[**Введение** 2](#_Toc8674000)

[**1.** **Теоретические основы кластерного анализа** 3](#_Toc8674001)

[**1.1** **Математическая модель кластерного анализа** 4](#_Toc8674002)

[**1.2** **Практическое применение кластерного анализа** 7](#_Toc8674003)

[**1.3** **Методы кластерного анализа** 9](#_Toc8674004)

[**1.4** **Методы вычисления расстояния между объектами** 12](#_Toc8674005)

[**1.5** **Методы объединения кластеров** 13](#_Toc8674006)

[**1.6**  **Кластерный анализ в сегментации потребителей на основе их интересов** 14](#_Toc8674007)

[**Литература** 15](#_Toc8674008)

# **Введение**

В современном мире, где количество данных очень велико, распространена проблема быстрой и эффективной обработки данных с большим объемом. В их числе проблема сегментации потребителей различных ресурсов, на группы, в котором отличие пользователей одной группы сводились бы к минимуму. Для решения таких задач, как правило, используются различные алгоритмы кластеризации. Кластеризация (или кластерный анализ) – задача разбиения множества объектов на непересекающиеся подмножества, называемых кластерами.

Алгоритмы кластеризации дают возможность разделения большого количества входных данных на неявные классы, посредством распределения каждого объекта в группы в котором, различия между объектами в одной группе сводится к минимуму.

Цель данной дипломной работы – исследование алгоритмов кластеризации различных объектов, по их характеристикам, для выделения схожих объектов в кластеры, на примере кластеризации профилей пользователей на основе их интересов. Основная цель работы реализация различных алгоритмов кластеризации данных, с использованием нескольких методов вычисления расстоянии между объектами. Основной задачей работы является построение автоматизированной информационной системы, которая обрабатывает входные, в различных форматах, и на выходе дает распределение по кластерам исходных данных, в удобном для изучения формате.

Объект исследования – различные алгоритмы кластеризации, инструменты необходимые для реализации и визуализации, такие как: c# - объектно-ориентированный язык программирования, ASP.NET Core 2 – платформа разработки веб-приложений.

Задачи для достижения поставленной цели:

1. Системный анализ и изучение предметной области
2. Формирование требований к автоматизированной системе кластеризации входных данных;
3. Создание структуры программы
   1. Реализация трехуровневой архитектуры программы
   2. Реализация алгоритмов кластеризации
   3. Разделение функционала вычисления расстояний между объектами от работы алгоритма
   4. Реализация обработки входных данных
   5. Создание интерфейса пользователя
   6. Создание БД, с использованием метода “Code First”
   7. Генерация тестовых данных
   8. Тестирование проекта
4. Анализ работы алгоритмов

# **Теоретические основы кластерного анализа**

Кластерный анализ - многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя. Первое применение кластерного анализа был в области социологии. Впервые в 1939 году был определен предмет кластерного анализа и сделано его описание исследователем Трионом (Tryon). Главное его назначение – разбиение данных на взаимно не пересекающиеся множества, с наименьшей попарной разницей между объектами каждого множества (кластера). Тем самым кластерный анализ решает задачу классификации и выявления структуры данных.

Достоинством кластерного анализа является, возможность разбиения объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличии от большинства математическо-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры, когда показатели имеют разнообразный вид, затрудняющий применение традиционных экономических подходов. Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы разреженных данных, путём разделения их по схожести на группы, с наименьшей отличием объектов одной группы, делая их компактными и наглядными.

# **Математическая модель кластерного анализа**

**Задача кластерного анализа**

Задача заключается в том, чтобы на основании данных, содержащихся во множестве , разбить множество объектов кластеров (подмножеств) , так, чтобы каждый объект принадлежал только одному подмножеству разбиения. А объекты, принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам, были однородными.

**Математические характеристики кластера**

Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера. Центр кластера – это среднее геометрическое место точек в пространстве переменных. Радиус кластера – максимальное расстояние точек от центра кластера. Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют спорными. Спорный объект – это объект, который по мере сходства может быть отнесен к нескольким кластерам. Размер кластера может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

**Неоднозначность задачи кластерного анализа**

Неоднозначность данной задачи может быть устранена экспертом или аналитиком. Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта в принципе допускают желательное разбиение пула (совокупности) объектов на кластеры. Второе предположение - правильность выбора масштаба или единиц измерения признаков. Выбор масштаба в кластерном анализе имеет большое значение. Рассмотрим пример. Представим себе, что данные признака х в наборе данных А на два порядка больше данных признака у: значения переменной х находятся в диапазоне от 100 до 700, а значения переменной у - в диапазоне от 0 до 1. Тогда, при расчете величины расстояния между точками, отражающими положение объектов в пространстве их свойств, переменная, имеющая большие значения, т.е. переменная х, будет практически полностью доминировать над переменной с малыми значениями, т.е. переменной у. Таким образом из-за неоднородности единиц измерения признаков становится невозможно корректно рассчитать расстояния между точками. Эта проблема решается при помощи предварительной стандартизации переменных.

**Стандартизация**

Стандартизация или нормирование приводит значения все преобразованных переменных к единому диапазону значений путем выражения через отношение этих значений к некой величине, отражающей определенные свойства конкретного признака. Существуют различные способы нормирования исходных данных:

, , где – соответственно среднее и среднеквадратическое отклонение – наибольшее и наименьшее значение . Наряду со стандартизацией переменных, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки, полученные в ходе опроса экспертов - специалистов предметной области. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом неодинакового веса переменных. В ходе экспериментов возможно сравнение результатов, полученных с учетом экспертных оценок и без них, и выбор лучшего из них.

**Формальное описание задачи**

Дано множество объектов данных , каждый из которых представлен набором атрибутов. Требуется построить множество кластеров , и отображение множества на множество , т.е. Отображение задает модель данных, являющуюся решением задачи. Множество определяется следующим образом:

Где – исследуемый объект. Каждый из объектов характеризуется набором параметров:

Каждая переменная может принимать значения из некоторого множества:

Задача кластеризации состоит в построении множества:

Здесь – кластер, содержащий похожие друг н друга объекты из множества :

Где – величина, определяющая меру близости для включения объектов в один кластер, – мера близости между объектами, называемая расстоянием.

**Расстояние**

Неотрицательное значение называется расстоянием между элементами , если выполняются следующие условия:

1. , для всех
2. , тогда и только тогда, когда
3. . Если расстояние меньше некоторого значения , то говорят, что элементы близки и помещаются в один кластер. В противном случае говорят, что элементы отличны друг от друга и их помещают в разные кластеры.

Алгоритмы(общее)

Большинство популярных алгоритмов, решающих задачу кластеризации, используют в качестве формата входных данных матрицу отличия . Строки и столбцы матрицы соответствуют элементам множества . Элементами матрицы являются значения в строке . Очевидно, что на главной диагонали значения будут равны нулю:

Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин:

1. Не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию “По построению”. Все они могут давать разные результаты.
2. Число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием.
3. Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

# **Практическое применение кластерного анализа**

**В биологии**

В биологии кластеризация имеет множество приложений в самых разны областях. Например, в биоинформатике с помощью неё анализируются сложные сети взаимодействующих генов, состоящие порой из сотен или даже тысяч элементов. Кластерный анализ позволяет выделить подсети, узкие места, концентраторы и другие скрытые свойства изучаемой системы, что позволяет в конечном счете узнать вклад каждого гена в формирование изучаемого феномена.

В области экологии широко применяется для выделения пространственно однородных групп организмов, сообществ и т. п. Реже методы кластерного анализа применяются для исследования сообществ во времени. Гетерогенность структуры сообществ приводит к возникновению нетривиальных методов кластерного анализа (например, метод Чекановского).

В общем стоит отметить, что исторически сложилось так, что в качестве мер близости в биологии чаще используются меры сходства, а не меры различия.

**В социологии**

При анализе результатов социологических исследований рекомендуется осуществлять анализ методами иерархического агломеративного семейства, а именно методом Уорда, при котором внутри кластеров оптимизируется минимальная дисперсия, в итоге создаются кластеры приблизительно равных размеров. Метод Уорда наиболее удачен для анализа социологических данных. В качестве меры различия лучше квадратичное евклидово расстояние, которое способствует увеличению контрастности кластеров. Главным итогом иерархического кластерного анализа является дендрограмма или «сосульчатая диаграмма». При её интерпретации исследователи сталкиваются с проблемой того же рода, что и толкование результатов факторного анализа — отсутствием однозначных критериев выделения кластеров. В качестве главных рекомендуется использовать два способа — визуальный анализ дендрограммы и сравнение результатов кластеризации, выполненной различными методами.

Визуальный анализ дендрограммы предполагает «обрезание» дерева на оптимальном уровне сходства элементов выборки. «Виноградную ветвь» (терминология Олдендерфера М. С. и Блэшфилда Р. К.) целесообразно «обрезать» на отметке 5 шкалы Rescaled Distance Cluster Combine, таким образом будет достигнут 80 % уровень сходства. Если выделение кластеров по этой метке затруднено (на ней происходит слияние нескольких мелких кластеров в один крупный), то можно выбрать другую метку. Такая методика предлагается Олдендерфером и Блэшфилдом.

Теперь возникает вопрос устойчивости принятого кластерного решения. По сути, проверка устойчивости кластеризации сводится к проверке её достоверности. Здесь существует эмпирическое правило — устойчивая типология сохраняется при изменении методов кластеризации. Результаты иерархического кластерного анализа можно проверять итеративным кластерным анализом по методу k-средних. Если сравниваемые классификации групп респондентов имеют долю совпадений более 70 % (более 2/3 совпадений), то кластерное решение принимается.

Проверить адекватность решения, не прибегая к помощи другого вида анализа, нельзя. По крайней мере, в теоретическом плане эта проблема не решена. В классической работе Олдендерфера и Блэшфилда «Кластерный анализ» подробно рассматриваются и в итоге отвергаются дополнительные пять методов проверки устойчивости:

1. кофенетическая корреляция — не рекомендуется и ограничена в использовании;
2. тесты значимости (дисперсионный анализ) — всегда дают значимый результат;
3. методика повторных (случайных) выборок, что, тем не менее, не доказывает обоснованность решения;
4. тесты значимости для внешних признаков пригодны только для повторных измерений;
5. методы Монте-Карло очень сложны и доступны только опытным математикам.

В информатике

Кластеризация результатов поиска – используется для “Интеллектуальной” группировки результатов при поиске файлов, веб-сайтов, других объектов, предоставляя пользователю возможность быстрой навигации, выбора заведомо более релевантного подмножества и исключения заведомо менее релевантного – что может повысить простоту и удобность пользовательского интерфейса по сравнению с выводом в виде просто сортированного списка.

Сегментация изображений – кластеризация может быть использована для разбиения цифрового изображения на отдельные области с целью обнаружения границ или распознавания объектов.

Интеллектуальный анализ данных (англ. *data mining)* — кластеризация в Data Mining приобретает ценность тогда, когда она выступает одним из этапов анализа данных, построения законченного аналитического решения. Аналитику часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель для всех данных. Таким приемом постоянно пользуются в маркетинге, выделяя группы клиентов, покупателей, товаров и разрабатывая для каждой из них отдельную стратегию.

# **Методы кластерного анализа**

Методы кластерного анализа можно разделить на две классификации:

1. **Иерархические и плоские**

Иерархические алгоритмы (алгоритмы таксономии) строят систему вложенных разбиений. Т.е. на выходе получается дерево кластеров, корнем которого является вся выборка, а листьями – наиболее мелкие кластеры.

Плоские алгоритмы строят одно разбиение объектов на кластеры.

1. **Четкие и нечеткие**

Четкие (или непересекающиеся) алгоритмы каждому объекту выборки ставят в соответствие номер кластера, т.е. каждый объект принадлежит одному и только одному кластеру.

Нечеткие (или пересекающиеся) алгоритмы каждому объекту ставят соответствие набор вещественных значений, показывающих степень отношения объекта к кластерам. Т.е. каждый объект относится к каждому кластеру с некоторой вероятностью.

**Алгоритмы иерархической кластеризации**

Среди алгоритмов иерархической кластеризации выделяются два основных типа. Восходящие и нисходящие алгоритмы.

Нисходящие алгоритмы работают по принципу “Сверху-вниз”: в начале все объекты помещаются в один кластер, который затем разбивается на все более мелкие кластеры.

Восходящие алгоритмы: в начале каждый объект помещается в отдельный кластер, а затем объединяют кластеры во все более крупные, пока все объекты выборки не будут содержатся в одном кластере. Таким образом строится система вложенных разбиений. Результаты таких алгоритмов обычно представляют в виде дерева- дендрограммы. Классический пример такого дерева – классификация животных и расстений.

**Алгоритмы квадратичной ошибки**

Задачу кластеризации можно рассматривать как построение оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование минимизации среднеквадратической ошибки разбиения:

Где – “Центр масс” кластера (точка со средними значения характеристик для данного кластера).

Алгоритмы квадратичной ошибки относятся к типу плоских алгоритмов. Самым распространенным алгоритмом этой категории является метод . Этот алгоритм строит заданное число кластеров, расположенных как можно дальше друг от друга. Работа алгоритма делится на несколько этапов:

1. Случайно выбрать точек, являющихся начальными центрами масс кластеров.
2. Отнести каждый объект к кластеру с ближайшим центром масс.
3. Пересчитать центры масс кластеров в зависимости от объектов внутри кластеров.
4. Если критерий остановки алгоритма не удовлетворен, вернутся к п. 2.

В качестве критерия остановки работы алгоритма обычно выбирают минимальное изменение среднеквадратической ошибки. Так же возможно останавливать работу алгоритма, если на шаге 2 не было объектов, переместившихся из кластера в кластер.

**Нечеткие алгоритмы**

Наиболее популярным алгоритмом нечеткой кластеризации является алгоритм . Он представляет собой модификатор метода .

Шаги работы алгоритма:

1. Выбрать начальное нечеткое разбиение объектов на кластеров путем выбора матрицы принадлежности размера .
2. Используя матрицу , найти значение критерия нечеткой ошибки: ,

Где – “центр масс” нечеткого кластера , .

4. Перегруппировать объекты с целью уменьшения этого значения критерия нечеткой ошибки.

3. Возвращаться в п.2 до тех пор, пока изменения матрицы не станут незначительными.

Этот алгоритм может не подойти, если заранее неизвестно число кластеров, либо необходимо однозначно отнести каждый объект к одному кластеру.

**Алгоритмы, основанные на теории графов**

Суть таких алгоритмов заключается в том, что выборка объектов представляется в виде графа , вершинами которого соответствуют объекты, а ребра имеют вес, равный расстоянию между объектами. Достоинством графовых алгоритмов кластеризации являются наглядность, относительная простота реализации и возможность внесения различных усовершенствований, основанные на геометрических соображениях. Основными алгоритмам являются алгоритм выделения связных компонент, алгоритм построения минимального покрывающего (остовного) дерева и алгоритм послойной кластеризации.

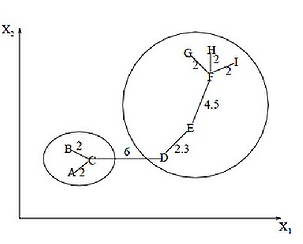
**Алгоритмы выделения связных компонент**

В алгоритме выделения связных компонент задается входной параметр и в графе удаляются все ребра, для которых расстояние больше . Соединенными остаются только наиболее близкие пары объектов. Смысл алгоритма заключается в том, чтобы подобрать такое значение , лежащее в диапазоне всех расстоянии, при котором граф разваливается на несколько связных компонент. Полученные компоненты и есть кластеры.

Для подбора параметра обычно строится гистограмма распределений попарных расстояний. В задачах с хорошо выраженной кластерной структурой данных на гистограмме будет два пика – один соответствует внутрикластерным расстояниям, второй – межкластерным расстояниям. Параметр подбирается из зоны минимума между этими пиками. При этом управлять количеством кластеров при помощи порога расстояния довольно затруднительно.

Алгоритм минимального покрывающего дерева

Алгоритм минимального покрывающего дерева сначала строит на графе минимальное покрывающее дерево, а затем последовательно удаляет ребра с наибольшим весом. На рисунке изображено минимальное покрывающее дерево, полученное для девяти объектов.



Путём удаления связи, помеченной CD, с длиной равной 6 единицам (ребро с максимальным расстоянием), получаем два кластера: {A, B, C} и {D, E, F, G, H, I}. Второй кластер в дальнейшем может быть разделён ещё на два кластера путём удаления ребра EF, которое имеет длину, равную 4,5 единицам.

# **Методы вычисления расстояния между объектами**

Для сравнения двух объектов необходим критерий, по которому будет происходить сравнение. Как правило, таким критерием является расстояние между объектами.

Евклидово расстояние – наиболее распространенная функция меры расстояния. Представляет собой геометрическое расстояние объектов в многомерно пространстве.

Квадрат Евклидового расстояния – применяется для увеличения расстояния для более отдаленных друг от друга объектов.

Расстояние городских кварталов (Манхэттенское расстояние) – Это расстояние суммы разностей по координатам. В большинстве случаев приводит к таким же результатам, как и для обычного расстояния Евклида. Однако, для этой меры влияние больших разностей уменьшается, т.к. не возводится в квадрат.

Расстояние Чебышева – Это расстояние может оказаться полезным, когда нужно определить два объекта разными, если они различаются по какой-либо одной координате. Расстояние Чебышева вычисляется по формуле:

Степенное расстояние – применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Степенное расстояние вычисляется по формуле:

Где – параметры, определяемы пользователем. Параметр ответственен за постепенное взвешивание разностей по отдельным координатам, параметр ответственен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра равны двум, то это расстояние совпадает с расстоянием Евклида.

# **1.5 Методы объединения кластеров**

В случае использования иерархических алгоритмов встает вопрос, как объединять между собой кластеры, как вычислять расстояния между ними. Существует несколько методов.

1. Одиночная связь (расстояния ближайшего соседа) – в этом методы расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями). Этот метод обычно работает очень хорошо, когда объекты происходят из отдельных групп. Если же кластеры имеют удлиненную форму или их естественный тип является “Цепочным”, то этот метод непригоден.
2. Полная связь (расстояние наиболее удаленных соседей) – В этом методе расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. наиболее удаленными соседями). Этот метод обычно работает очень хорошо, когда объекты происходят из отдельных групп. Если же кластеры имеют удлиненную форму или их естественный тип является «цепочечным», то этот метод непригоден.
3. Невзвешенное попарное среднее - В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты формируют различные группы, однако он работает одинаково хорошо и в случаях протяженных, “цепочечного» типа, кластеров.
4. Взвешенное попарное среднее - Метод идентичен методу невзвешенного попарного среднего, за исключением того, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента. Поэтому данный метод должен быть использован, когда предполагаются неравные размеры кластеров.
5. Невзвешенный центроидный метод – в этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.
6. Взвешенный центроидный метод (медиана) – этот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учета разницы между размерами кластеров. Поэтому, если имеются или подозреваются значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.

# **Кластерный анализ в сегментации потребителей на основе их интересов**

# **Литература**

1. Кластерный анализ [Электронный ресурс]: Материал из Национальной библиотеки им. Н. Э. Баумана. Режим доступа:URL - <https://ru.bmstu.wiki/Кластерный_Анализ> .
2. Обзор алгоритмов кластеризации данных [Электронный ресурс]. Режим доступа: URL - <https://habr.com/ru/post/37185/>
3. Кластерный анализ [Электронный ресурс]. Режим доступа: URL - <https://ru.wikipedia.org/wiki/Кластерный_Анализ>