структур признакового пространства, а затем выбор из них наилучшей. Но как уже было отмечено этот путь ведёт к существенным вычислительным затратам. Поэтому имеет смысл предложить некоторые эвристические решения, позволяющие сразу найти одну оптимальную структуру или структуру близкую к оптимальной. Эвристики сводятся к оптимальному выбору единственного подходящего распределения кластерных зон на координатных профилях. Параметрами такого выбора являются: объем покрытия кластерными зонами, число кластерных зон и протяжённость участка постоянства функции числа кластерных зон.

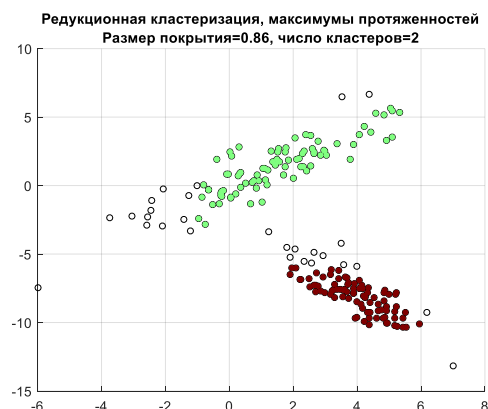


Рис. 6. Выбор по максимуму протяжённости участков постоянства в профилях. Кластерная структура 1

Удовлетворительные результаты можно получить, если размер зонового покрытия установить как ограничение (например, > 0.75) и выбрать при этом пороговый уровень профиля с максимальным числом кластерных зон или с

максимальной протяжённостью участка постоянства числа кластерных зон. На рис. 6 и рис. 7 показаны кластерные структуры, построенные с использованием двух эвристик для двумерного признакового пространства.

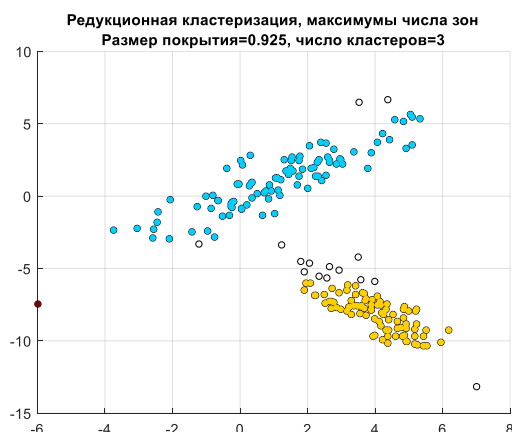


Рис. 7. Выбор по максимуму числа кластерных зон в профилях. Кластерная структура 2

Не кластеризованные точки на рисунках представлены не закрашенными маркерами.

А. Эталоны кластеров

Результаты кластерного анализа часто используются для построения метрических классификаторов данных. В этом применении принадлежность точки классу определяется по минимальному расстоянию между распознаваемой точкой и кластерами классов. Для Больших данных такой подход неэффективен по вычислительным затратам, поскольку требует оценки расстояния между распознаваемой точкой и всеми точками обучающей выборки. Более эффективным решением является сравнения распознаваемой точки с эталонами (модами) класса. В простейшем варианте эталоном класса может служить средний вектор кластера. Число эталонов в классе в этом случае равно числу кластеров. При использовании эталонов более подходящей мерой межклассового разделения является дискриминант для пары векторов.

IV. ЗАКЛЮЧЕНИЕ

Выполненное исследование показывает, что с ростом размерности признакового пространства объем покрытия кластерной структуры уменьшается, т. е. происходит редукция числа точек к объёму кластерного покрытия при сохранении структуры данных. Поэтому данный метод кластеризации уместно назвать редукционным. Для Больших данных это свойство алгоритма может играть положительную роль, поскольку редукция делает Большие

данные обозримыми без потери информации о кластерной структуре [9, 10].

Оценкой качества кластеризации может служить суммарный объем покрытия кластерной структурой точек класса. Чем он выше, тем лучше качество кластеризации. Максимальное значения данного критерия, очевидно, не превышает число образов в признаковом пространстве.

Следует заметить, что алгоритм редукционной кластеризации не использует какую-либо меру близости в признаковом пространстве, поэтому он применим к выборкам данных представленных в любой шкале от интервальной до номинальной.

СПИСОК ЛИТЕРАТУРЫ

- [1] Ester Martin, Kriegel Hans-Peter, Sander Jörg, Xu Xiaowei (1996). Simoudis Evangelos, Han Jiawei, Fayyad Usama M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.
- [2] Ankerst Mihael, Breunig Markus M., Kriegel Hans-Peter, Sander Jörg (1999). "OPTICS: Ordering Points To Identify the Clustering Structure" // ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.
- [3] Achtert E., Böhm C., Kröger P. (2006). "DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking". Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science. 3918. pp. 119–128.
- [4] Зорич В.А. Многомерная геометрия, функции очень многих переменных и вероятность // Теория вероятн. и ее примен., 2014, том 59, выпуск 3, С. 436–451/ DOI:https://doi.org/10.4213/tvp4578
- [5] Мильман В.Д. Явления, возникающие в высоких размерностях. Успехи матем. наук, 2004, т.59, №1, с.157–168.
- [6] Гисин В.Б., Цаленко М.Ш. Алгебраическая теория систем и её приложения. Системные исследования. Методологические проблемы: Ежегодник // 1984. М.: Наука, 1984. С. 130–151.
- [7] Макаренко Н.Г., Князева И.С. Топология и данные. Труды XX международной научно-технической конференции «Нейроинформатика-2018». Лекции по нейроинформатике. М.: НИЯУ МИФИ, 2018. С. 72-108.
- [8] Дорогов А.Ю., Курбанов Р.Г. Шестопалов М.Ю. Редукция выборки в задачах многомерной кластеризации // Сборник докладов Всероссийской научной конференции «Управление и информационные технологии (УИТ-2005)». Санкт-Петербург 30 июня-2 июля 2005. Т.1. С. 284-288.
- [9] Дорогов А.Ю. Кластерный анализ обучающих выборок // «Информационные системы и технологии в моделировании и управлении»: сборник трудов V Международной научно-практической конференции (20-22 мая 2020 г.) / отв. редактор К.А. Маковейчук г. Симферополь: ООО «Издательство типография "Ариал"», 2020. С. 48-56.
- [10] Дорогов А.Ю. Кластерный анализ высокоразмерных данных в задаче классификации образов // Материалы XXVIII Всероссийского семинара Нейроинформатика, её приложения и анализ данных, 25 сентября – 27 сентября 2020 года / Под ред. М.Г. Садовского, отв. за вып. М.Ю. Сенашова; Красноярск: Институт вычислительного моделирования СО РАН, 2020. с. 42-49.