УДК 519.7 Научная статья

Алгоритм кластеризации на основе разбиения пространства признаков

М. А. Казаков

Институт прикладной математики и автоматизации КБНЦ РАН
360000, КБР, г. Нальчик, ул. Шоратнова 89 А
E-mail: kasakow.muchamed@gmail.com

В данной статье предлагается новый способ робастной кластеризации на основе рекурсивного разбиения пространства признаков и анализа плотностей. Представлен алгоритм робастной кластеризации линейно неразделимых точек, его программная реализация, а также результаты тестирования на классических наборах данных.

Ключевые слова: кластеризация, робастная кластеризация, машинное обучение.

DOI: 10.26117/2079-6641-2022-39-2-136-149

Поступила в редакцию: 05.07.2022 В окончательном варианте: 22.08.2022

Для цитирования. Казаков М. А. Алгоритм кластеризации на основе разбиения пространства признаков // *Вестник КРАУНЦ. Физ.-мат. науки.* 2022. Т. 39. № 2. С. 136-149. DOI: 10.26117/2079-6641-2022-39-2-136-149

Введение

Системы машинного обучения можно классифицировать в зависимости от количества и типа контроля, который они получают во время обучения. Существует четыре основных категории: обучение с учителем, обучение без учителя, обучение с частичным привлечением учителя и обучение с подкреплением. В обучении с учителем обучающий набор, который передается алгоритму, включает в себя правильные решения, называемые метками. Типичными задачами обучения с учителем являются классификация и регрессия. При обучении без учителя обучающие данные не помечены. Система пытается учиться самостоятельно. Поскольку маркировка данных обычно требует много времени и средств, часто в наборах данных

Финансирование. Работа выполнена без финансовой поддержки

много экземпляров без меток и несколько экземпляров с метками. Некоторые алгоритмы могут работать с данными, которые частично помечены (обучением с частичным привлечением учителя). Обучение с подкреплением представляет собой совсем иной подход. Система обучения может наблюдать за окружающей средой, выбирать и выполнять действия, а взамен получать вознаграждения или штрафы в виде отрицательных вознаграждений [1].

Некоторые из наиболее важных задач обучения без учителя: кластеризация, обнаружение аномалий и обнаружение новизны, уменьшение размерности, изучение ассоциативных правил. Кластеризация — это метод исследовательского анализа данных, который позволяет организовать множество данных в значимые подгруппы (кластеры), не зная заранее об их принадлежностях к группам. Каждый кластер, возникающий в ходе анализа, определяет группу объектов, имеющих определенную степень сходства и более непохожих на объекты других кластеров. Кластеризация позволяет структурировать информацию и получать значимые отношения в данных [2]. Кластеризация — отличный инструмент для анализа данных, сегментации клиентов, рекомендательных систем, поисковых систем, сегментации изображений, обучения с частичным привлечением учителя, уменьшения размерности и многого другого.

Многие алгоритмы кластеризации реализованы в высокоуровневых библиотеках, работа с которыми описана в замечательных учебниках [1]-[3]. Среди множества алгоритмов кластеризации, можно отметить такие популярные методы, как метод k-средних [4], [5], агломеративная кластеризация [6], [7], метод DBSCAN [8], [9], робастные алгоритмы оценки averaging aggregation functions [10]-[12]. Интересные результаты получены в работах [13]-[17].

При кластеризации данных возникает множество проблем. В силу отсутствия размеченных данных, алгоритму необходимо выявлять закономерности, присутствующие только в свойствах объектов. Первая проблема связана с адекватным выявлением количества кластеров. Более простым алгоритмам, таким как k-means, требуется явное задание количества кластеров. В некоторых случаях это число может быть заранее известным, либо полученным в результате предварительного анализа данных. Это может существенно упростить задачу кластеризации. Более сложные алгоритмы способны самостоятельно определить количество кластеров. Вторая проблема заключается в том, что количество кластеров, на которые следует разбивать множество точек может быть не однозначным. Более того, в большинстве случаев распределения точек реальных database может и не быть единственно корректного разбиения на кластеры, хотя почти всегда есть более предпочтительные разбиения. Третья проблема связана с тем, что точки могут не быть линейно разделимыми. Для разбиения на кластеры линейно разделимые распределения точек достаточно использовать гиперплоскости. В случае линейно неразделимых распределений возникает необходимость использовать более специфические алгоритмы. Четвертой проблемой можно отметить проблему выбросов. С одной стороны, здесь поднимается вопрос интерпретации: какие точки считать выбросами, а какие нет? С другой стороны, при кластеризации мы можем отвлечься от этих

вопросов, и оставаться наедине с самими точками. Выбросы могут нежелательно повлиять на формирование кластеров. Для алгоритмов, не учитывающих влияние выбросов, наличие выбросов может существенно исказить результат. И даже алгоритмы, учитывающие влияние выбросов, не всегда могут нивелировать их влияние. Особое место занимает пятая проблема: во многих наборах данных может оказаться уместным поиск подкластеров внутри кластеров. Это приводит к необходимости иерархической кластеризации, при котором выстраивается дерево кластеров. Приведенное перечисление проблем не является исчерпывающим, однако их решение позволяет создавать эффективные алгоритмы кластеризации. Стоит так же отметить, что не всегда выбор сложных алгоритмов при решении задачи является оправданным. Сложные алгоритмы, как правило, работают гораздо медленнее, чем простые. В некоторых задачах, когда нам заранее известно (или мы хотя бы с уверенностью предполагаем) количество кластеров, либо известно, что точки линейно разделимы, может оказаться гораздо более уместным использование простых алгоритмов кластеризации. Подробнее эти вопросы рассмотрены в [18].

В данной статье предлагается алгоритм кластеризации, устойчивый к выбросам. В основе алгоритма лежит рекурсивное разбиение пространства признаков на ячейки, анализ плотности и однородности и слияние ячеек в общие кластеры. Произведены тестовые вычисления на классических двумерных наборах данных. Среди существующих методов кластеризации, предлагаемый метод имеет схожие черты с методом DBSCAN, в котором кластеризация так же производится на основе анализа плотности. Программная реализация данного алгоритма была выполнена на языке Python, с использованием библиотеки numpy. Ссылка на файлы в https://github.com/ordevoir/clustering.

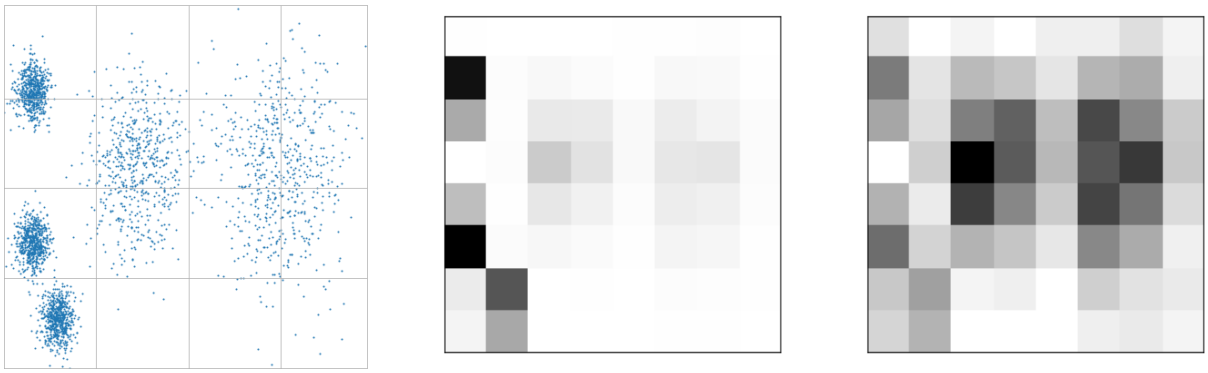
Описание алгоритма

Главный принцип алгоритма заключается в рекурсивном разбиении пространства признаков на равные кубические области. В общем случае для n-мерного пространства признаков это будут n-мерные гиперкубы, поэтому для определенности я буду называть их клетками. Так, при разбиении клетки образуются его субклетки, а родительскую клетку для заданной клетки можно назвать суперклеткой. На каждом уровне разбиения анализируются две характеристики клетки: средние плотности субклеток и равномерность распределения плотности разбиваемой клетки. На основе этих характеристик можно принять решение о том, является ли данная клетка целиком частью некоторого кластера. Главное предположение заключается в том, что клетка считается частью кластера в том случае, если его средняя плотность и равномерность распределения превышают определенные пороговые значения. Так, при каждом разбиении образуется множество (возможно пустое) клеток, удовлетворяющих условию. В результате анализа соседствующих областей, клетка из этого множества либо примыкает к существующему класте-

ру, либо образует новый кластер и сливается с другими клетками. Рассмотрим алгоритм подробней:

Этап инициализации. В первую очередь пространство признаков масштабируется для приведения к кубическому виду. В общем случае, в масштабировании пространства признаков нет необходимости, но в любом случае исходная область значений должна быть выбрана кубической. Эта область значений будет исходной клеткой нулевого уровня разбиения. Вычисляется средняя плотность клетки нулевого уровня.

Первый этап. Производится разбиение на клетки первого уровня. Вычисляются их средние плотности. Клетки, плотность которых ниже некоторого порогового значения помечаются как пустые и отсеиваются. Дальнейшая работа ведется только с непустыми клетками. На основе значений средних плотностей клеток первого уровня вычисляется равномерность распределения плотностей для клетки нулевого уровня (его однородность). На рис.1 представлены карты распределения плотностей и однородностей клеток для распределения капель.



а) б) в)

Рис. 1. Карты распределения плотностей и однородностей клеток для распределения точек: а) распределение точек, б) распределение плотностей клеток, в) распределение однородностей клеток.

[Figure 1.Distribution maps of cell densities and homogeneity for droplet distribution: a) dots distribution, b) cell densities distribution, c) cell homogeneity distribution.]

Второй этап. Производится разбиение непустых клеток первого уровня на клетки второго уровня. Вычисляются средние плотности клеток второго уровня, отсеиваются пустые клетки. На основе средних плотностей клеток второго уровня вычисляется однородность для каждой непустой клетки первого уровня. На данном этапе имеются данные о плотностях и однородностях клеток первого уровня. Непустые клетки первого уровня помечаются как однородные и неоднородные.

Третий этап. Из множества однородных клеток выбирается одна, которая образует первый кластер. Другие клетки из этого множества также могут оказаться элементами этого кластера. Поэтому кластер производит экспансию на множестве оставшихся однородных клеток. Алгоритм экспансии кластера будет рассмотрен отдельно. В результате экспансии кластера подмножество однородных клеток входит в состав кластера. После из оставшихся клеток выбирается произвольная, ко-

торая образует второй кластер. Второй кластер аналогично первому производит экспансию. Этот процесс продолжается до полного исчерпания множества однородных клеток первого уровня.

Четвертый этап. Переход на клетки второго уровня. Производится разбиение неоднородных клеток второго уровня. Вычисляются их средние плотности, отсеиваются пустые клетки, оставшиеся клетки помечаются как однородные и неоднородные, аналогично тому, как это производилось с клетками первого уровня на втором этапе.

Пятый этап. Производится экспансия уже существующих кластеров на однородные клетки второго уровня. В результате этого остается множество клеток, не примкнувших к кластерам. Эти клетки должны образовать новые кластеры по тому же принципу, который описан на третьем этапе.

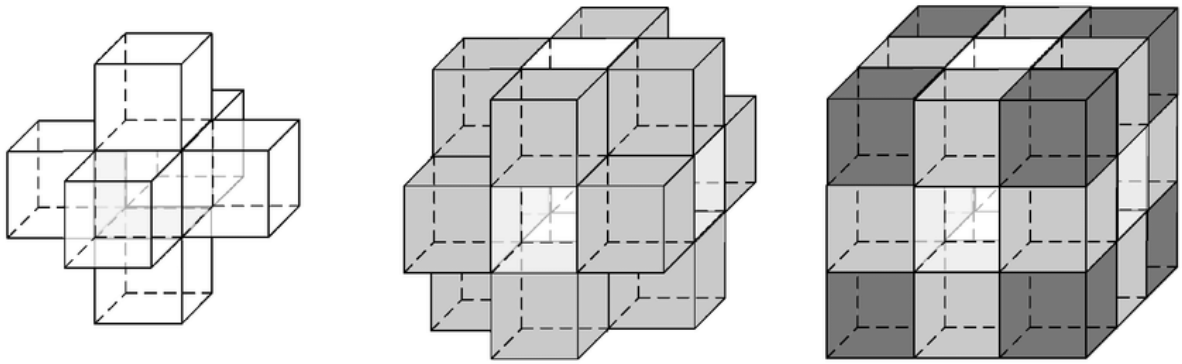
При переходе на третий уровень и далее производятся алгоритмы, идентичные четвертому и пятому этапам. Таким образом обеспечивается рекурсивное погружение до необходимой глубины, позволяющее обнаруживать мелкие кластеры и повышать детализацию.

Алгоритм экспансии кластера

При экспансии кластера клетки, входящие в кластер помечаются как активные и пассивные. Это делается для того, чтобы исключить из дальнейших вычислений клетки, которые не будут участвовать в дальнейшей экспансии. В частности, это внутренние клетки, которые окружены со всех сторон клетками собственного кластера, а также внешние клетки, окруженные клетками собственного кластера и пустыми клетками. Таким образом, после того, как кластер полностью будет расширен на данном уровне, активные клетки кластера будут граничить с неоднородными клетками, на которые кластер потенциально может расшириться на следующем уровне.

Для гиперкуба количество различных видов соседей окрестности первого порядка определяется размерностью. В двумерном случае $(n = 2)$, для квадратов существует два типа соседства: квадраты, смежные по сторонам и квадраты, смежные по вершинам. В трехмерном случае $(n = 3)$ существует три типа соседства: кубы, смежные по граням; кубы, смежные по ребрам и кубы, смежные по вершинам. Каждый из этих типов может также характеризовать «близость» смежных кубов, в зависимости от того, какой размерности будет множество общих точек кубов. Множество общих точек для кубов, смежных по сторонам, представляет поверхность (размерность $n-1$). Множество точек для кубов, смежных по ребрам, образуют линию (размерность $n-2$). Смежные по вершинам кубы имеют по одной общей точке $(n-3)$. В общем случае, число различных типов соседства в окрестности первого порядка для гиперкуба равно его размерности. Функция экспансии использует два ближайших типа соседства: клетки, смежные по гиперграням $(n-1)$ и по гиперребрам $(n-2)$. На рис.2 представлены соседи для трехмерной клетк (cube). На первом изображении представлены клетки, смежные по гиперграням,

на втором изображении – клетки, смежные по гиперребрам, а на третьем изображении – клетки, смежные по гипервершинам [19]. Условия расширения на клетки,



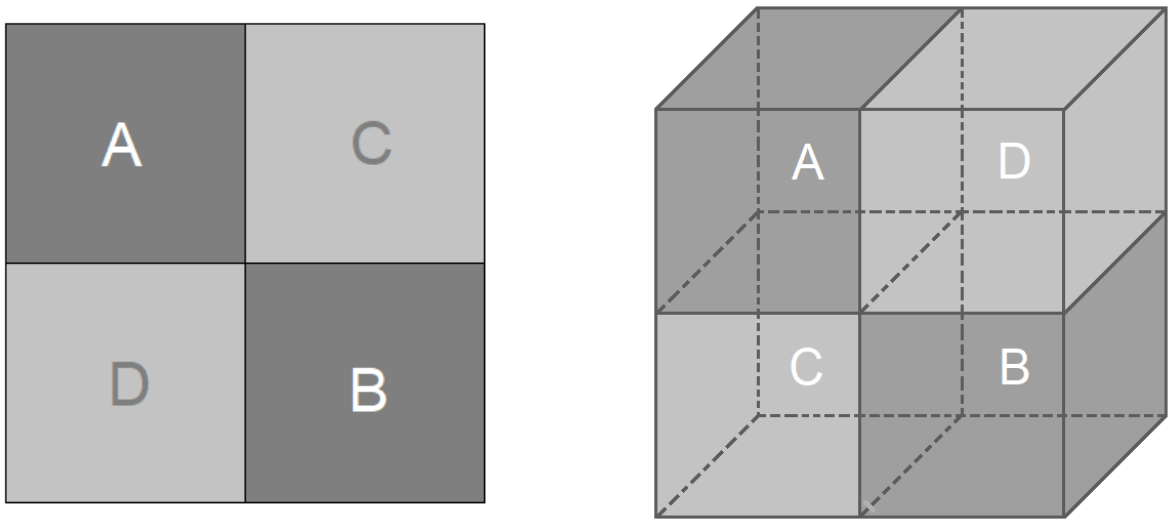
а) б) в)

Рис. 2. Соседи для 3D клеток: а) соседи по гипергранням, б) соседи по гиперребрам, в) соседи по гипервершинам.
[Figure 2. Neighbors for 3D cells: a), b), c).]

соседние по гипергранням и гиперребрам различны. В первом случае достаточным условием для расширения является плотность и однородность клетки, превышающие заданные значения. Во втором случая добавляется дополнительное условие. Для пары клеток, смежных по гиперребрам существуют клетки, смежные по гипергранням с каждой из этой пары клеток. Назовем их клетками дополнения для пары клеток. Дополнительное условие заключается в том, чтобы хотя бы одна из клеток дополнения, которая будет заведомо неоднородной, обладала плотностью, превышающей заданное значение (это значение должно быть меньше, чем пороговое значение плотности, используемое при расширении на клетки, соседние по гипергранням). Без этого дополнительного условия существенно возрастает риск нежелательного слияния кластеров. На рис.3 представлены клетки дополнения для двух клеток, смежных по гиперребрам.

Алгоритм:

препарирование данных
инициализация
for i = 1 **to** max_depth **do**
  cells = список клеток i-го уровня
  homo_cells = {cells[j] | cells[j].homogeneity > threshold}
  экспансия кластеров на однородные клетки i-го уровня
  while homo_cells ≠ ∅ **do**
    cluster = создание нового кластера
    экспансия cluster на однородные клетки i-го уровня
  end while
end for



а) б)

Рис. 3. Для клеток A и B, смежных по гиперребрам, клетками дополнения будут клетки C и D: а) двумерные клетки, б) трехмерные клетки.
[Figure 3. For cells A and B adjacent along hyperedges, the cells of the complement will be cells C and D: a) 2D cells, b) 3D cells.]

Параметры алгоритма

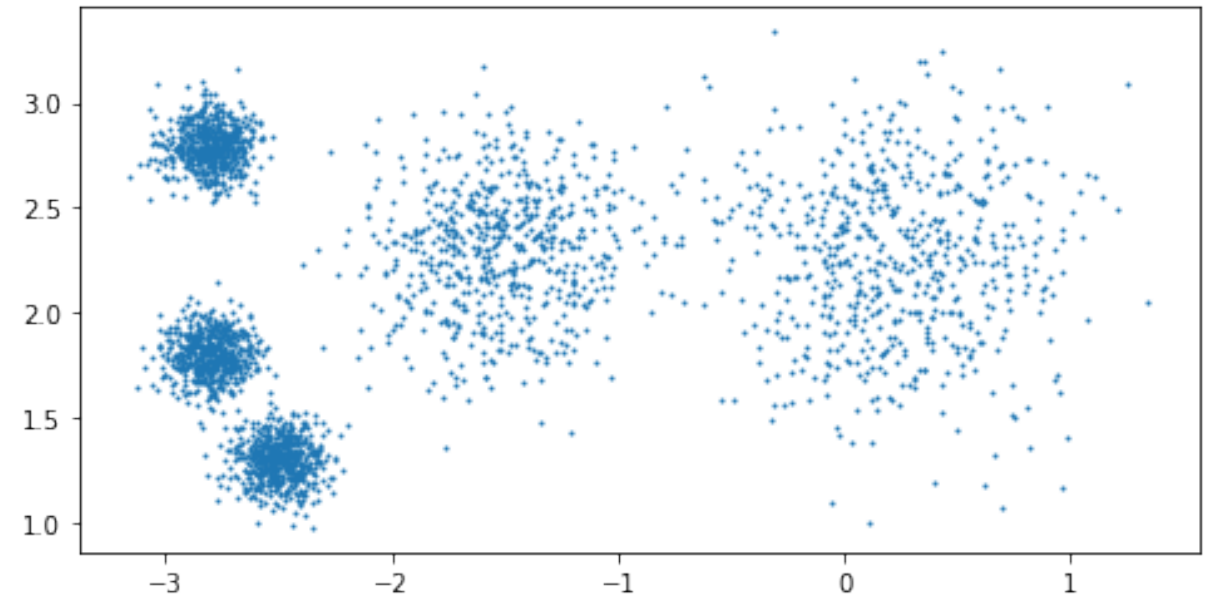
В алгоритме содержится 4 параметра, задаваемые перед началом кластеризации. Основной параметр алгоритма указывает, на сколько частей будет разбиваться ребро клетки (в программном коде этот параметр представлен как DENOMINATOR). Чем выше это значение – тем более раздробленным будет пространство на каждом этапе разбиения. Параметр DENSITY определяет порог плотности, ниже которого клетка помечается как мертвая. Параметр HOMOGENEITY определяет значение, в соответствии с которым клетка будет помечена как однородная или неоднородная. Параметр ADJACENCY определяет порог плотности для клеток дополнения, используемых для пар клеток, смежных по гиперребру. От этих параметров существенно зависит качество результата. От параметра DENSITY зависит робастность алгоритма, высокое значение пороговой плотности позволяет лучше отсеивать выбросы, но вместе с тем возникает риск отбросить клетки, которые должны были бы образовать кластеры. Параметры HOMOGENEITY и ADJACENCY характеризуют чувствительность алгоритма: при высоких значениях может происходить слишком сильное разбиение, при низких – напротив, нежелательно слияние кластеров.

Тестирования на наборах данных

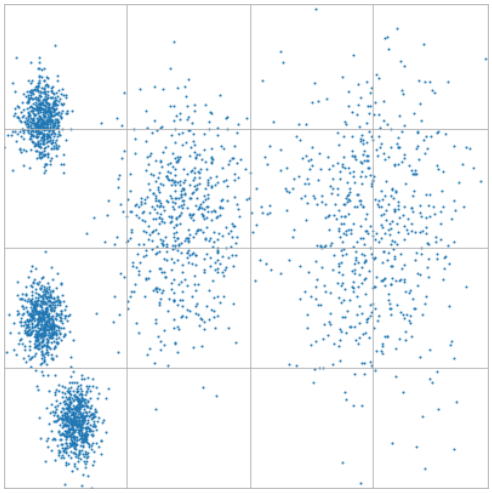
Для тестирования работы алгоритма были сгенерированы наборы данных bubbles, moons, circles. В каждом наборе данных содержится по 2500 точек. Для

каждого случая значения параметров DENOMINATOR и ADJACENCY выбраны равными 4 и 0.1 соответственно.

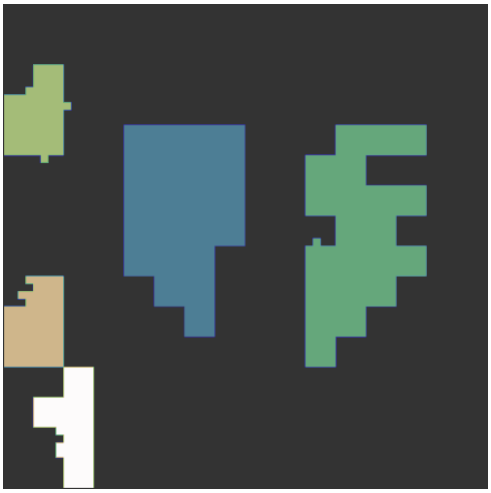
На рис.4 представлена работа алгоритма на распределении точек типа bubles. На первой изображении представлено непосредственно распределение точек. На втором изображении – результат масштабирования. На третьем изображении представлены кластеры, сформированные алгоритмом. Как и ожидается, алгоритм сформировал 5 кластеров. Значения параметров DENSITY и HOMOGENEITY равны 2 и 6.5 соответственно. На рис.5 представлена работа алгоритма на распреде-



а)



б)

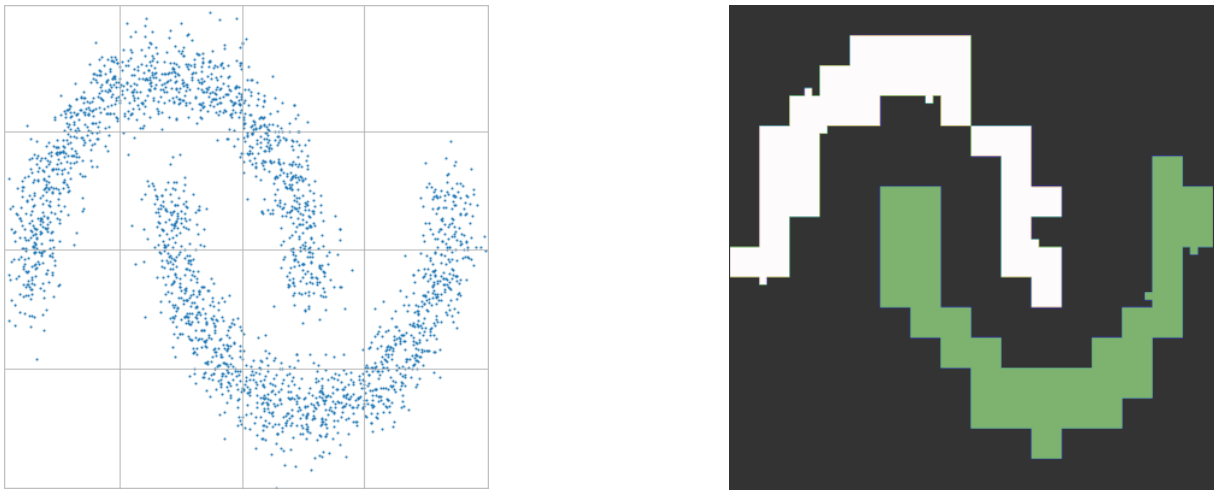


в)

Рис. 4. а) начальное распределение точек, б) результат масштабирования, выполненного для приведения признакового пространства к квадратному виду, в) результат кластеризации.

[Figure 4.a) The initial distribution of points, b) is the result of the scaling performed to bring the feature space to a square form, c) the result of clustering.]

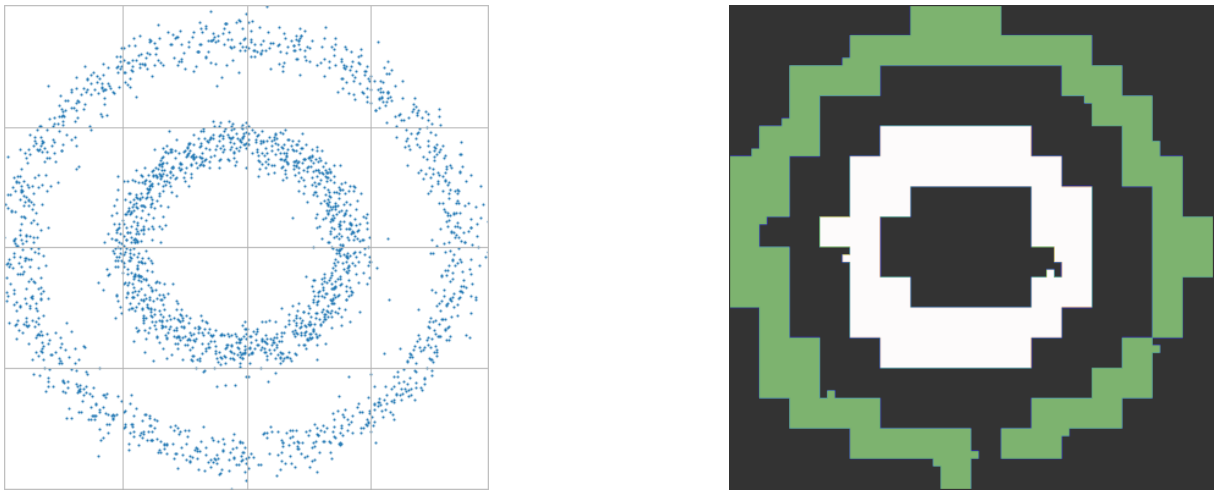
лении точек типа moons. На первой изображении представлено непосредственно распределение точек. На втором изображении представлены кластеры, сформированные алгоритмом. Алгоритм сформировал 2 кластера. Значения параметров DENSITY и HOMOGENEITY равны 2 и 6.5 соответственно. Заметим, что в данном случае можно наблюдать связность кластера за счет клеток, смежных по гиперребру, в то время как на рис.4 смежность по гиперребру клеток разных кластеров не привела к их слиянию. На рис.6 представлена работа алгоритма на распределении



а) б)

Рис. 5. а) Распределение точек moons, б) результат кластеризации.
[Figure 5.a) Moons point distribution, b) result of clustering.]

точек типа circles. На первой изображении представлено непосредственно распределение точек. На втором изображении представлены кластеры, сформированные алгоритмом. Алгоритм сформировал 2 кластера. Значения параметров DENSITY и HOMOGENEITY равны 2 и 6.5 соответственно (так же как и для moons). Как и на рис.5 связность кластера образуется за счет клеток, смежных по гиперребру. В нижней части рисунка можно заметить разрыв области кластера. Очевидно, что это произошло из-за того, что в этом месте образовалась разреженность в данных. Можно сказать, что эта проблема противоположная проблеме выбросов. В ином случае это могло привести к нежелательному расщеплению кластера. Как видно из диаграмм, предлагаемый метод кластеризации способен найти линейно неразделимые кластеры. При этом, в отличие от большинства методов кластеризации (в том числе и в методе DBSCAN), в предлагаемом методе не используется расстояние между точками пространства признаков. Алгоритм можно отнести как к Density-based clustering [20], [21] так и к Connectivity-based clustering, так как с одной стороны алгоритм использует области с высокой плотностью, а с другой стороны объекты, расположенные рядом считаются более близкими, чем объекты, далекие друг от друга. Соответственно, дальнейшее развитие алгоритма может позволить производить иерархическую кластеризацию. Очевидна устойчивость к выбросам, достигаемая во многом за счет того, что клетки плотность которых ниже пороговой исключаются из вычислений. Алгоритм не нуждается в предва-



а) б)

Рис. 6. а) Распределение точек circles, б) результат кластеризации.
[Figure 6.a) Circles point distribution, b) result of clustering.]

рительном задании количества кластеров и при оптимальных параметрах успешно формирует корректное количество кластеров. Алгоритм является довольно устойчивым к выбросам.

Недостатки. Главным недостатком метода является существенная зависимость результата кластеризации от параметров. В зависимости от распределения плотностей и неоднородностей, характерного для всего набора данных, может возникнуть необходимость коррекции параметров, для достижения желаемого результата. При этом не возникает необходимости корректирования параметров в зависимости от формы и размеров искомых кластеров. Это обстоятельство может в дальнейшем позволить разработать алгоритм априорной оценки оптимальных значений параметров. К другому недостатку следует отнести недостаточная исследованность алгоритма, в частности, на многомерных пространствах признаков, хотя программная реализация позволяет работать с произвольным количеством измерений. Также стоит сказать, что нет уверенности в правильности анализа пары клеток смежных по гиперребрам. Возможно, вместо проверки клеток, смежных с каждой клеткой из рассматриваемой пары, следует проверять субклетки, а не клетки того же уровня либо измерять градиент плотности клеток. Остается множество вопросов для дальнейших исследований.

Заключение

Предлагаемый в работе алгоритм работающий на основе рабиения пространства признаков и анализе плотностей позволяет производить робастную кластеризацию линейно неразделимых точек. Алгоритм автоматически определяет корректное количество кластеров. Представлена программная реализация алгоритма. Произведенные тесты на классических распределениях точек позволяют оценить эффективность алгоритма. Алгоритм не лишен недостатков, однако в перспективе

видится возможность работы над ними, а также возможность производить иерархическую кластеризацию.

Конкурирующие интересы. Конфликтов интересов в отношении авторства и публикации нет.

Авторский вклад и ответственность. Автор участвовал в написании статьи и полностью несет ответственность за предоставление окончательной версии статьи в печать.

Благодарность. Также авторы могут выразить благодарности своим коллегам за обсуждение и подготовку статьи к печати, а также рецензентам за ценные замечания.

Список литературы

1. Géron A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.*. O'Reilly Media, Inc.: 2st edition, 2019. 856 pp.
2. Raschka S. *Python machine learning*. Packt publishing ltd: 1st edition, 2015. 456 pp.
3. Müller A. C., Guido S. *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media: 1st edition, 2016. 398 pp.
4. MacQueen J. B Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967. vol. 1, pp. 281–297.
5. Lloyd S., Stuart P. Least square quantization in PCM, *IEEE Transactions on Information Theory*, 1982. vol. 28, no. 2, pp. 129–137.
6. Sibson R. SLINK: an optimally efficient algorithm for the single-link cluster method, *The Computer Journal. British Computer Society*, 1973. vol. 16, no. 1, pp. 30–34.
7. Defays D. An efficient algorithm for a complete link method, *The Computer Journal. British Computer Society*, 1977. vol. 20, no. 4, pp. 364–366.
8. Ester M., Kriegel H. P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise, *KDD.*, 1996. vol. 96, no. 34, pp. 226–231.
9. Sander J. et al. Density-based clustering in spatial databases: The algorithm gdbscan and its applications, *Data mining and knowledge discovery*, 1998. vol. 2, no. 2, pp. 169–194.
10. Shibzukhov Z. M. On the Principle of Empirical Risk Minimization Based on Averaging Aggregation Functions, *Doklady Mathematics*, 2017. vol. 96, no. 2, pp. 494–497 DOI: 10.1134/S106456241705026X.
11. Shibzukhov Z. M. On a Robust Approach to Search for Cluster Centers, *Automation and Remote Control*, 2021. vol. 82, no. 10, pp. 1742–1751 DOI: 10.1134/S0005117921100118.
12. Shibzukhov Z. M. Machine Learning Based on the Principle of Minimizing Robust Mean Estimates, *Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA*AI 2020*, 2020. vol. 1310, pp. 472–477 DOI: 10.1007/978-3-030-65596-956.
13. Kharinov M. V. Superpixel Clustering, *International Russian Automation Conference (RusAutoCon). – IEEE*, 2021, pp. 303–308 DOI: 10.1109/RusAutoCon52004.2021.9537461.
14. Huang D., Wang C. D., Lai J. H. Locally weighted ensemble clustering, *IEEE transactions on cybernetics*, 2017. vol. 48, no. 5, pp. 1460–1473 DOI: 10.1109/TCYB.2017.2702343.
15. Debnath T., Song M. Fast Optimal Circular Clustering and Applications on Round Genomes, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021. vol. 18, no. 6, pp. 2061–2071 DOI: 10.1109/TCBB.2021.3077573.
16. Nock R., Nielsen F. On Weighting Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006. vol. 28, no. 8, pp. 1223–1235 DOI: 10.1109/TPAMI.2006.168.
17. Kaur P. J. et al. Cluster quality based performance evaluation of hierarchical clustering method, *1st International Conference on Next Generation Computing Technologies (NGCT). – IEEE*, 2015, pp. 649–653 DOI: 10.1109/NGCT.2015.7375201.
18. Flach P. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press: 1st edition, 2012. 416 pp.

19. Shu M. L. et al. Planning the obstacle-avoidance trajectory of mobile anchor in 3D sensor networks, *Science China Information Sciences*, 2015. vol. 58, no. 10, pp. 1–10 DOI: 10.1007/s11432-015-5354-2.

20. Ankerst M., Breunig M., Kriegel H. P., Sander J. OPTICS: Ordering Points To Identify the Clustering Structure, *ACM SIGMOD international conference on Management of data. ACM Press.*, 1999. vol. 28, no. 2, pp. 49–60 DOI: 10.1145/304181.304187.

21. Achtert, E., Böhm, C., Kröger, P. 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*, 2006. vol. 3918, pp. 119–128 DOI: 10.1007/1173113916.

Казаков Мухамед Анатольевич – младший научный сотрудник отдела Нейроинформатики и машинного обучения, Институт прикладной математики и автоматизации, Кабардино-Балкарская Республика, г. Нальчик, Россия, ORCID 0000-0002-5112-5079.

MSC 68T27 Research Article

Clustering algorithm based on feature space partitioning

M. A. Kazakov

Institute of Applied Mathematics and Automation KBSC RAS,
360000, Nalchik, Shortanova st., 89a, Russia
E-mail: kasakow.muchamed@gmail.com

A new approach to robust clustering is proposed based on recursive partitioning of the feature space and density analysis. An algorithm for robust clustering of linearly inseparable points, its software implementation, as well as test results on classical data distributions are presented.

Key words: clustering, robust clustering, machine learning.

DOI: 10.26117/2079-6641-2022-39-2-136-149

Original article submitted: 05.07.2022 Revision submitted: 22.08.2022

For citation. Kazakov M. A. Clustering algorithm based on feature space partitioning. *Vestnik KRAUNC. Fiz.-mat. nauki.* 2022, **39**: 2, 136-149. DOI: 10.26117/2079-6641-2022-39-2-136-149

Competing interests. The author declares that there are no conflicts of interest with respect to authorship and publication.

Contribution and responsibility. The author contributed to the writing of the article and is solely responsible for submitting the final version of the article to the press. The final version of the manuscript was approved by the author.

References

[1] Geron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc., 2019, pp. 856.

[2] Raschka S. Python machine learning. Packt publishing ltd: 1st edition, 2015, pp. 456.

[3] Muller A. C., Guido S. Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media: 1st edition, 2016, pp. 398.

[4] MacQueen J. B. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, vol. 1, pp. 281–297.

[5] Lloyd S., Stuart P. Least square quantization in PCM, IEEE Transactions on Information Theory, 1982, vol. 28, no. 2, pp. 129–137.

Funding. The work was done without financial support

[6] Sibson R. SLINK: an optimally efficient algorithm for the single-link cluster method, The Computer Journal. British Computer Society, 1973, vol. 16, no. 1, pp. 30–34.
[7] Defays D. An efficient algorithm for a complete link method, The Computer Journal. British Computer Society, 1977. vol. 20, no. 4, pp. 364–366.
[8] Ester M., Kriegel H. P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise, KDD, 1996. vol. 96, no. 34, pp. 226–231.
[9] Sander J., et al. Density-based clustering in spatial databases: The algorithm gdbscan and its applications, Data mining and knowledge discovery, 1998, vol. 2, no. 2, pp. 169–194.
[10] Shibzukhov Z. M. On the Principle of Empirical Risk Minimization Based on Averaging Aggregation Functions, Doklady Mathematics, 2017, vol. 96, no. 2, pp. 494–497.
[11] Shibzukhov Z. M. On a Robust Approach to Search for Cluster Centers, Automation and Remote Control, 2021. vol. 82, No 10, pp.1742–1751 DOI: 10.1134/S0005117921100118.
[12] Shibzukhov Z. M. Machine Learning Based on the Principle of Minimizing Robust Mean Estimates, Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA*AI 2020, 2020. vol. 1310, pp. 472–477. DOI:10.1007/978-3-030-65596-956.
[13] Kharinov M. V. Superpixel Clustering, International Russian Automation Conference (RusAutoCon). IEEE, 2021, pp. 303–308. DOI: 10.1109/RusAutoCon52004.2021.9537461.
[14] Huang D., Wang C. D., Lai J. H. Locally weighted ensemble clustering, IEEE transactions on cybernetics, 2017. vol. 48, no. 5, pp. 1460–1473.DOI: 10.1109/TCYB.2017.2702343.
[15] Debnath T., Song M. Fast Optimal Circular Clustering and Applications on Round Genomes, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021. vol. 18, no. 6, pp. 2061–2071. DOI: 10.1109/TCBB.2021.3077573.
[16] Nock R., Nielsen F. On Weighting Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006. vol. 28, No 8, pp. 1223–1235. DOI: 10.1109/TPAMI.2006.168.
[17] Kaur P. J., et al. Cluster quality based performance evaluation of hierarchical clustering method, IEEE, 2015, pp. 649–653. DOI: 10.1109/NGCT.2015.7375201.
[18] Flach P. Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press: 1st edition, 2012, pp. 416.
[19] Shu M. L., et al. Planning the obstacle-avoidance trajectory of mobile anchor in 3D sensor networks, Science China Inform. Sci., 2015, vol. 58, no. 10, pp. 1–10.
[20] Ankerst M., Breunig M., Kriegel H. P., Sander J. OPTICS: Ordering Points To Identify the Clustering Structure, ACM SIGMOD international conference on Management of data. ACM Press., 1999. vol. 28, No 2, pp. 49–60. DOI: 10.1145/304181.304187.
[21] Achtert E., Bohm C., Kroger P. 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, 2006, vol. 3918, pp. 119–128. DOI: 10.1007/1173113916.

Kazakov Mukhamed Anatolevich – Junior Researcher of the Department of Neural Networks and Machine Learning, Institute of Applied Mathematics and Automation, Kabardino-Balkar Republic, Nalchik, Russia, ORCID 0000-0002-1576-1860.