*Федеральное государственное бюджетное образовательное учреждение высшего образования*

***«Московский государственный технический университет***

***имени Н.Э. Баумана***

***(национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)***

ФАКУЛЬТЕТ «Фундаментальные науки»

КАФЕДРА «Вычислительная математика и математическая физика» (ФН-11)

**Р А С Ч Ё Т Н О - П О Я С Н И Т Е Л Ь Н А Я З А П И С К А**

**к курсовой работе на тему:**

Кластерный анализ в биоинформатических приложениях.

Дисциплина: Основы сеточных методов

Студент

А.А. Соколов

(И.О.Фамилия)

Руководитель курсовой работы

Т.В. Облакова

(И.О.Фамилия)

Москва, 2019

Список исполнителей

|  |  |  |
| --- | --- | --- |
| Руководитель курсовой работы | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Т.В. Облакова |
| Исполнитель | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | А.А. Соколов |

Реферат

Отчет выполнен в 1 части и содержит: страниц — , иллюстраций — , таблиц — , приложений — , в отчете использовано источников — 12.

Содержание

[Список исполнителей 2](#_Toc40476698)

[Реферат 3](#_Toc40476699)

[Содержание 4](#_Toc40476700)

[Термины и определения 6](#_Toc40476701)

[Перечень сокращений и обозначений 7](#_Toc40476702)

[Введение 8](#_Toc40476703)

[Основная часть КУРСОВОЙ РАБОТЫ 9](#_Toc40476704)

[1 Предпосылки 9](#_Toc40476705)

[1.1 Машинное обучение и интеллектуальный анализ данных 9](#_Toc40476706)

[1.1.1 Обучение с учителем 10](#_Toc40476707)

[1.1.2 Обучение без учителя 11](#_Toc40476708)

[1.2 Биоинформатика 12](#_Toc40476709)

[2 Обзор современных семейств кластеризации 15](#_Toc40476710)

[2.1 Введение 15](#_Toc40476711)

[2.2 Меры близости 16](#_Toc40476712)

[2.3 Метрики для дискретных признаков 18](#_Toc40476713)

[2.3.1 Расстояние Хэмминга 18](#_Toc40476714)

[2.3.2 Коэффициент соответствия 18](#_Toc40476715)

[2.4 Метрики для непрерывных признаков 19](#_Toc40476716)

[2.5 Семейства кластеризации 20](#_Toc40476717)

[2.5.1 Partitional Clustering 21](#_Toc40476718)

[2.5.2 Иерархическая кластеризация 21](#_Toc40476719)

[2.5.3 Нечеткая кластеризация 21](#_Toc40476720)

[2.5.4 Кластеризации на основе нейронный сетей 22](#_Toc40476721)

[2.5.5 Примесные модели кластеризации 22](#_Toc40476722)

[2.5.6 Графовые модели кластеризации 23](#_Toc40476723)

[2.5.7 Ансамблевая кластеризация 23](#_Toc40476724)

[2.5.8 Бикластеризация 23](#_Toc40476725)

[2.6 Кластеры и разбиения (Clusters and Partitions) 24](#_Toc40476726)

[2.7 Подытог 24](#_Toc40476727)

[3 Partional clustering 25](#_Toc40476728)

[3.1 Введение 25](#_Toc40476729)

[3.2 k-Means 27](#_Toc40476730)

[3.2.1 Принцип 27](#_Toc40476731)

[3.2.2 Ядерный k-means 34](#_Toc40476732)

[3.3 k-Medoids 37](#_Toc40476733)

[3.3.1 Принцип 37](#_Toc40476734)

[3.3.2 Сравнение k-means и k-medoids. Confusion matrix. Метрики ошибок 38](#_Toc40476735)

[3.3.3 Замечание по поводу метрик ошибок 44](#_Toc40476736)

[Заключение 46](#_Toc40476737)

[Список использованных источников 47](#_Toc40476738)

[Приложения 53](#_Toc40476739)

Термины и определения

Перечень сокращений и обозначений

Введение

Основная часть КУРСОВОЙ РАБОТЫ

Предпосылки

. Машинное обучение и интеллектуальный анализ данных

Машинное обучение — это область искусственного интеллекта, которая занимается разработкой алгоритмов и методов, которые учатся на эмпирических данных. По наличию или отсутствию верно размеченных исходных данных данные методы в основном разделяют на две группы: обучение с учителем (supervised learning) и обучение без учителя (unsupervised learning). В случае обучения с учителем, мы должны предоставить алгоритму репрезентативный набор данных с верно размеченными исходами train), на котором алгоритм должен будет научиться предсказывать неизвестные новые, неразмеченные данные (test data). Обучение без учителя же, напротив, направлено на изучение скрытых структур в данном алгоритму наборе немаркированных данных. Стоит отметить, что существует множество других отраслей машинного обучения, в том числе обучение с частичным привлечением учителя (semi-supervised learning), обучение с подкреплением (reinforcement learning) и многие другие.

Целью интеллектуального анализа данных (data mining) является обнаружение закономерностей и полезных фрагментов информации в большом наборе данных. Методы машинного обучения используются при извлечении данных в дополнение к другим методам, например, из прикладной статистики, дискретной математики и методам оптимизации. Ключевым аспектом, который позволяет анализу интеллектуального анализа данных использовать методы машинного обучения, является возможность форматировать и представлять добытые данные в структуре, доступной для методов машинного обучения. Такие данные обычно организованы как набор объектов с количественными характеристиками.

.. Обучение с учителем

Репрезентативный набор известных и верно размеченных данных, которые используются в обучении с учителем принято называть обучающая выборка или train data, или learning data. Обучающая выборка обычно представляет собой набор примеров, отклик которых уже известен. Отклики примеров представляются собой «правильные ответы» на вопрос, который исследуется для данного конкретного объекта. Иллюстрированным примером является набор фотографий котов и собак, который используется для обучения с учителем, чтобы ответить на вопрос «на фотографии представлен кот или собака?». Обучающая выборка изначально считается верно размеченной, то есть правильный ответ на вопрос о том, какое животное представлено на фотографии, имеется для каждой фотографии. Если обучение прошло успешно, то данные алгоритм сможет с высокой точностью идентифицировать животное на невидимой ранее фотографии, то есть на объекте из тестовой выборки.

В качестве примеров приведем некоторые наиболее популярные алгоритмы, использующиеся в задачах линейной и нелинейной классификации:

* метод k-ближайших соседей [1];
* искусственные нейронные сети [2];
* метод опорных векторов [3];
* линейный дискриминантный анализ [4];
* наивный байесовский классификатор [5];
* дерево решений [6];
* генетическое программирование [7].

Это далеко не полный список существующих алгоритмов и в настоящей курсовой работе будут рассмотрены лишь некоторые из представленных алгоритмов.

Многие линейные методы, такие как метод опорных векторов (SVM), можно сделать нелинейными, применив так называемый «ядерный трюк», который преобразует данные в другое пространство, в котором применяется первоначально линейный метод [8]. Кроме того, из-за различий между этими методами в их слабых и сильных сторонах используются ансамбли классификаторов, в которых разные методы голосуют за ответ, и ответ, который получает большинство голосов, считается окончательным консенсусным ответом [9], [10]. Общий вид ансамбля представлен на

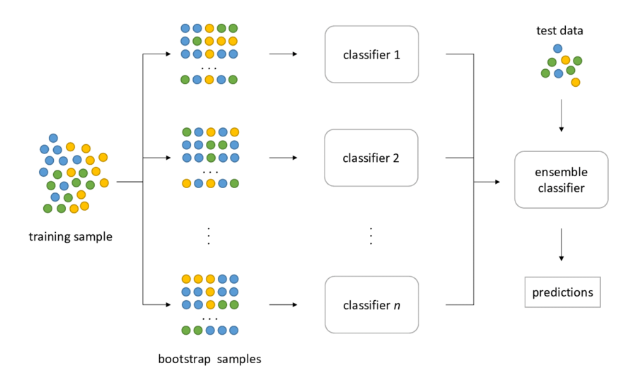


Рисунок 1 — общая структура ансамбля классификаторов.

.. Обучение без учителя

В отличие от обучения с учителем, данные, исследуемые методами обучения без учителя, не имеют изначальной маркировки верных ответов. В биоинформатических приложениях данный класс алгоритмов является доминирующим и представляет собой фокус данной курсовой работы. При кластеризации набор наблюдаемых объектов с измеренными признаками разделяется на несколько кластеров объектов так, что те объекты, которые включены в один кластер, похожи друг на друга, но не похожи на объекты, включенные в другие кластеры. Степень схожести объектов определяется изначально исследователем, исходя из различных эвристик [11]. Одной из проблем, возникающих при обучении без учителя является определение правильного числа кластеров. Это может быть оценено на основе априорных знаний, специфичных для конкретной области, или это может стать одним из этапов обучения.

Существует множество семейств методов кластеризации, которые основываются на обучении без учителя. Можно привести в качестве примеров наиболее популярные семейства с некоторыми примерами:

1. partitional clustering — метод k-means [12];
2. нейросетевые методы кластеризации — самоорганизующаяся карта Кохонена (SOM) [13] и self-organising oscillator networks [14];
3. mixture model clustering (примесные модели) — finite mixture model (модель конечных примесей) [15];
4. иерархическая кластеризация — агломеративные и дивизионные методы [16];
5. нечеткая кластеризация — метод нечёткой кластеризации C-средних [17];
6. графовая кластеризация — спектральная кластеризация [18];
7. консенсусная кластеризация [19];
8. бикластеризация [20];

а также многие другие методы.

. Биоинформатика

Современные прорывные области исследований появляются благодаря сотрудничеству исследователей из разных, порой, на первый взгляд, далеких дисциплин. Примерами могут служить квантовая информатика, нейролингвистика, биофизика, биохимия, системная инженерия, бизнес-информационные системы управления и другие. Этот союз между дисциплинами дает результаты, который смешивают взгляды ученых из разных областей на одну и ту же тему или некоторую область тем.

Стимулы, ведущие к таком сотрудничеству, многочисленны. Например, одна дисциплина может разрабатывать инструменты, которые генерируют типы данных, которые анализируются методами из другой дисциплины. В других случаях исследователи могут прийти к тому, что обнаружат, что значительная часть области их действия фактически основана на принципах другой области, например, исследование биологических процессов в клетке на уровне химических взаимодействий, привели к возникновению междисциплинарной биохимии. Междисциплинарные области, такие как бизнес-информационные системы и управленческие информационные системы, возникли из-за высокого спроса бизнеса на возникшее в недавнее время существенный рост объемов данных, который привел к невозможности использования стандартных классических методов анализа данных.

Междисциплинарным направлением, которое рассматривается в данной курсовой работы с точки зрения применимости методов машинного обучения, в частности обучения без учителя, является биоинформатика. Причиной появления этой области является все более растущие объемы необработанных биологических данных, которые связаны с развитием новых технологий секвенирования [21].

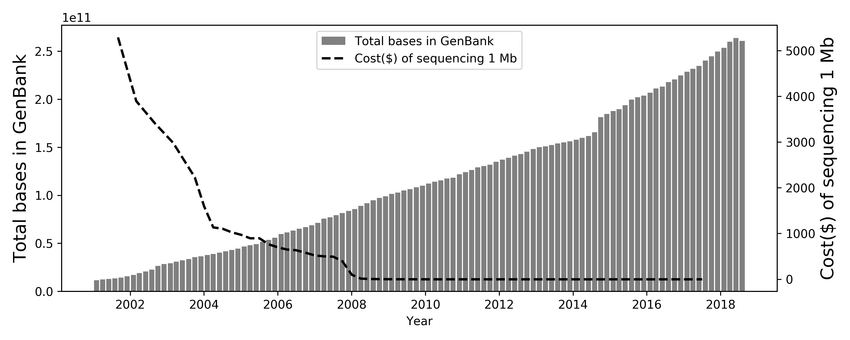


Рисунок — Растущий объем базы данных GenBank (Ncbi.nlm.nih.gov 2018), следующий за уменьшением стоимости секвенирования.

Масштаб этих данных на порядки выше тем то, что может быть эффективно обработано вручную. Следовательно к данной проблеме начали подключаться специалисты по компьютерным наукам и исследователи в области анализа данных. Совместная работа ученых из разных областей, обмен результатами и проведение междисциплинарных конференций привели и по-прежнему ведут ко многим ключевым открытиям в биологии и медицине. Такой успех привлек многих информатиков и математиков к биологии и многих биологов к компьютерным наукам, чтобы на лоне коллективного взаимодействия специалистов из разных областей привести к созданию новой междисциплинарной науки – биоинформатики.

Область биоинформатики включает разработку методов и инструментов, предназначенных для хранения, поиска, организации, анализа и представления высокопроизводительных биологических данных.

Фундаментальной основой биоинформатики принято считать молекулярную биологию. В самом общем виде каждая часть молекулярной биологии, которая связана с большими объемами данных, подвергается анализу с помощью биоинформатических методов. Это включает, например, аспекты анализа последовательностей ДНК, РН и белков, экспрессии генов и белков, анализ генетики заболеваний, включая рак, анализ регуляции генов, регуляции химических взаимодействий, сопоставимый анализ разнообразия геномов между индивидуумами или видами и многое другое.

Обзор современных семейств кластеризации

. Введение

Задача кластеризации состоит в том, чтобы сгруппировать объекты данных в набор непересекающихся классов, называемых кластерами так, чтобы объекты внутри кластеров имели высокое сходство друг с другом, тогда как объекты в отдельных классах можно было бы считать различными. Прежде чем углубиться в алгоритмы кластеризации, в этом пункте рассмотрим обзорно некоторые методы кластеризации и кратко объясним общее обоснование всех методов кластеризации.

Будем считать, что биологические данные, будь то данные об экспрессии генов или экспрессии белков, данные о последовательности ДНК или белков или другие данные представлены в виде матрицы , которую называют матрицей признаков. строк матрицы представляют собой объектов (гены, белки, метаболиты, гликаны и т.д.), столбцов — объектов (данные из разных временных точек, тканей или видов). Имея эту матрицу, можно кластеризовать ее тремя способами: во-первых, можно кластеризовать ее по строкам, чтобы сгруппировать гены или белки в кластеры, основываясь на том, насколько похожи их признаки; во-вторых, можно разбить ее по столбцам, чтобы сгруппировать признаки в кластеры, основываясь на том, сколько генов или белков ведут себя одинаково с такими признаками; и наконец, можно сгруппировать матрицу как по строкам, там и по столбцам, чтобы посмотреть на гены, которые ведут себя одинаково со многими, если не се всеми признаками вместе.

С другой стороны, биологические данные также могут быть организованы в квадратную матрицу , которая называется матрицей подобия или матрицей близости, где как ее строки, так и столбцы представляют собой объекты данных, и запись обозначает сходство или различие между -м и -м объектами. Зачастую, эта матрица является симметричной. Далее в этом пункте будут рассмотрены меры близости, которые являются фундаментальным элементом кластеризации, а затем перейдем к обзору мира алгоритмов кластеризации.

. Меры близости

Существует множество разных терминов для измерения близости, например: сходство, различие, расстояние, корреляция. Концептуально, это одно и тоже — более высокая схожесть или большая корреляция означает, что два признака более похожи или геометрически ближе. Поэтому, естественно, задать следующие вопросы: какие стандарты нужно использовать для измерения расстояния или сходства между двумя объектами и как это измерять. Иногда также требуется измерить расстояние не только между одним объектов и кластером, но и между двумя кластерами.

В литературе описано множество метрик. Они могут быть использованы в различных приложениях на основе их различных особенностей. Поскольку данные могут быть количественными и качественными, непрерывными или двоичными, номинальными и ранговыми, метрики можно грубо разделить на две категории: 1) измерение расстояний для объектов с дискретными признаками и 2) измерение расстояний для объектов с непрерывными признаками. Тем не менее, эти метрики расстояния имеют некоторые общие свойства [22]:

1. Симметричность: или для любых ;
2. Положительная определенность: или для любых ;
3. Неравенство треугольника: или для любых ;
4. Рефлексивность: или , если ,

где и обозначают меры различия и схожести между и соответственно.

Поясним эти понятия. Мера различия (dissimilarity) — это численная характеристика, которая отражает, насколько различны два объекта. Чем ниже значение, тем более похожи объекты. Минимальное значение обычно равняется . Максимальное значение может быть разным для разных реализаций конкретных мер. Мера схожести (similarity) — это численная характеристика, которая отражает, насколько два объекта схожи. Чем выше значение меры, тем более схожи объекты. Обычно . Когда говорится о близости каких-либо объектов, предполагается ясным из контекста, о чем конкретно идет речь.

Приведем простой пример для номинативного признака. Предположим, что рассматривается цвет глаз испытуемого, а также будем считать, что мы можем точно установить цвет каждого испытуемого. Тогда рассмотрим меры различия и схожести:

, ()

, ()

где и обозначают соответствующее значение признака у двух различных объектов — цвет глаз испытуемых.

. Метрики для дискретных признаков

Предположим, что дискретный признак представлен в виде , где — алфавит длины , Например, бинарный алфавит или 4-буквенный генетический код, или 20-буквенный

.. Расстояние Хэмминга

Расстояние Хэмминга между двумя дискретными признаками объекта, имеющими одинаковую длину, представляет собой число позиций, в которых соответствующие дискретные значения различны.

Пример: расстояние Хэмминга между двумя бинарными векторами `001011101` и `101010010` равняется 5, а между векторами `144321` и `144231` — 2. Наконец, если мы в предыдущем примере заменим , то получим две последовательности ДНК `ACCGTA` и `ACCTGA` с расстоянием Хэмминга, равным 3, для них.

.. Коэффициент соответствия

Определим вектор соответствий, который отражает, сколько позиций в двух векторах одинаковой длины совпадают. Элементами вектора соответствий являются , если элементы на соответствующих позициях в двух векторах совпадают, и 0 в противном случае. Обозначим количество единиц в полученном векторе за , а количество нулей за . Тогда коэффициент соответствия определяется как Например, в примере с бинарными векторами из предыдущего подпункта вектор соответствия примет вид , , и коэффициент соответствия равен , а в примере с последовательностью ДНК имеем и коэффициент соответствия равен .

. Метрики для непрерывных признаков

В случае дискретных величин была рассмотрена только метрика Хэмминга, как наиболее популярная в приложениях. Для непрерывных случаев список различных метрик куда шире. Приведем наиболее популярные в виде таблицы:

Таблица схожести — Метрики схожести и различия для непрерывных признаков

|  |  |  |
| --- | --- | --- |
| Метрика | Формула | Комментарий |
| Расстояние Минковского |  | Расстояние Минковского — это метрика в Евклидовом пространстве, которая является обобщением евклидова, манхэттенского и чебышёвского расстояний. |
| Евклидово расстояние |  | Частный случай метрики Минковского для |
| Манхэттенское расстояние |  | Частный случай метрики Минковского для |
| Чебышёвское расстояние |  | Частный случай метрики Минковского при |
| Расстояние Махаланобиса |  | **—** внутрикластерная матрица ковариации |
| Пирсоновская корреляция |  | Коэффициент корреляции Пирсона не имеет свойства положительной определенности, так как принимает значения на отрезке [-1;1]. Тем не менее, мы можем его зарескейлить до [0;1] для и |
| Jacknife корреляция |  | Jacknife корреляци в соответствии с методом складного ножа устойчива к выбросам. обозначает корреляцию между -ым и -ым значениями признака при условии удаления -ого значения. |
| Мера Оцуки — Отиаи  (косинусный коэффициент) |  | Косинусный коэффициент широко используется в анализе текстов. |

Можно отметить, что наибольшей популярностью в биоинформатических приложениях имею Евклидова метрика, корреляция Пирсона и расстояние Махаланобиса [23], [24], [25].

. Семейства кластеризации

Различные стартовые критерии могут привести к различным таксономиям алгоритмов кластеризации. Мы рассмотрим наиболее распространенную классификацию наиболее важных алгоритмов кластеризации в биоинформатических приложениях в восемь семейств, а именно: partitional clustering, иерархическая кластеризация, нечеткая кластеризация, кластеризация на основе нейронных сетей, кластеризация на основе примесных моделей, графовая кластеризация, консенсусная и бикластеризация.

.. Partitional Clustering

Частичня кластеризация, в буквальном смысле, пытается напрямую разбить входной набор данных на непересекающиеся части на основе заранее заданного некоторого критерия. Это определение предельно широкое, и может включать многие семейства кластеризации, такие как самоорганизующаяся кластеризация или примесные модели и так далее. Но на практике это семейство ограничивают алгоритмами на основе критерия минимизации квадрата ошибки. Наиболее значимым представителем данного семейства являются методы k-means и k-medoids [26]. Детальное рассмотрение алгоритмов частичной кластеризации представлено в пункте 3.

.. Иерархическая кластеризация

В отличие от частичной кластеризации, иерархическая кластеризация разбивает набор данных в иерархическую структуру вложенных кластеров, которые можно графически представить в виде дерева. Древовидная структура называется дендрограммой. Окончательные результаты кластеризации получаются путем «разрезания» полученного дерева. Иерархическая кластеризация может быть далее разделена на дивизионные (нисходящие) подходы и агломеративные (восходящие) подходы. Кроме того, в каждом из этих двух классов есть много разных алгоритмов, основанных на разных методах разделения или методах связи. Мы подробно рассмотрим иерархическую кластеризацию в пункте 4.

.. Нечеткая кластеризация

Основным преимуществом нечеткой кластеризации является то, что она ослабляет требования непересекаемости кластеров, позволяя каждому объекту быть связанным со всеми кластерами со степенью принадлежности, что означает, что каждый объект данных может быть назначен нескольким кластерам, а не только одному. Концепция данного подхода основана на нечеткой логике, следовательно, любой алгоритм, который предполагает возможность пересекаться кластерам, можно отнести к данному семейству. Наиболее популярными представителями можно считать fuzzy c-means (FCM), fuzzy k-medoids, possibilistic c-means (PCM) и прочие. Подробное рассмотрение методов нечеткой кластеризации представлено в пункте 4.

.. Кластеризации на основе нейронный сетей

Кластеризация на основе нейронных сетей начинается с набора узлов (также называемых нейронами), которые являются одинаковыми для каждого слоя, за исключением некоторых параметров, инициализированных случайным образом, которые заставляю каждый узел вести себя немного по-разному. Затем эти узлы извлекают уроки из данных на конкурентной основе: активные узлы усиливают свое присутствие, подавляя при этом активность других узлов. Типичными примерами нейросетевой кластеризации являются самоорганизующиеся карты Кохонена (SOM), самоорганизующиеся осциляторные сети (SOON), теория адаптивного резонанса (ART) и другие алгоритмы машинного обучения. Детальное рассмотрение данного семейства представлено в пункте 5.

.. Примесные модели кластеризации

Основной методов данного семейства является определение ядра кластеризации для каждого отдельного компонента с точки зрения выборочной плотности , где — это неизвестный набор параметров. Примесные модели дают значимые результаты во многих случаях, когда алгоритмы, основанные на Евклидовых метриках, терпят неудачу, особенно в анализе временных рядов (экспрессия генов во времени) и категориальных данных. В данном семействе выделяют две группы методов, а именно, модели конечной примеси (параметрические модели) и модели бесконечной примеси (непараметрические модели). Подробное рассмотрение примесных методов кластеризации представлено в пункте 6.

.. Графовые модели кластеризации

Кластеризация, основанная на графовых методах, кроме того, что способна выполнять ту же работу, что и другие методы, но также может кластеризовать реляционные данные, что обычно не под силу другим алгоритмам кластеризации. В исследованиях по молекулярной и клеточной биологии графовые модели кластеризации широко используются при анализе сетей белок-белкового взаимодействия и метаболических сетей. Подробное рассмотрение данного семейства представлено в пункте 7.

.. Ансамблевая кластеризация

Консенсусная кластеризация, также известная как ансамблевая кластеризация, используется для объединения различных результатов кластеризации и рассматривается как решение проблемы несогласованности алгоритмов стохастической кластеризации или кластеризации с различными параметрами. Ансамблевая кластеризация формализует идею о том, что объединение различных кластеризаций должно отражать общую организацию данных в различных результатах кластеризации. Подробное рассмотрение ансамблевой кластеризации представлено в пункте 8.

.. Бикластеризация

Концепция одновременной кластеризации данных в нескольких измерениях восходит к середине второй половины прошлого века, и она не привлекала большого внимания до тех пор, пока Cheng, Y. и Church, G.M. в 2000 году не разработали алгоритм двухкластерного анализа данных для экспресии генов [27]. В пункте 9 рассматривается основная концепция бикластеризации и приводятся основные типы бикластеров.

2.6 Кластеры и разбиения (Clusters and Partitions)

Кластеры и разбиения являются разными форматами представления результатов работы алгоритмов кластеризации. Первые представляют собой вектор индексов , называемый разбиением и , где — это число кластеров. Вторая форма представляет собой матрицу разбиений . При кластеризации, когда объект может быть отнесен только к одному кластеру (чёткое разбиение), -ый объект принадлежит -ому кластеру, если равняется 1, и не принадлежит в противном случае. В ситуации нечеткой кластеризации значение отражает “веру” алгоритма в то, что -ый объект принадлежит -ому кластеру. Наконец, разбиения могут быть представлены в виде набора кластеров , где обозначает -ый кластер и . Число элементов в -ом кластере обозначается за , тогда .

. Подытог

В этом пункте были введены метрики расстояния для дискретных и непрерывных признаков. Расстояние Хэмминга и коэффициент соответствия являются наиболее популярными метриками для дискретных данных. Метрики непрерывных данных приведены в таблице 1. Выбор метрики имеет решающее значение для конечных результатов кластеризации, поскольку различные расстояния могут привести к разным результатам. Поэтому метрика должна выбираться так, чтобы как можно лучше отражать реальную близость данных. Затем мы кратко представили восемь семейств кластеризации. Эти алгоритмы широко используются в молекулярной биологии, генетике, геномике, протеомике и прочих биоинформатических приложениях. Мы также определили понятия кластеров и разбиений, которые будут использоваться далее. Теперь можно приступить к технической реализации методов кластеризации.

Partional clustering

3.1 Введение

Суть кластеризации разбиениями может быть сформулирована следующим образом. Имеется набор данных из объектов в -мерном пространстве признаков, и требуется определить такое разделение на групп или кластеров, что объекты внутри кластера более схожи друг с другом, чем объекты из разных кластеров. Всего количество различных разбиений -элементного множества на непустых подмножеств определяется числом Стирлинга второго рода [28]:

, ()

где — биномиальный коэффициент (число сочетаний из по .

Например, число различных разбиений 10 объектов в 4 кластера равняется , а в случае 20 объектов уже . Поэтому полный перебор всех вариантов с целью поиска наилучшего неосуществим в вычислительном отношении. С другой стороны, введенное так определение получается слишком широким. В таком случае принято разделение разных алгоритмов с помощью различных критериев кластеризации. Для кластеризации разбиениями в основном пользуются критерием на основе ошибки квадрата. Далее будет разбираться именно этот случай.

Лучшими примерами семейства кластеризации разбиениями являются k-means и k-medoids (также известного, как разделение вокруг медоидов (PAM)). Как один из самых старых алгоритмов кластеризации [29], [30], k-means был разработан для группировки по сходству. Продолжением этой идеи стала работа, посвященной алгоритму k-medoids [31], [32].

С появлением высокопроизводительных методов сбора данных в областях молекулярной биологии, например, ДНК-микрочипы, новые методы секвенирования, в том числе методы секвенирования нового поколения (NGS), k-means и k-medoids стали широко использоваться в биоинформатических приложениях. Herwing с коллегами разработали процедуру кластеризации, основанную на модифицированном k-means, названным последовательный k-means, для высокопроизводительного анализа данных кДНК [33]. Tavazoie c коллегами использовали k-means в качестве инструмента кластеризации для анализа данных с микрочипа об экспресии генов для Saccharomyces cerevisiae (почкующиеся дрожжи) [34]. В своем исследовании хроматиновых сигнатур промоторов и энхансеров Heintzman и его коллеги использовали k-means в качестве одного из этапов своего пайплайна [35], [36]. Pollard и van de Laan проанализировали данные с микрочипа по экспрессии генов с использованием k-medoids и обосновали это выбор лучшей устойчивостью к выбросам у данного метода [37]. Harbison с коллегами разработали вычислительный метод обнаружения сайтов связывания регулятора транскрипции, где k-medoids использовался для кластеризации значимых мотивов регуляторов транскрипции [38]. Elsheikh с коллегами использовали k-means и k-medoids в исследовании рака молочной железы для выявления наличия вариантов на глобальных уровнях гистоновых маркеров в различных классах инвазивного рака молочной железы [39].

Позже появилось много обобщений и вариаций, основанных на оригинальных алгоритмах k-means и k-medoids, например, нечеткие c-means, нечеткие k-medoids (буду рассмотрены в пункте 4), ядерные k-means [40], [41], генетические k-medoids [42], генетические k-means [43], [44], сферические k-medoids [45] и многие другие. В данном пункте мы подробно рассмотрим принципы k-means и k-medoids и некоторые их вариации, а также применения.

. k-Means

3.2.1 Принцип

Алгоритм k-means является наиболее распространенным алгоритмом секционированной кластеризации. Основная идея состоит в получении разбиений, которые минимизирую квадратичную ошибку для заданного числа кластеров . Квадратичная ошибка также известна как внутригрупповая дисперсия. Предположим, что имеющееся разбиение образует набор кластеров для набора из объектов. Определим центроид (он же барицентр, он же центр тяжести) -ого кластера как средний вектор членов кластера :

. ()

Тогда квадратичная ошибка или внутриклассовая дисперсия представляет собой сумму квадратов евклидовых расстояний между всеми членами в и центроидом :

. ()

И полная квадратичная ошибка для всего разбиения представляет собой сумму внутригрупповых дисперсий по всем кластерам:

()

Критерий кластеризации на основе квадратичной ошибки состоит в том, чтобы найти оптимальное разбиение путем минимизации общей квадратичной ошибки . Данная задача имеет недетерминированную полиномиальную сложность (NP-сложная задача) [46].

На практике применяются вычислительно эффективные и быстро сходящиеся эвристики. Базовая процедура алгоритма k-means представлена в таблице 2. На первом шаге кластеры инициализируются случайным образом. Детерминированная инициализации дает нерепрезентативные результаты, которые будут зависеть от начального разбиения. Существует множество вариантов начальной инициализации алгоритма. Одним из первых является самым очевидным — кластеры выбираются абсолютно случайным образом и обновляются в соответствии с последующими шагами алгоритма. Далее был предложен метод (Forgy), заключающийся в произвольном выборе объектов в качестве центроидов и назначении остальных объектов ближайшим кластерам [47], который позже был продолжен Маккуином (MacQueen) [48].

Таблица 2 — Базовая процедура k-means

|  |  |
| --- | --- |
| Шаг 1 | Случайная (или детерминированная) инициализация кластеров; |
| Шаг 2 | Назначение каждого объекта ближайшему кластеру , где ; |
| Шаг 3 | Обновление центроидов каждого кластера, ; |
| Шаг 4 | Повторять шаги 2 и 3 до момента отсутствия изменений в кластерах. |

Понятно, что второй шаг алгоритма k-means разбивает наблюдения на диаграмму Вороного. Каждый объект данных в определенной ячейке находится ближе к центроиду в своем кластере, чем к любому другому центроиду.

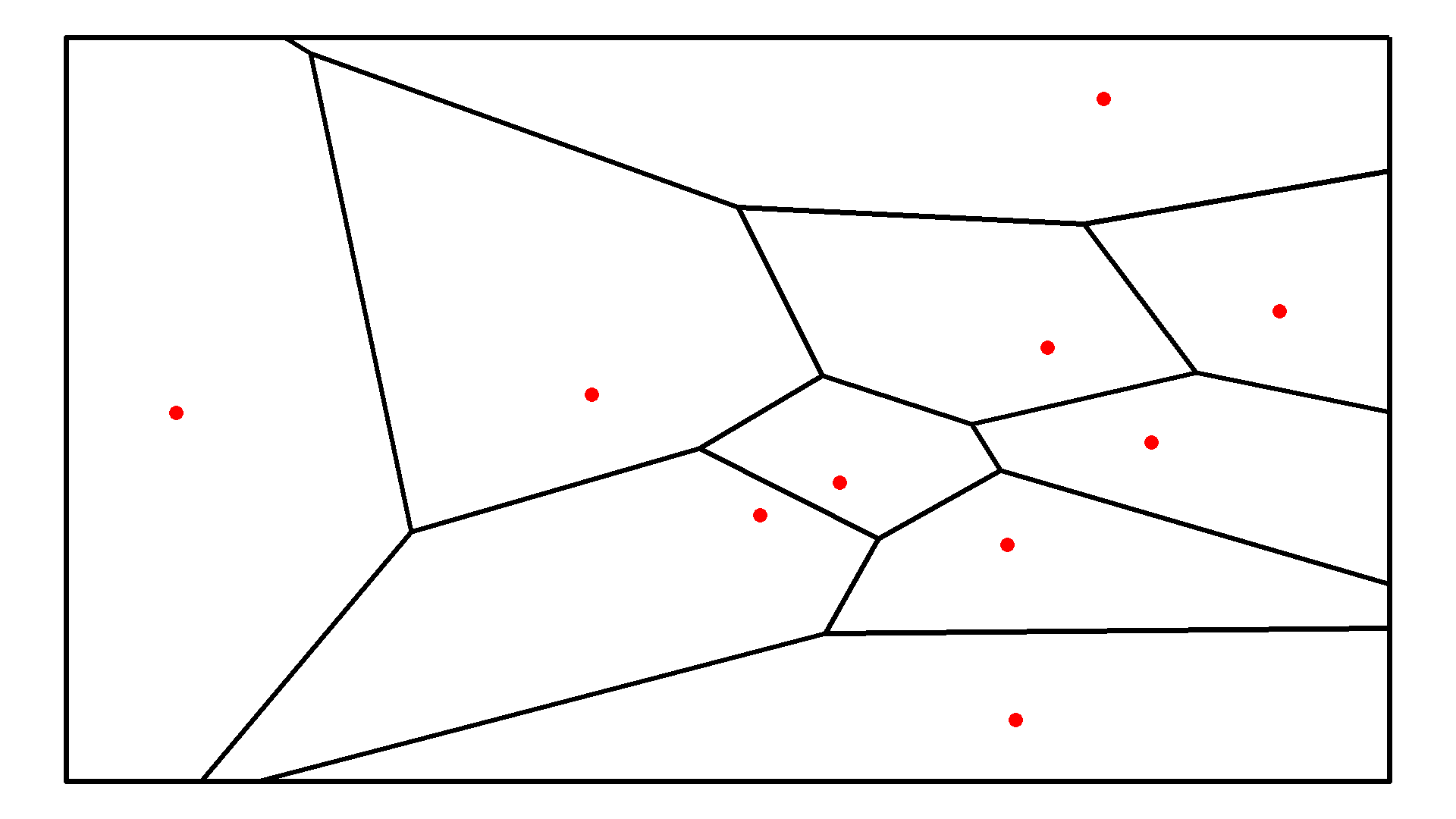


Рисунок 3 — Иллюстрация диаграммы Воронова в двумерном случае. Пространство признаков разделено на 10 ячеек. Разделяющей гиперплоскостью признаков в данном случае служат отрезки прямой.

Пример. Рассмотрим некоторую абстрактную задачу, на которой пошагово продемонстрируем принцип метода k-means (построение диаграммы Воронова). Создадим вручную 5 кластеров с 1000 точек в каждом (рис. 3).

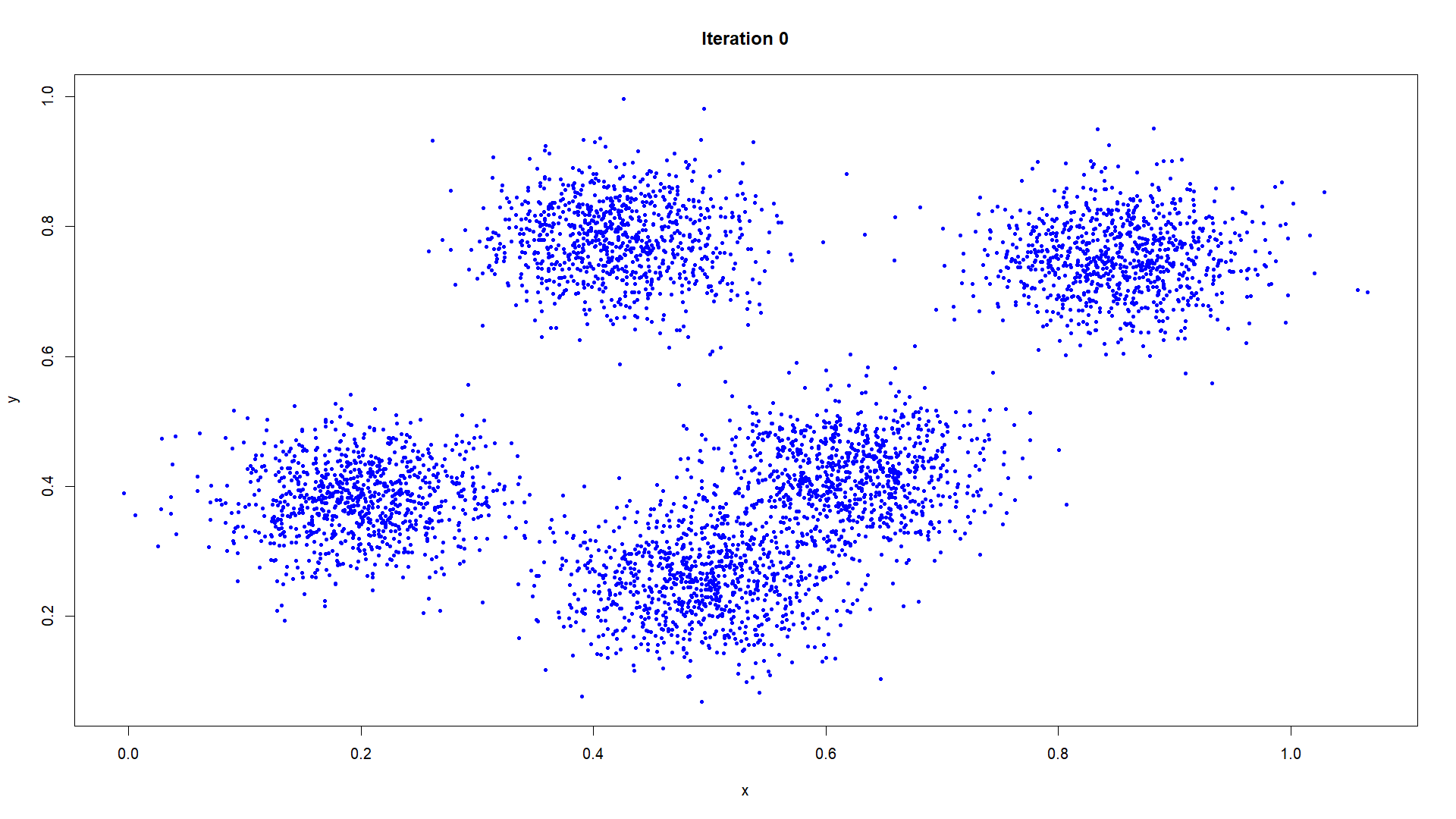


Рисунок — Начальные данные.

Каждая синяя точка является одним объектом. Далее вручную выберем 5 начальных центров кластеризации и нанесем их красными точками на рисунок (рис. 4).

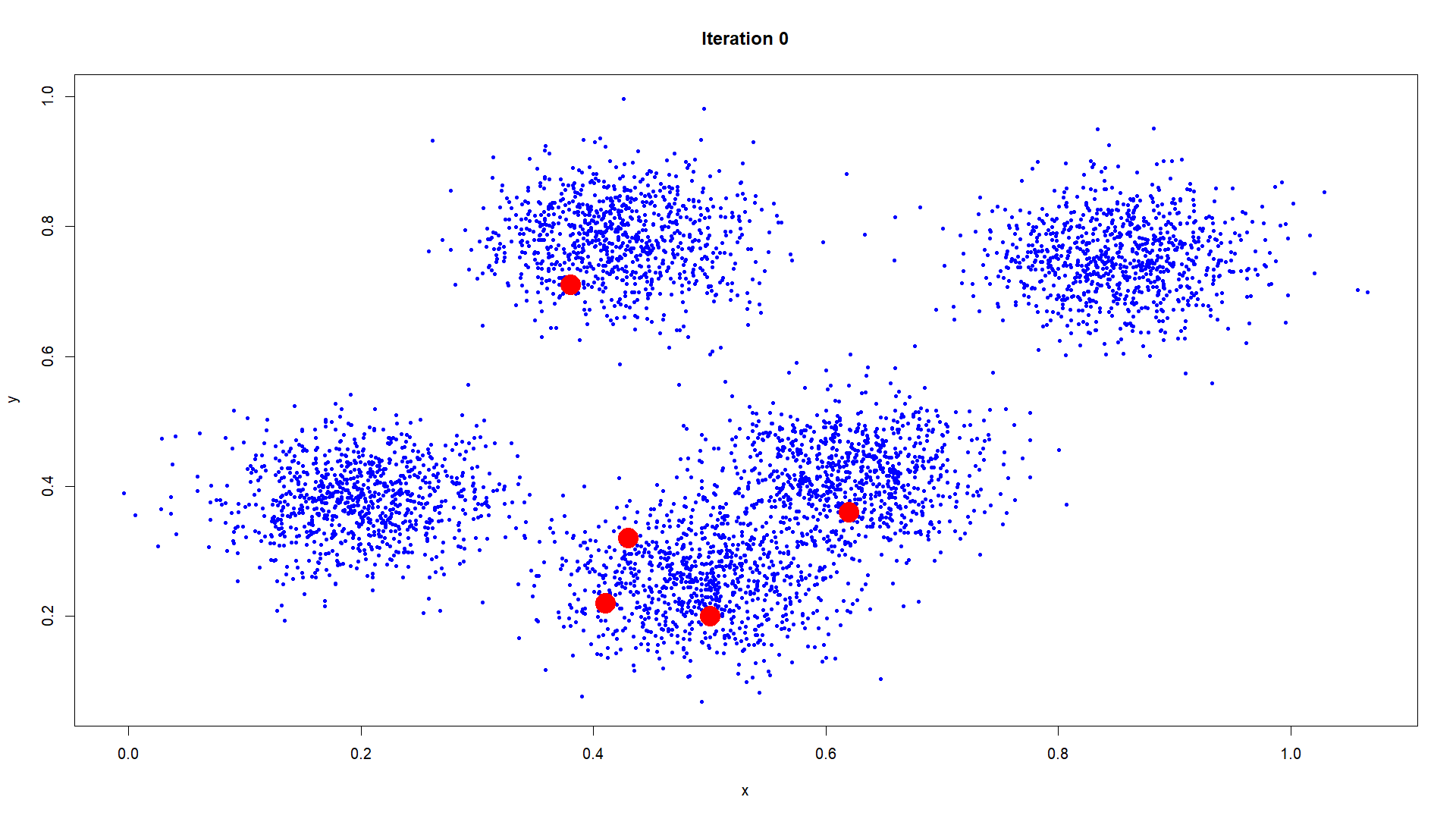


Рисунок — Начальные данные с вручную выбранными начальными центрами кластеризации.

Теперь распределим данные согласно алгоритму, представленному в таблице 2. Каждый цвет отражает принадлежности точки в определенному кластеру (рис. 5).

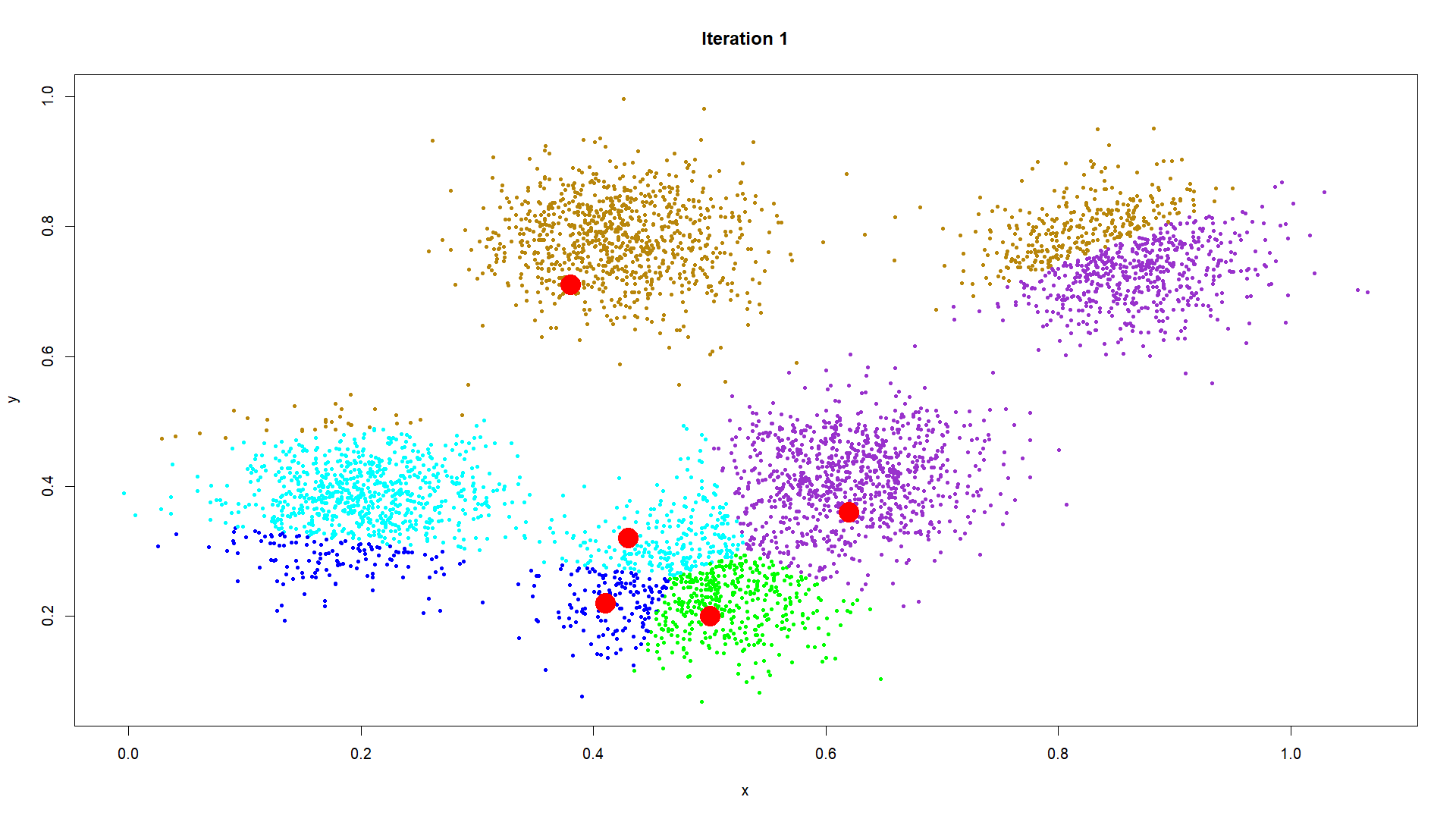


Рисунок — Первая итерация алгоритма кластеризации.

Далее отметим новые центры кластеров, как центры тяжести соответствующих кластеров, и соединим их стрелочками со старыми центрами, чтобы увидеть динамику итерации (рис. 6).

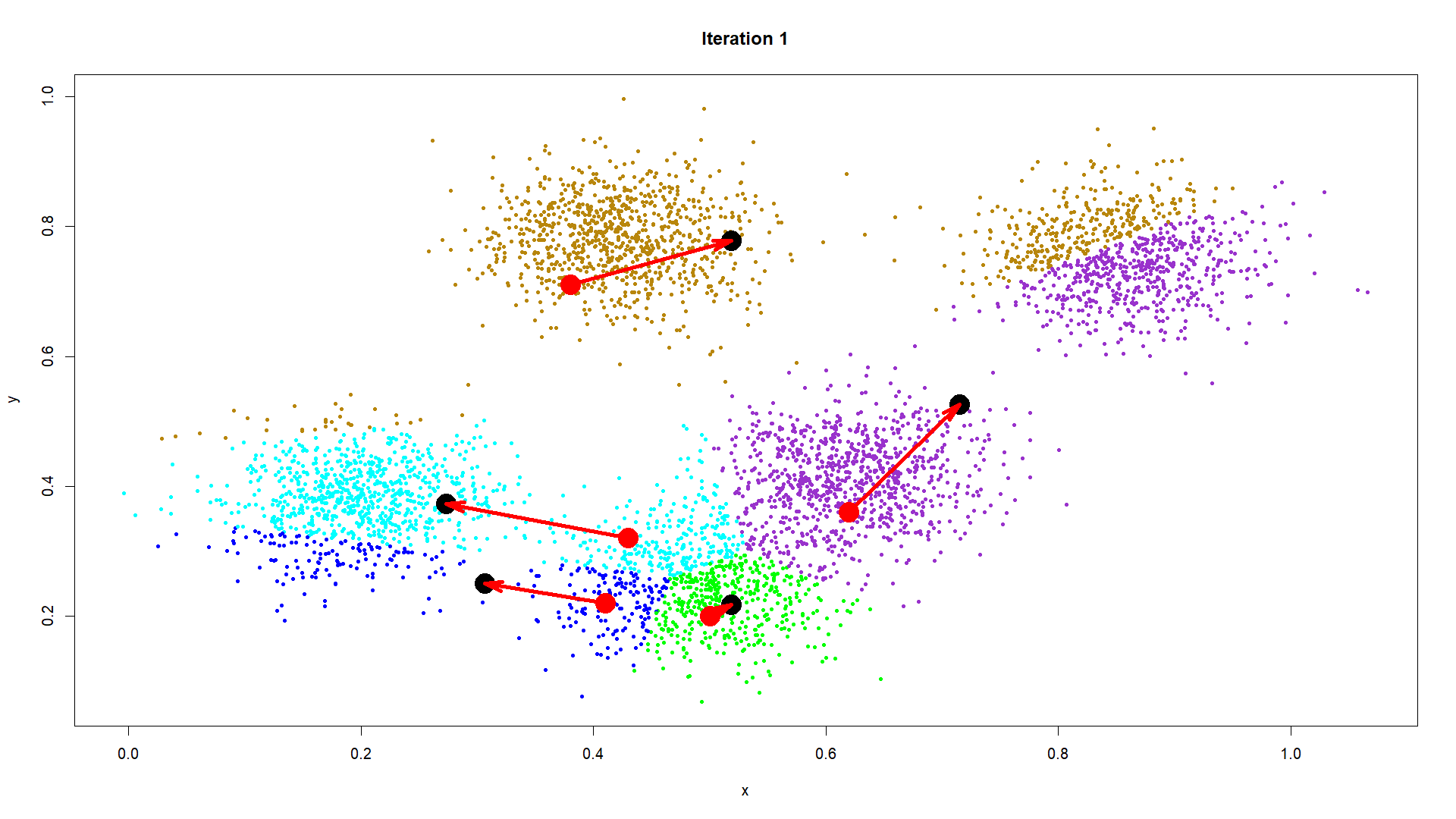


Рисунок — Первая итерация алгоритма с новыми центроидами.

На второй итерации кластеры будут формироваться исходя из центров, отмеченных черными точками на рисунке 6. Тогда центроиды, соответствующие второй итерации будут отмечены снова красным, а следующие центры (3 итерация) — черным (рис 7).

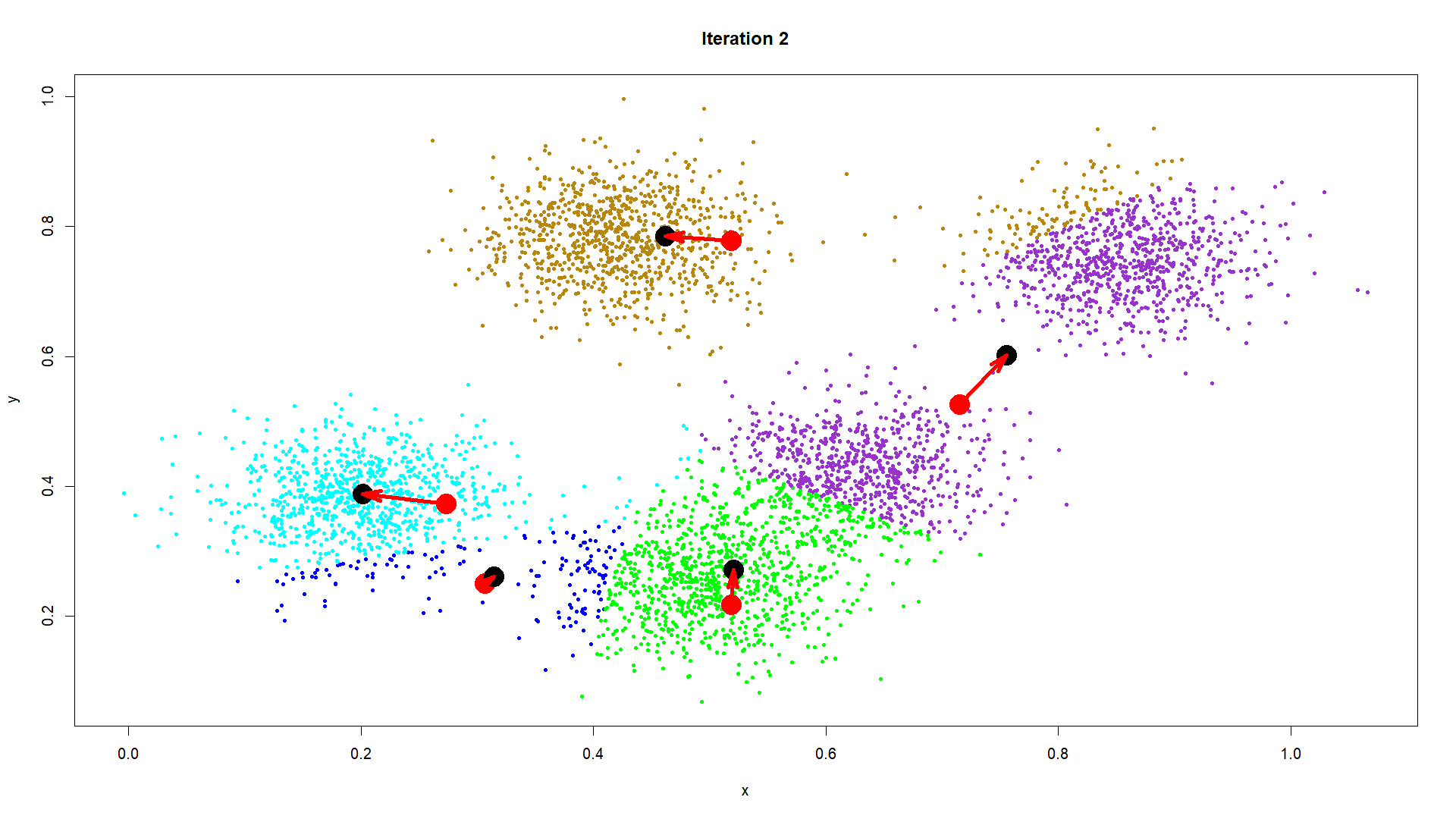


Рисунок 8 — Вторая итерация.

Аналогично происходят все последующие итерации алгоритма.

На рисунках 8–12 представлены первые 16 итераций алгоритма. Можем заметить, что уже начиная с 7 итерации различия трудно заметны. Очевидно, что это лишь показательный пример и в реальных задачах обычно не бывает такой ярко выраженной изначальной отделенности кластеров, и количество итераций может быть существенно больше.

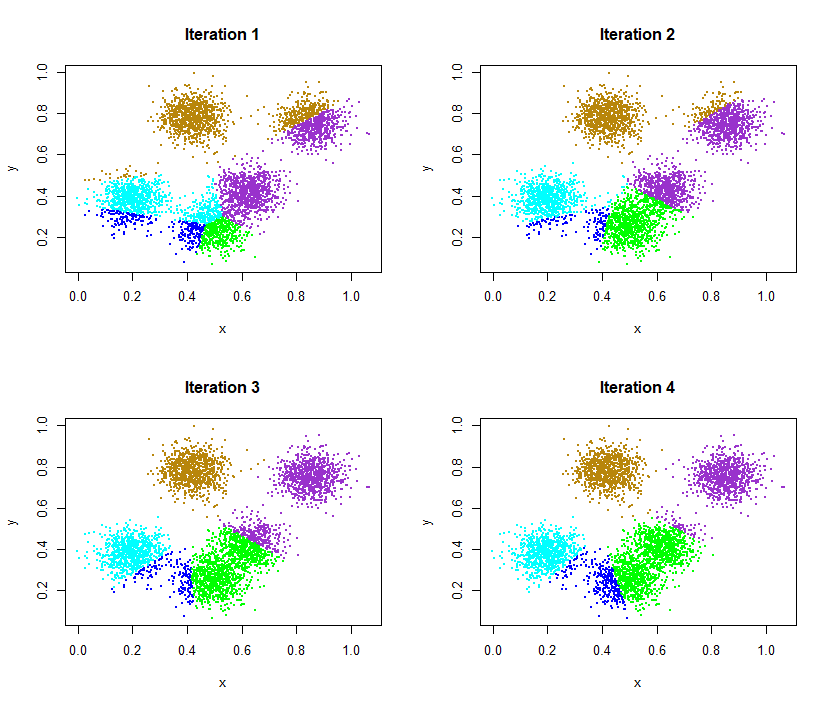


Рисунок — Итерации 1–4.

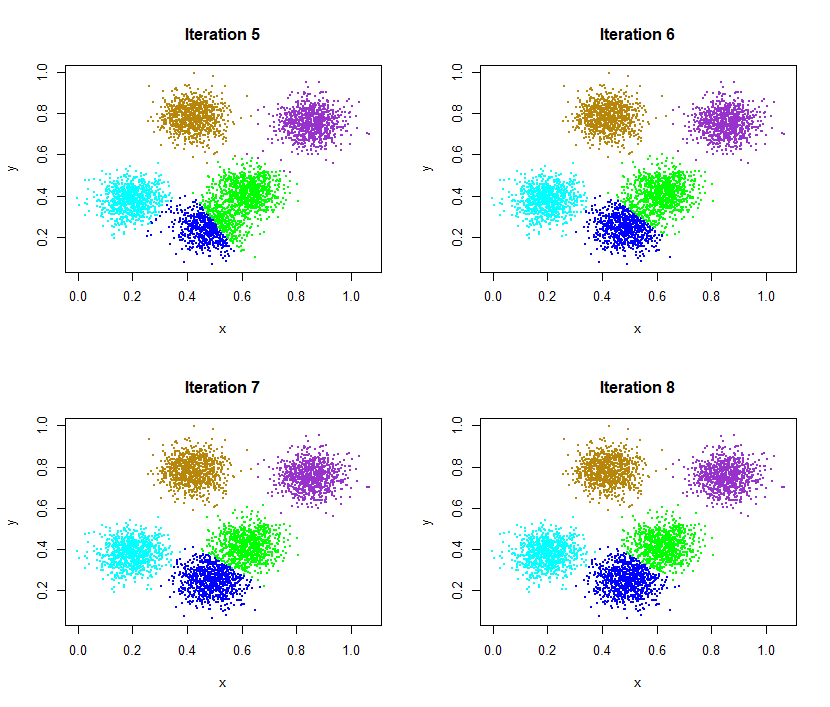


Рисунок 10 — Итерации 5–8.

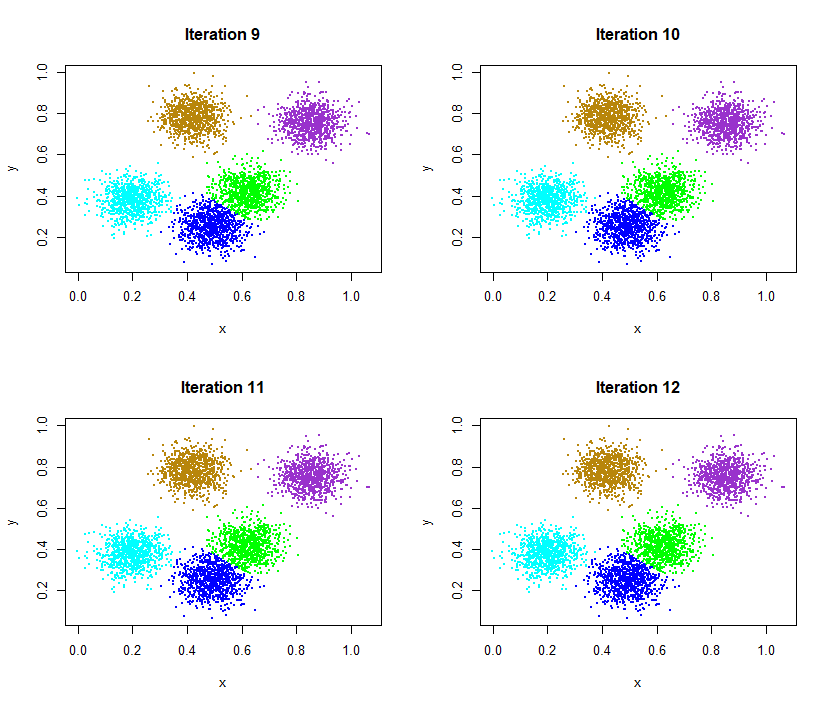


Рисунок — Итерации 9–12.

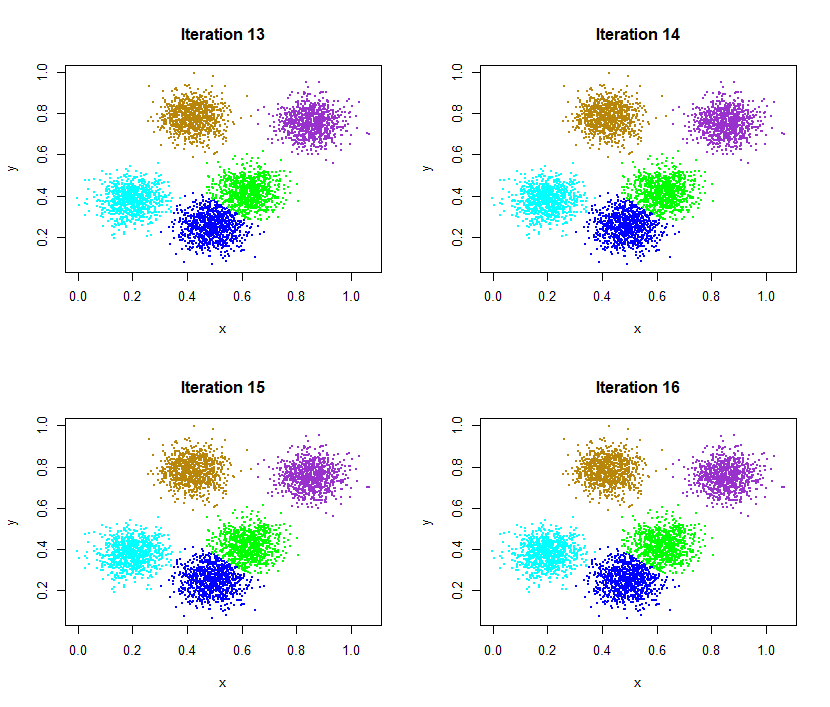


Рисунок — Итерации 13-16.

.. Ядерный k-means

Ядерный k-means (kernel k-means) является расширением стандартного k-means путем нелинейного отображения объектов данных в многомерное пространство признаков. Таким образом, он обладает возможность обнаруживать кластеры, которые не являются линейно разделимыми в исходном пространстве (рисунок 12).

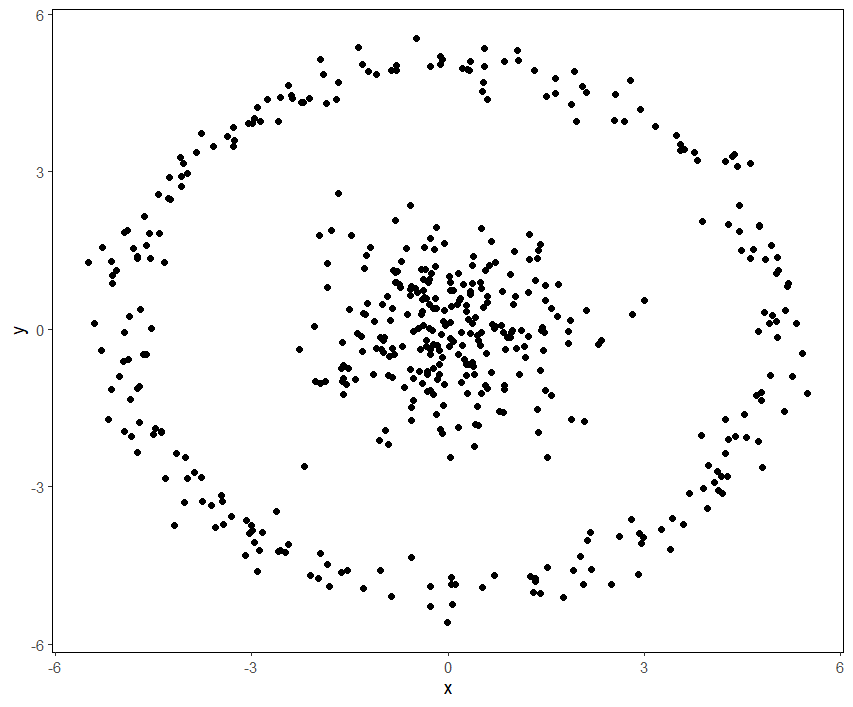


Рисунок — Пример линейно не разделимого множества.

Обычно расширение k-means до kernel k-means реализуется путем выражения расстояния в виде ядерной функции, как

, ()

где — нелинейная функция отображения. Некоторые популярные ядерный функции перечислены в таблице 3.

Таблица — Примеры популярный ядерных функций

|  |  |
| --- | --- |
| Гауссово ядро  (радиально-базисная функция) |  |
| Полиномиальное ядро |  |
| Сигмоидальное ядро |  |
| Экспоненциальное ядро |  |
| Лапласианово ядро |  |

Тогда имеем простой способ преобразовать вычисление евклидова расстояния в пространстве признаков в ядерном случае — использовать ядерный трюк [50], как показано в уравнении 8.

()

Или в случае Пирсоновской корреляции, как показано в уравнении 9.

()

Ключевая идея метода kernel k-means заключается в вычислении метрики между объектами и центроидами кластеров в пространстве признаков:

()

Остальные шаги алгоритма такие же, как и для обычного k-means. Основным недостатком данного алгоритма можно считать повышение вычислительной стоимости за счет многократных вычислений значений ядерной функции.

На рисунках 13 и 14 представлено сравнение результата работы k-means и kernel k-means для линейно не разделимого множества, рассмотренного выше (рис. 12).

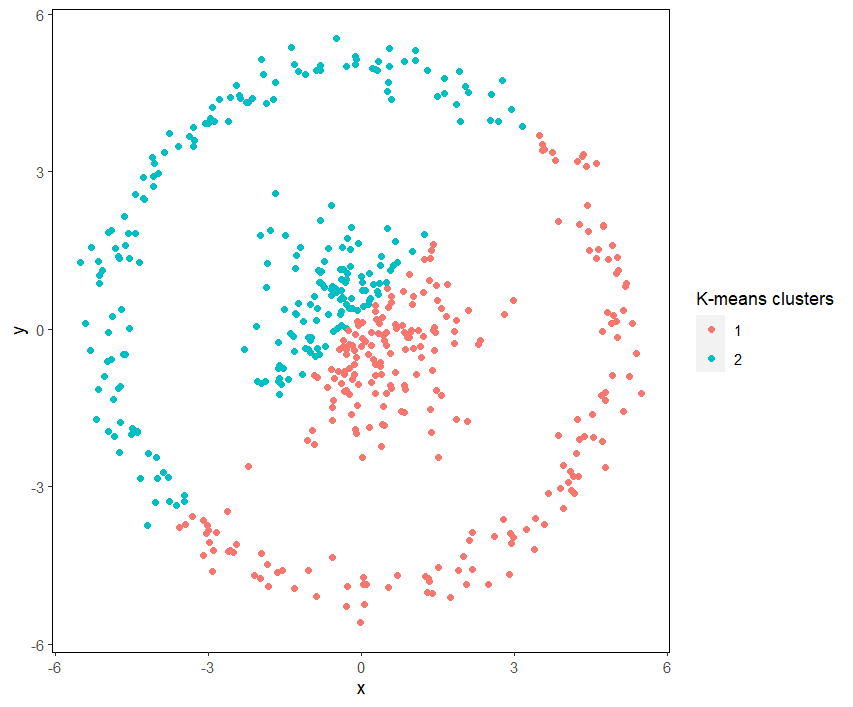


Рисунок — Базовый k-means.

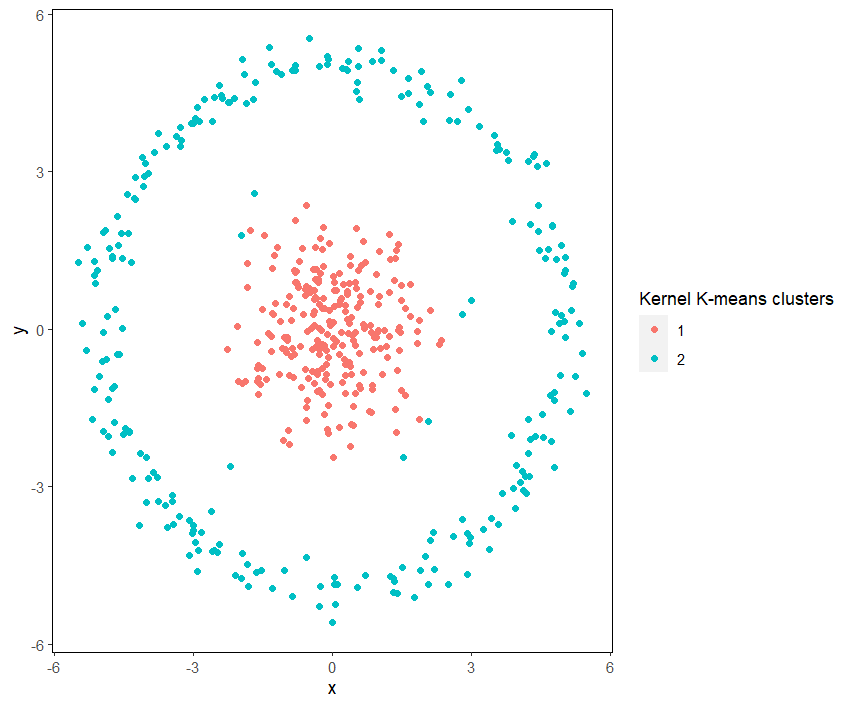


Рисунок 15 — Ядерный k-means.

В данном случае было выбрано Гауссово ядро с параметром .

3.3 k-Medoids

3.3.1 Принцип

Алгоритм k-medoids, также известный как PAM, основан на принципе, аналогичном k-means. В отличие от k-means подход k-medoids ищет среди объектов центров, которые называются medoids (медоиды), то есть центрами кластеров могут выступать не любые точки, а лишь точки, соответствующие данным. Это позволяет методу k-medoids быть устойчивым к выбросам, поскольку выбросы могут сильно влиять на центроиды, но почти не влияют на медоиды. Процедура k-medoids аналогична процедуре k-means, показанной в таблице 2, за исключением того, что на шаге 3 она обновляет медоиды каждого кластера.

.. Сравнение k-means и k-medoids. Confusion matrix. Метрики ошибок

Пример. Рассмотрим 2 выборки, в одну из которых домешаны выбросы (рис. 15).

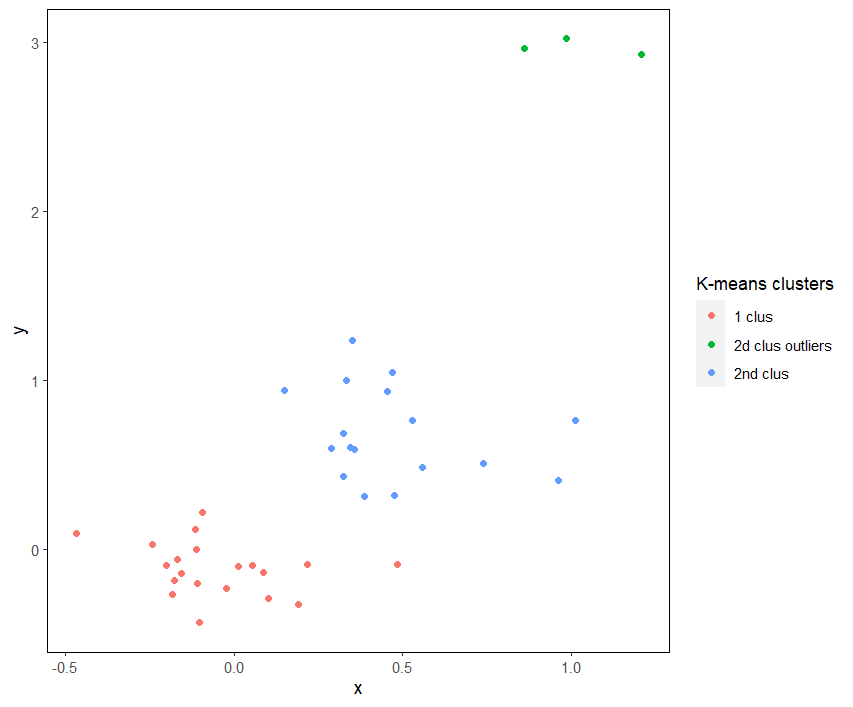


Рисунок — Начальные кластеры (красные и синие точки) и выбросы (зеленые точки), в действительности относящиеся ко второму кластеру.

Сравним работу алгоритмов k-means и k-medoids на этом примере (рис. 16, 17). В обоих случаях вручную выбрано количество будущих кластеров . Дополнительными точками отметим центры кластеризации: центроиды для k-means и медоиды для k-medoids. Для начала рассмотрим подход, при котором выбросы просто исключаются. Тогда редуцированный набор данных будет выглядеть, как показано на рисунке 16.

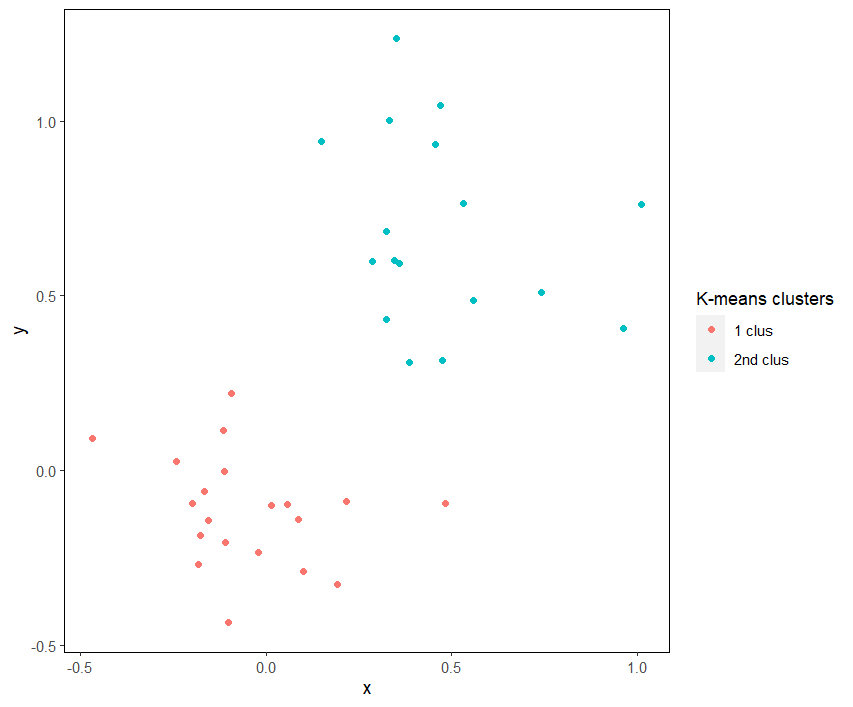


Рисунок 17 — Редуцированный на выбросы набор данных.

Тогда в этом случае применим k-means (рис. 17) и k-medoids (рис. 18).

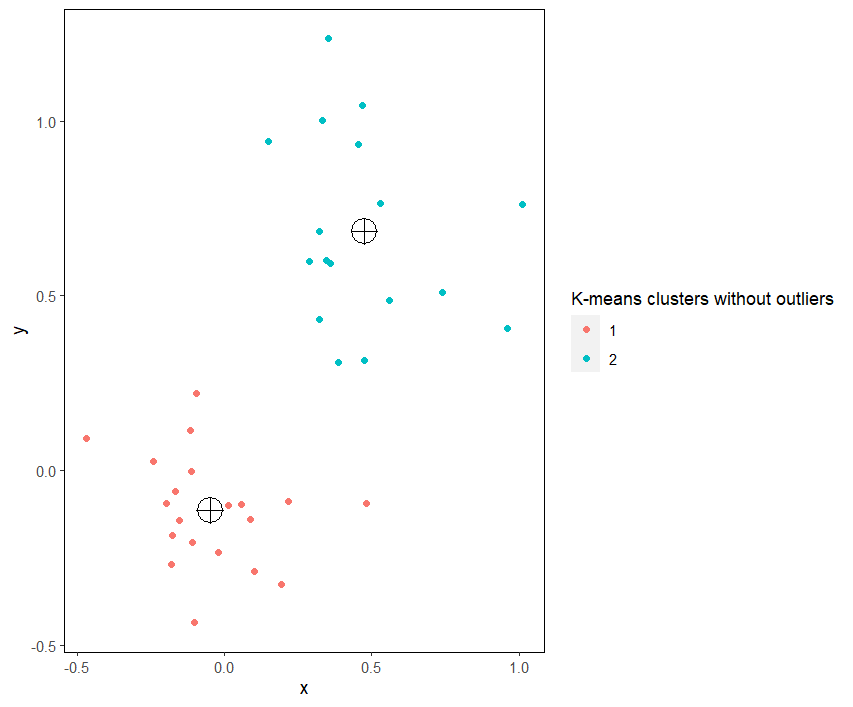


Рисунок 18 — Алгоритм k-means, примененный на редуцированном наборе данных.

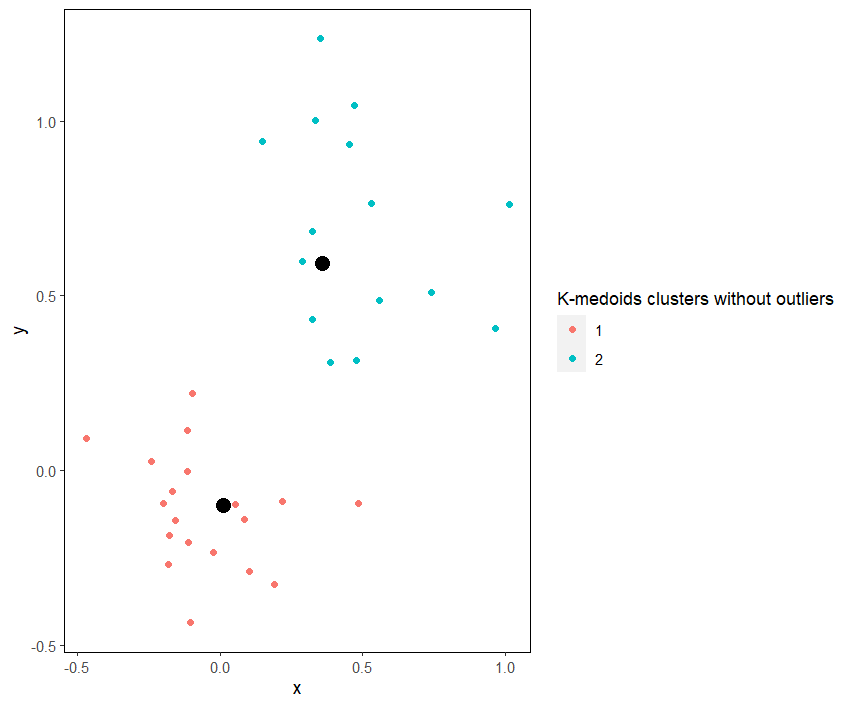


Рисунок 19 — Алгоритм k-medoids, примененный на редуцированном наборе данных.

Рассмотрим матрицу ошибок (confusion matrix) полученной кластеризации (табл. 5). Общий вид построения матрицы ошибок приведен в таблице 4.

|  |  |  |
| --- | --- | --- |
| Предсказанные величины | Действительные величины | |
| 1–ый кластер | 2–ой кластер |
| 1­–ый кластер | True Positive (TP) | False Positive (FP) |
| 2–ой кластер | False Negative (FN) | True Negative (TN) |

Таблица 4 — Матрица ошибок (confusion matrix).

|  |  |  |
| --- | --- | --- |
| Предсказанные величины | Действительные величины | |
| 1–ый кластер | 2–ой кластер |
| 1­–ый кластер | 20 | 0 |
| 2–ой кластер | 0 | 17 |

Таблица 5 — Матрица ошибок для k-means, примененного к редуцированному набору данных.

Из таблицы 4 становится ясно, что все объекты распределились в исходные кластеры. Можно формализовать это, введя несколько метрик, которые бы отражали верность предсказаний. Первой рассмотрим метрику Accuracy (ACC):

(11)

где — общее число наблюдений.

Тогда в нашем случае значение , так как все данные кластеризовались верно.

Далее введем метрики precision (точность) и recall (полнота).

, (12)

. (13)

Тогда получим и .

Можно заметить, что accuracy и precision имеют одинаковый перевод — точность, но вследствие того, что ACC, как будет показано далее, практически не используется, можно остановиться на использовании слова “точность” в отношении только precision.

Такие же результаты получаем для k-medoids в случае редуцированного набора данных (табл. 6), .

|  |  |  |
| --- | --- | --- |
| Предсказанные величины | Действительные величины | |
| 1–ый кластер | 2–ой кластер |
| 1­–ый кластер | 20 | 0 |
| 2–ой кластер | 0 | 17 |

Таблица 6 — Матрица ошибок для k-medoids, примененного к редуцированному набору данных.

Ситуации обстоит иначе, если в данных появляются выброс. Рассмотрим изначальные данные, в которых присутствовало 3 выброса, соответствующих второму кластеру. Кластеризуем их по методу k-means (рис. 19) и k-medoids (рис. 20).

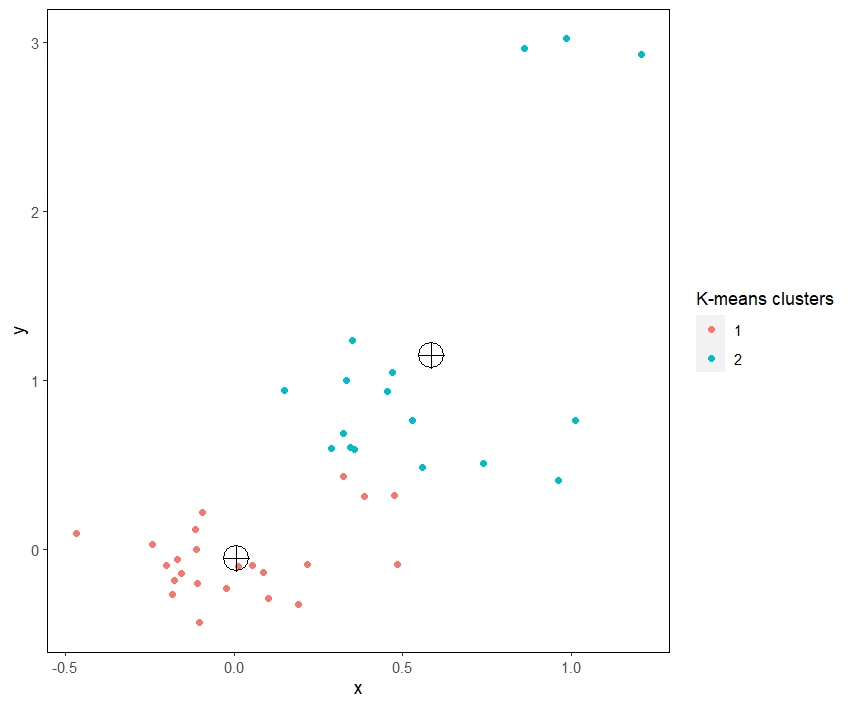


Рисунок 20 — Алгоритм k-means, примененный к исходному набору данных, содержащим выбросы.

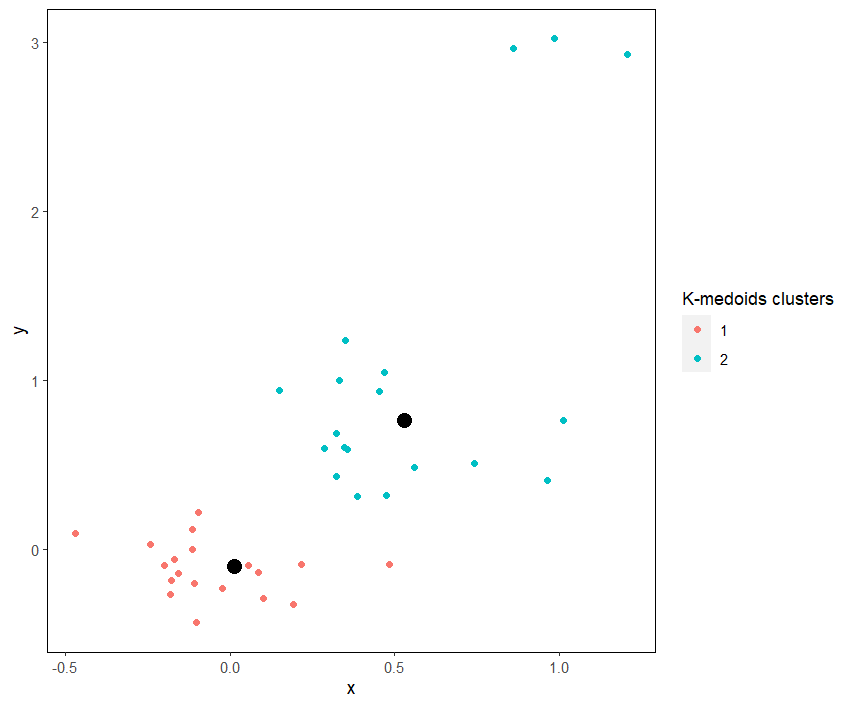


Рисунок 21 — Алгоритм k-medoids, примененный к исходному набору данных, содержащим выбросы.

Уже по рисунку становится понятно основное преимущество k-medoids — устойчивость к выбросам. Если в случае k-means, центроид достаточно сильно сместился в сторону выбросов, то медоид хоть и также сместился в сторону выбросов, но незначительно и это не помешало верно определить кластеры (табл. 6,7,8).

|  |  |  |
| --- | --- | --- |
| Предсказанные величины | Действительные величины | |
| 1–ый кластер | 2–ой кластер |
| 1­–ый кластер | 20 | 3 |
| 2–ой кластер | 0 | 17 |

Таблица 7 — Матрица ошибок для k-means, примененного к полному набору данных.

|  |  |  |
| --- | --- | --- |
| Предсказанные величины | Действительные величины | |
| 1–ый кластер | 2–ой кластер |
| 1–ый кластер | 20 | 0 |
| 2–ой кластер | 0 | 20 |

Таблица 8 — Матрица ошибок для k-medoids, примененного к полному набору данных.

|  |  |  |  |
| --- | --- | --- | --- |
| Метод кластеризации | Accuracy (ACC) | Precision (точность) | Recall (полнота) |
| k-means | 0.925 | 0.8695652 | 1 |
| k-medoids | 1 | 1 | 1 |

Таблица 9 — Сравнение значений метрик ACC, точности и полноты для k-means и k-medoids в случае полного набора данных.

3.3.3 Замечание по поводу метрик ошибок

Основываясь только на матрице ошибок, можно рассматривать множество различных метрик точности. Точность (precision) можно интерпретировать как долю объектов, определенных алгоритмом, как относящихся к первому кластеру, и при этом, действительно, являющимися представителями первого кластера. Полнота (recall) показывает, какую долю объектов первого кластера из всех объектов первого кластера обнаружил алгоритм. Две эти метрики менее нестабильны по отношению к сильно неравным классам. Рассмотрим следующий пример. Допустим, у нас имеется антифрод система и мы пытаемся оценить ее качество. Предположим, имеются данные о 110 банковских транзакциях с заранее известными реальными данными об их достоверности, из которых 100 являются нормальными и 10 — мошенническими. В результате работы нашего классификатора, который пытается определить мошеннические операции, были получены следующие результаты: . Тогда ACC:

. (14)

Однако, если мы будем считать все операции, как не мошеннические, то значение accuracy вырастет:

0.909, (15)

но при этом модель потеряет какой-либо смысл.

Введение метрики точности (precision) не дает нам определять все объекты в один кластер, так как тогда растет уровень . Можно считать, что precision демонстрирует способность алгоритм отличать один кластер от другого, а recall — способность в принципе обнаруживать данный кластер.

Случаи метрик для многоклассовой кластеризации описаны, например, здесь [53], [54].

Заключение

Список использованных источников

11. Altman, N.S. (1992). An introduction to kernel and nearest-neighbour nonparametric regression. The American Statistician, 46(3), pp. 175–185.

22. Haykin, S. (1999). Neural Networks – A Comprehensive Foundation, 3rd Edition, Pearson, Prentice Hall, Singapore.

3. Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), pp. 273–297.

4. Duda, R.O., Hart, P.E. and Stork, D.G. (2000). Pattern Classification. 2nd ed. John Wiley & Sons, Inc, New York.

5. Devroye, L., Györfi, L. and Lugosi, G. (1997). A Probabilistic Theory of Pattern Recognition. 2nd ed. Springer, New York.

6. Quinlan, J.R. (1986). Induction of decision trees. Machine Learning, 1(1), pp. 81–106.

7. Liu, K.-H. and Xu, C.-G. (2009). A genetic programming-based approach to the classification of multiclass microarray datasets. Bioinformatics, 25(3), pp. 331–337.

8. Shawe-Taylor, J. and Cristianini, N. (2004). Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge.

9. Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), pp. 123–140.

10. www.kaggle.com/amrmahmoud123/1-guide-to-ensembling-methods

11. Abu-Jamous, B., Fa, R., Roberts, D.J. and Nandi, A.K. (2013). Paradigm of tunable clustering using binarization of consensus partition matrices (Bi-CoPaM) for gene discovery. Plos One, 8(2), e56432.

12. Pena, J.M., Lozano, J.A. and Larranaga, P. (1999). An empirical comparison of four initialization methods for the K-Means algorithm. Pattern Recognition Letters, 20(10), pp. 1027–1040.

13. Xiao, X., Dow, E.R., Eberhart, R. et al. (2003). Gene Clustering Using Self-organizing Maps and Particle Swarm Optimization. Proceedings of the International Parallel and Distributed Processing Symposium, Nice, France, 22–26 April 2003. IEEE, pp. 154–163.

14. Salem, S.A., Jack, L.B. and Nandi, A.K. (2008). Investigation of self-organizing oscillator networks for use in clustering microarray data. IEEE Transactions on Nanobioscience, 7(1), pp. 65–79.

15. Bailey, T.L. and Elkan, C. (1994). Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, Stanford, CA, 14–17 August 1994. AAAI Press, Menlo Park, CA, pp. 28–36.

16. Eisen, M. (1999). Cluster and TreeView, http://rana.lbl.gov/manuals/ClusterTreeView.pdf.

17. Dimitriadou, E., Weingessel, A. and Hornik, K. (2002). A combination scheme for fuzzy clustering. International Journal of Pattern Recognition and Artificial Intelligence, 16, pp. 901–912.

18. Ng, A.Y., Jordan, M.I. and Weiss, Y. (2001). On Spectral Clustering: Analysis and an algorithm, in Advances in Neural Information Processing Systems (eds T.G. Dietterich, S. Becker and Z. Ghahramani), MIT Press, Cambridge, MA, 14, pp. 849–856.

19. Avogadri, R. and Valentini, G. (2008). Ensemble clustering with a fuzzy approach, in Supervised and Unsupervised Ensemble Methods and their Applications Studies in Computational Intelligence (ed O. Okun), Springer-Verlag, Berlin.

20. Cheng, Y. and Church, G.M. (2000). Biclustering of expression data. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, Boston, MA, 16–23 August 2000. AAAI Press, Menlo Park, CA, pp. 93–103.

21. Dash, Sajal & Rahman, Sarthok Rasique & Hines, Heather & Feng, Wu. (2018). Incremental BLAST: incremental addition of new sequence databases through e-value correction. 10.1101/476218.

22. Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), pp. 645–678.

23. Scherf, U., Ross, D.T., Waltham, M. et al. (2000). A gene expression database for the molecular pharmacology of cancer. Nature Genetics, 24(3), pp. 236–244.

24. Stein, T., Morris, J.S., Davies, C.R. et al. (2004). Involution of the mouse mammary gland is associated with an immune cascade and an acute-phase response, involving LBP, CD14 and STAT3. Breast Cancer Research, 6(2), pp. R75–R91.

25. D’haeseleer, P. (2005). How does gene expression clustering work? Nature Biotechnology, 23(12), pp. 1499–1502.

26. Xu, R. and Wunsch, D. (2010). Clustering algorithms in biomedical research: a review. IEEE Reviews in Biomedical Engineering, 3, pp. 120–154.

27. Cheng, Y. and Church, G.M. (2000). Biclustering of expression data. Proceedings of the International Conference on Intelligent Systems for Molecular Biology, San Diego, 19–23 August 2000,. International Society for Computational Biology (ISCB), pp. 93–103.

28. math.mit.edu/~levine/18.312/alg-comb-lecture-4.pdf.

29. Steinhaus, H. (1956). Sur la division des corps matériels en parties. Bulletin de l’Académie Polonaise des Sciences, 1, pp. 801–804.

30. MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 21 June–18 July 1965 and 27 December 1965–7 January 1966, University of California Press, Berkeley, CA, 14, pp. 281–297.

31. Kaufman, L. and Rousseeuw, P.J. (1987). Clustering by Means of Medoids, Delft Faculty of Mathematics and Informatics.

32. Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, Inc, New York.

33. Herwig, R., Poustka, A.J., Müller, C. et al. (1999). Large-scale clustering of cDNA -fingerprinting data. Genome Research, 9(11), pp. 1093–1105.

34. Tavazoie, S., Hughes, J.D., Campbell, M.J. et al. (1999). Systematic determination of genetic network architecture. Nature Genetics, 22(3), pp. 281–285.

35. Heintzman, N.D., Stuart, R.K., Hon, G.C. et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics, 39(3), pp. 311–318.

36. Heintzman, N.D., Hon, G.C., Hawkins, R.D. et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature, 459(7243), pp. 108–112.

37. Pollard, K.S. and van de Laan, M.J. (2002). Statistical inference for simultaneous clustering of gene expression data. Mathematical Biosciences, 176(1), pp. 99–121.

38. Harbison, C.T., Gordon, D.B., Lee, T.I. et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature, 431(7004), pp. 99–104.

39. Elsheikh, S.E., Green, A.R., Rakha, E.A. et al. (2009). Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. Cancer Research, 69(9), pp. 3802–3809.

40. Zhang, R. and Rudnicky, A. (2002) A Large Scale Clustering Scheme for Kernel K-Means. Proceedings of the 16th International Conference on Pattern Recognition, 11–15 August 2002, Québec City, Canada, 4, pp. 289–293.

41. Yu, S., Tranchevent, L.C., Liu, X. et al. (2012). Optimized data fusion for kernel K-means clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(5), pp. 1031–1039.

42. Sheng, W. and Liu, X. (2006). A genetic k-medoids clustering algorithm. Journal of Heuristics, 12(6), pp. 447–466.

43. Krishna, K. and Murty, M.N. (1999). Genetic K-means algorithm. Systems, Man, and Cybernetics, Part B; IEEE Transactions on Cybernetics, 29(3), pp. 433–439.

44. Lu, Y., Lu, S., Fotouhi, F. et al. (2004). Incremental genetic K-means algorithm and its application in gene expression data analysis. BMC Bioinformatics, 5(1), p. 172.

45. Dikmen, M., Hoiem, D. and Huang, T.S. (2012). A Data Driven Method for Feature Transformation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 16–21 June 2012, IEEE, Providence, RI, pp. 3314–3321.

46. Garey, M.; Johnson, D.; Witsenhausen, H. (1982-03-01). "The complexity of the generalized Lloyd - Max problem (Corresp.)". IEEE Transactions on Information Theory. 28 (2): 255–256.

47. Forgy, E.W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics, 21, pp. 768–769.

48. MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 21 June–18 July 1965 and 27 December 1965–7 January 1966, University of California Press, Berkeley, CA, 14, pp. 281–297.

49. Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, Inc, New York.

50. Girolami, M. (2002). Mercer kernel-based clustering in feature space. IEEE Transactions on Neural Networks, 13(3), pp. 780–784.

51. Holland, J.H. (1975). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, University of Michigan Press, Ann Arbor, MI.

52. Krishna, K. and Murty, M.N. (1999). Genetic K-means algorithm. Systems, Man, and Cybernetics, Part B; IEEE Transactions on Cybernetics, 29(3), pp. 433–439.

53. scikit-learn.org/stable/modules/model\_evaluation.html#multiclass-and-multilabel-classification.

54. towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

Приложения

Приложение

2191531